EasyChair Preprint

№ 2317

Robust MUSIC Based TDOA Estimation in Competing-Speaker Scenarios

Md Ahsan Habib, Yi Zhou and Feng Ni

January 5, 2020

ROBUST MUSIC BASED TDOA ESTIMATION IN COMPETING-SPEAKER SCENARIOS

Habib Md Ahsan, Yi Zhou, Feng Ni

Chongqing University of Posts and Telecommunications,
School of Communication and Information Engineering, Chongqing, China
mdahsanhabib@qq.com, zhouy@cqupt.edu.cn, chris.ni@foxmail.com

ABSTRACT

Deep neural network (DNN) based time difference of arrival (TDOA) estimation methods such as Multiple Signal Classification (MUSIC) report superior performance in noisy and reverberation environments but the degradation is observed in the presence of competing for interference. This study investigates it's potential for robust MUSIC-based TDOA estimation in competing-Speaker scenarios. First, a time-frequency (TF) mask which is 0 for non-speech TF bins and 1 for speech TF bins based on the phase and DNN is proposed to accurately estimate the spatial covariance matrix (SCM) that are relatively clean for the MUSIC algorithm in this paper. Second, the proposed approach further reduces the search space to drastically decrease the computation cost by leveraging phase information above. Experimental results on simulated and recorded data confirm the effectiveness and the superiority of the proposed MUSIC-based TDOA estimation method in competing-Speaker scenarios, in comparison with baseline methods.

***Index Terms*—** TDOA, MUSIC, DNN, phase, computation cost

1. INTRODUCTION

The time difference of arrival (TDOA) estimation plays a core role in array signal processing such as adaptive beamforming, acoustic source localization and tracking which are widely employed in human-computer interaction, surveillance and other applications [1, 2]. However, the existence of multiple sources in the noisy and reverberant environments degrades the performance of TDOA estimation significantly, resulting in the erroneous or biased target source location.

Over the past years, a few TDOA estimation algorithms have been devised. The generalized cross-correlation with phase transform (GCC-PHAT) [3] or the steered response power with phase transform (SRP-PHAT) [4] algorithm and the MUSIC [5] algorithm are the most popular techniques in sound source localization. GCC-PHAT algorithm has the limitation which was that errors were accentuated where the signal power was low. The basic idea of MUSIC algorithm is to conduct eigenvalue decomposition for the SCM of microphone array input data, resulting in a target signal subspace orthogonal with a noise subspace corresponding to the target signal component. However, in noisy and reverberant environments with competing interferences, the target signal subspace constructed from the eigenvectors corresponding to the largest eigenvalues of the input signal SCM in the MUSIC algorithm would be biased. To solve this problem, several approaches were proposed. For example, Schwartz et al. [6] used complex gaussian mixture model to perform clustering of multiple speakers' TDOAs, assuming each gaussian model been associated with one source. But it generally suffers from resolution problem and microphone geometry mismatches. In [7], the weighted SCM-based MUSIC method for robust TDOA estimation is proposed, which selects speech dominated TF bins through a long short term memory (LSTM) based mask predictor and can achieve better performance in very challenging noisy and reverberant conditions. However, some noisy TF bins might be wrongly selected due to the presence of competing interferences.

To address the above problem, this paper proposes a SCM weighted by a TF mask approach based on the cross-correlation the power spectrum of the observed signal in the MUSIC algorithm. Considering the SCM of TF bins dominated by noise and reverberation may alter the estimation of the TDOA in low SNR and high reverberant scenarios, the SCM is further multiplied by a mask predictor based on the LSTM, like [7], through leveraging the strong learning power of DNN. In this way, the contributions of competing interferences, background noise and reverberation in the SCM are heavily attenuated. Then the TDOA is estimated by finding a peak from the summed pseudo spectrum based on the estimated SCM to overcome the spatial aliasing ambiguity occurring at high frequencies. However, searching the peak through scanning all possible source locations on a discrete 2-D space is computationally expensive when using the MUSIC algorithm, so in real applications its implementation can be difficult. In this context, this paper proposes to search the peak at larger interval to localize rough target speech source. Then the TDOA difference between precise localization and rough localization is estimated based on the above phase information and partial differential mathematical principle. In doing so, the proposed approach drastically decreases the computation cost by reducing the search space while preserves the high resolution, accuracy, and stability of the MUSIC algorithm. The experimental results on simulated and real conditions show that the performance of the proposed method is better than the GCC-PHAT [3], MUSIC [5] and its variant based on the LSTM [7] in noisy and reverberant environments with competing interferences.

In Section 2, the TDOA estimation problem is formulated. The proposed method is introduced in Section 3. Experimental performance evaluations and the conclusions of this paper are given in Section 4 and Section 5, respectively

2. PROBLEM FORMULATION

In this section the signal model is described and the MUSIC-based TDOA estimation problem is formulated.

2.1. Signal Model

Considering a planar and circular array with M microphones in a 2D geometry as shown in Fig 1. The received time domain signals

are denoted by $y_m(t), m = 1, 2, ..., M$ in the noisy and reverberant environments. The signal at microphone m is modeled as

$$y_m(t) = \sum_{r=1}^{R} h_{m,r} * s_r(t) + v_m(t) \tag{1}$$

Where $h_{m,r}$ represents the relative transfer function (RTF) associated with the r-th source $\mathbf{s_r}(t), r = 1, 2, ..., R(R < M)$ from the reference microphone to microphone $m.v_m(t) = \delta_{\mathbf{m}}(t) + \eta_{\mathbf{m}}(t)$ is modeled as an uncorrelated noise where $\delta_{\mathbf{m}}(t)$ and $\eta_{\mathbf{m}}(t)$ denote the late reverberation and additive background noise signal at the m-th microphone, respectively.

Eq.(1) can be transformed into frequency domain to obtain

$$\mathbf{Y}(t, f) = \mathbf{H}(f)\mathbf{S}(t, f) + \mathbf{V}(t, f) \tag{2}$$

where

$$\mathbf{Y}(t, f)=[Y_1(t, f), ..., Y_M(t, f)]^T, \tag{3}$$

$$\mathbf{H}(f)=[H_{1,r}(f), ..., H_{M,r}(f)]^T, \tag{4}$$

$$\mathbf{V}(t, f)= [V_1(t, f), ..., V_M(t, f)]^T \tag{5}$$

and $Y_M(t, f)$, $H_{M,r}(f)$ and $V_M(t, f)$are the short time Fourier –transformer (STFT), respectively. (t, f) represents time frequency index of signal in shift and the superscript T denotes nonconjugate transposition. The RTF $\mathbf{H}(f)$ takes the flowing form

$$\mathbf{H}(f) = \exp(-j\frac{2\pi f}{N} f_s \tau_{m,r})] \tag{6}$$

where $j = \sqrt{-1}$, N is the number of STFT frequencies, f_s is the sampling rate in Hz and $\tau_{m,r}$ denotes the TDOA of the r-th source between microphone m and the reference microphone. The TODA is given by

$$\tau_{m,r} = \frac{d_m \cos(\theta_r)}{c} \tag{7}$$

where d_m denotes the distance between microphone m and the reference microphone, θ_r is the angle of arrival of the r-th source and c is the sound velocity.

2.2. MUSIC-Based TDOA Estimation

The SCM based on the multi-channel signals shown in eq.(2) is defined as

$$\mathbf{R}_y(t, f) = E[\mathbf{Y}(t, f)\mathbf{Y}^{\mathbf{H}}(t, f)] = \mathbf{R}_s + \mathbf{R}_\delta + \mathbf{R}_\eta \tag{8}$$

where $E[\cdot]$ is the expectation operation. $\mathbf{R}_s$, $\mathbf{R}_\delta$ and $\mathbf{R}_\eta$ are corresponding to the SCMs of point sources, late reverberation and noise.

Eigenvalue decomposition is applied to the SCM $\mathbf{R}_s(t, f)$ which needs to be estimated. Given R source signals are considered in the paper, according to the order of eigenvalues, the eigenvectors corresponding to the largest R eigenvalues are obtained to compose the signal subspace as $U_s(t, f) = [u_1, \cdots, u_R]$. The rest, $M - R$ eigenvalues and their corresponding eigenvectors, as the noise subspace $U_n(t, f) = [u_{R+1}, \cdots, u_M]$. Then the pseudo spatial spectrum is defined as

$$P(t, f, \theta) = \frac{1}{a^H(\theta)U_n(t, f)fU_n^H(t, f)a(\theta)} \tag{9}$$

where θ is the arrival angle and $a(\theta)$ is the corresponding steering vector perpendicular to the noise subspace $U_n(t, f)$. In practice, the denominator of eq.(9) will not be zero, because noise exists and the signal is discretely sampled. The estimation of DOA is obtained by searching the peak from the summed pseudo spectrum which can overcome the spatial aliasing ambiguity occurring at high frequencies, i.e.,

$$P(t, \theta) = \sum_f P(t, f, \theta) \tag{10}$$

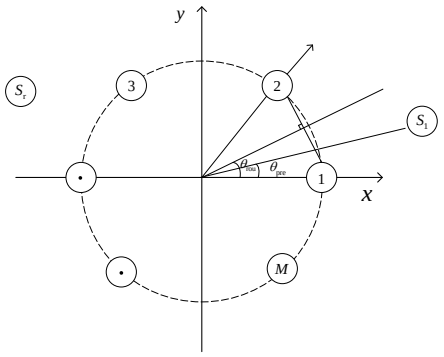

Figure 1: Diagram of uniform circular array.

3. PROPOSED APPROACH

Under the assumption of noise-free/reverberation-free environments and the single signal point source, the TDOA of the target signal can be accurately estimated based on the MUSIC algorithm. However, the presence of competing interferences and reverberation introduces a secondary peak in SCM. Furthermore, diffused background noise may flatten the peak, causing high pseudo spectrum values to span over TDOA intervals, which map to connected regions instead of point locations. Therefore, the reliable TF bins that carry relatively clean for TDOA estimation are extracted to alleviate this problem. This is realized by implementing a mask predictor based on the phase and LSTM to estimate the speech mask. In addition, the MUSIC algorithm has high computational cost when searching the precise peak from 0° to 359° with a 1° step. Hence, the proposed approach increases the search step size to estimate roughly and further estimates precise angle based on the above phase information and partial differential mathematical principle in the subspace that is likely to contain the target source.

3.1. Mask Prediction

DNN-based TF mask has dramatically advanced monaural speech separation [8]. This paper firstly trains the mask predictor which is capable of accurately determining the speech dominance at each TF bin only using the reference microphone channel speech data, estimates masks for all channels using the same predictor. Among various types of neural networks, the LSTM [7] has been shown to generate consistently better separation results and is thus employed in this paper. Depending on using the speech signal as the target to define the IRM

$$W_{\mathrm{IRM}}(t, f) = \frac{|H(f)S(t, f)|^2}{|H(f)S(t, f)|^2 + |V(t, f)|^2}. \tag{11}$$

Although the LSTM-based $W_{\mathrm{IRM}}(t, f)$ mask predictor can selected reliable TF bins and distinguish speech signal from late reverberation and background noise, the presence of competing inter-

ferences will introduce multiple peaks. In this section, we propose a method to combat the competing interferences by multiplying the phase-based mask predator with $W_{\mathrm{IRM}}(t,f)$.

For a given sound source, the adjacent pair of microphones is denoted as $\{m_i, m_{i+1}\}, i = 1, 2, ..., M-1$ and the last pair of microphone pairs is represented as $\{m_M, m_1\}$. The plane of the circular array is divided into M uniform sectors whose serial numbers are denoted as g=1 . . . M, and the preset phase of each sector can be calculated by the following formula [9]

$$G^{(t,f)}_{m_i \to m_1}(\phi_g) \triangleq e^{-j\omega\tau_{m_i \to m_1}(\phi_g)} \tag{12}$$

where $\tau_{m_i \to m_1}(\phi_g)$=$\tau_{m_1 m_2}(\phi_g)$−$\tau_{\mathrm{m_i m_{i+1}}}(\phi_g)$ is the relative delay between the signals received at the microphone pair $\{m_1, m_2\}$ and $\{m_i, m_{i+1}\}$. ϕ_g is the angle between the centerline of each uniform sector and the positive direction of the x-axis.

The phase of the cross-spectrum signal received at adjacent two microphones is calculated as

$$G_{m_i m_{i+1}}(t,f) = \frac{Y_i(t,f)Y^*_{i+1}(t,f)}{\left|Y_i(t,f)Y^*_{i+1}(t,f)\right|}. \tag{13}$$

The preset angle phase and the observed signal phase are provided by (12) and (13), respectively. Therefore the TF mask index can be expressed as

$$I^{(t,f)}_{\mathrm{phase}} \triangleq \arg\min \sum_{i=1}^{m} \left\| G_{m_i m_{i+1}}(t,f) - G^{(t,f)}_{m_i \to m_1}(\phi_g) \right\|_2 \tag{14}$$

Assuming the target source comes from the g-th sector, therefore the phase TF mask is

$$W_{\mathrm{phase}}(t,f) = \begin{cases} 1, & I_{\mathrm{phase}}(t,f) = g \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

The final TF mask on the LSTM $W_{\mathrm{IRM}}(t,f)$ and by the phase information $W_{\mathrm{phase}}(t,f)$, is defined as

$$W(t,f) = W_{\mathrm{IRM}}(t,f) \times W_{\mathrm{Phase}}(t,f) \tag{16}$$

According to the MUSIC algorithm formula derivation, the performance of TDOA estimation depends on the estimated SCM how close to the real SCM which only includes the target signal. The relatively clean estimated SCM in this paper, which contains more target signal spatial information while minimizes the effect of competing interferences, late reverberation and background noise, is denoted as

$$\hat{R}(t,f) = E[W(t,f)Y(t,f)Y^H(t,f)W^{\mathrm{H}}(t,f)]. \tag{17}$$

3.2. Low Complexity TDOA Estimation

Although the MUSIC approach is robust and reliable, it is computationally expensive as it requires a fine discretization of the space to achieve a good localization precision. In this section the paper searches the peak at larger interval to localize the target signal source in the subspace. Then the TDOA estimation difference $\tau(\theta_{\mathrm{diff}})$ between the precise source TDOA (θ_{pre}) and the rough source TDOA $\tau(\theta_{\mathrm{rou}})$, as $\tau(\theta_{\mathrm{diff}}) = \tau(\theta_{\mathrm{rou}}) - \tau(\theta_{\mathrm{pre}})$ using the partial differential mathematical strategy is estimated to preserve high solution of algorithm

$$\begin{aligned} \sum_{\omega} \omega G^{(t,f)}_{m_i m_{i+1}} [G^{(t,f)}_{m_i \to m_1}]^* &= \sum_{w} \omega e^{-j\omega\tau(\theta_{pre})} e^{j\omega\tau(\theta_{rou})} \\ &= \sum_{w} \omega e^{j\omega\tau(\theta_{diff})} \end{aligned} \tag{18}$$

According to Euler's formula, (18) is equal to

$$\sum_{\omega} \omega e^{j\omega\tau(\theta_{\mathrm{diff}})} = \sum_{\omega} \omega(\cos(\omega\tau(\theta_{\mathrm{diff}})) + j\sin(\omega\tau(\theta_{\mathrm{diff}}))) \tag{19}$$

The TDOA estimation difference $\tau(\theta_{\mathrm{diff}})$ can be estimated as

$$\tau(\theta_{\mathrm{diff}}) = \frac{imag[\sum_{\omega} \omega e^{j\omega\tau(\theta_{\mathrm{diff}})}]}{\sum_{\omega} \omega^2} \approx \frac{\sum_{\omega} \omega^2 \tau(\theta_{\mathrm{diff}})}{\sum_{\omega} \omega^2} \tag{20}$$

where $\sin(\omega\tau(\theta_{\mathrm{diff}})) \approx \omega\tau(\theta_{\mathrm{diff}})$ when $\omega\tau(\theta_{diff})$ is small.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

In this section, the experiments of MUSIC-based TDOA estimation algorithm are conducted on simulated and real data. The 6-channel uniform circular microphone array with a radius of 4.5cm is utilized. The array is placed at the center of the room with a height of 1m. To create the simulation training data, the 7000 clean speech signals from the CHiME-3 [10] are convolved with the room impulse responses generated with different settings using the image method [11]. based on the given circular array geometry. The simulated room size is fixed at 7 m×5 m×3 m. T60 ranges from 0.0 s to 1.0 s with 0.1 s step size. Then the randomly selected noises from the CHiME-3 [10] database are added to each utterance at the SNR levels randomly chosen from 0dB to 20dB. LSTM network settings are the same as [7].

In the test of simulation data, the clean utterances from the CHiME-3 [10] are convolved with the designed 24 scenarios (3 room sizes × 2 distances × 4 SNRs). The three room sizes are respectively small (6 m×6 m×3 m), medium (10 m×10 m×3 m) and large (14 m×14 m×3 m). Each room has two distances that near distances are 1m and far distances are 1 m and far distances are 1.5 m, 3 m and 5 m respectively. The non-stationary diffused noise is added to the clean utterances and the input SNRs are -10dB, 0dB, 10dB and 20dB, respectively. To simulate competing-Speaker scenarios, the target and the interference sources speak from 0° and 180°, respectively.

In the test of real data, the room size is fixed at 7 m×5 m×3 m. Some utterances are recorded to generate 27 scenarios using different settings (3 distances × 3 angles × 2 interferences). The target source at 0°, 1 m, 2 m and 3 m from the center of the array, respectively. Interference sources which include single human interference source (60°, 120° and 180°) single music point source (60°, 120° and 180°) and dual human interference sources (60°-300°, 120°-240° and 160°-200°) stay 2 m from the center of the array.

The sampling rate for all speech signals and noise is 16kHz and the frame size is set to 512. Hamming window and 75% overlap between adjacent frames are applied. The proposed MUSIC-based TDOA estimation method is compared with the state-of-the-art TDOA estimation algorithms consisting of GCC-PHAT[3], Music [5] and LSTM-based WMUSIC [7] in the competing Speaker scenarios with background noise and reverberation. The root mean square error (RMSE) which calculates the difference between the real matrix and corresponding estimation is used as evaluation metric

SNR (dB)	Method	RMSE(×0.0001)					
		T60=0.3s		T60=0.6s		T60=0.9s	
		1m	1.5m	1m	3m	1m	5m
-10	GCC-PHAT	2.29	2.36	2.36	2.11	2.63	2.19
	MUSIC	3.70	3.70	3.31	3.70	3.51	3.51
	WMUSIC	3.70	4.03	3.42	3.42	3.49	3.68
	Proposed	**0.05**	**0.07**	**1.18**	**0.06**	**1.17**	**0.09**
0	GCC-PHAT	2.86	2.62	2.87	2.87	2.73	2.92
	MUSIC	3.70	3.70	3.27	3.63	3.43	3.51
	WMUSIC	4.05	4.05	3.86	4.16	3.52	2.87
	Proposed	**0.03**	**1.17**	**0.06**	**0.03**	**0.05**	**0.04**
10	GCC-PHAT	3.24	3.09	3.10	2.76	3.16	3.15
	MUSIC	3.52	3.89	3.71	3.71	3.52	3.71
	WMUSIC	3.71	3.71	3.52	4.21	3.70	3.89
	Proposed	**1.16**	**0.01**	**0.03**	**1.15**	**1.17**	**0.02**
20	GCC-PHAT	3.37	3.05	3.16	3.45	3.32	2.90
	MUSIC	3.52	3.71	3.32	3.52	3.71	3.52
	WMUSIC	3.32	3.71	3.32	3.71	3.52	3.52
	Proposed	**1.66**	**1.66**	**1.17**	**0.04**	**1.66**	**1.66**

Table 1: The TDOA estimation results on the simulated data.

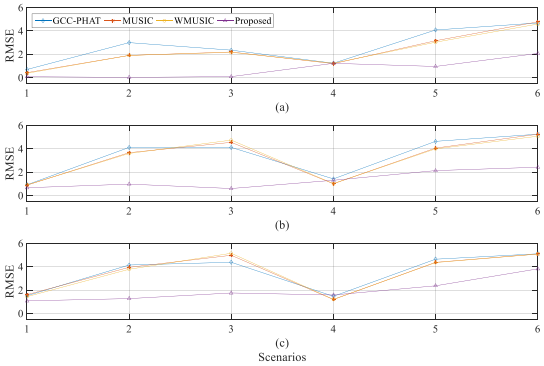

Figure 2: The TDOA estimation results on the real data.

4.2. Results and Comparison on Simulated Data

The comparative results of TDOA estimation in several simulated scenarios are shown in Table 1. It can be seen that the performances of GCC-PHAT, MUSIC and WMUSIC approaches are seriously affected by competing interferences. In addition, the non-stationary noise further degrades the TDOA estimation of GCC-PHAT method to cause inaccurate results that neither belong to target source nor to interference sources while MUSIC and WMUSIC method are more robust to background noise. In contrast, the performance of proposed method has been improved a lot by selecting reliable TF bins through a mask predator based on the phase and LSTM

4.3. Results and Comparison on Real Data

The RMSE results of comparative TDOA estimation algorithms in the real environment are illustrated in Figure 2. Horizontal coordinate 1-3 expresses the single human interference source in 60°, 120° and 180° respectively, 4-6 expresses the single music interference source in 60°, 120° and 180° respectively and 7-9 express the dual human interference sources in 60°-300°, 120°-240° and 160°-200° respectively. Sub figure (a), (b) and (c) present that target signal source are 1m, 2m and 3m from the array, respectively. There is little stationary background noise in the recording. Therefore, the performance of TDOA estimation mainly is affected by competing interferences. As shown in the interval of Figure 2 horizontal coordinate 1-3, the proposed algorithm has significantly improved performance compared to GCC-PHAT, MUSIC and WMUSIC algorithms. However, the advantage of the proposed algorithm is limited when the competing interference is music point source. This was due to the fact that music signal runs through the overall spectrum, which results in inaccurate mask predator based on the phase because of high frequency aliasing.

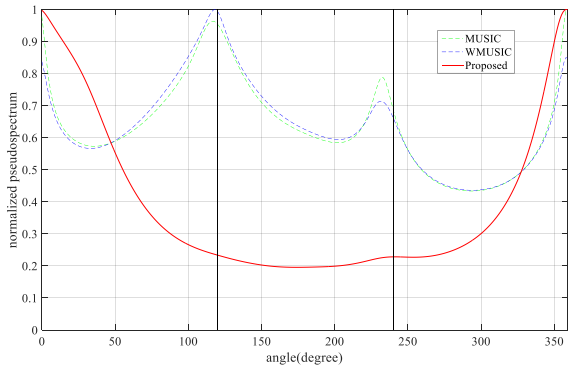

Figure 3: The pseudo spectrum of MUSIC based methods. True doa is at 0° while competing interference sources are at 120° and 240° marked with vertical line

An example of the summed pseudo spectrum in real environment is shown in Fig 3. The proposed method estimates the arrival angle exactly same as true angle ($= 0^\circ$) since the highest peak is there. Competing interference sources in 120° and 240° introduce two peaks to cause wrong estimation, although MUSIC and WMUSIC methods have a peak at 0°. The proposed method overcomes the problem by filtering competing interference, noise and reverberation. In addition, the proposed method is approximately the numbers of searching subspace degrees times faster than classical MUSIC methods in Matlab 2015a due to reducing search space while partial differential mathematical operation requires little computation.

5. CONCLUSIONS

This paper proposed a MUSIC-based TDOA estimation method in challenging conditions. Through experimental evaluations in simulated and real data, the robustness of the method to competing interferences, background noise and reverberation was shown. In addition, the proposed method has a faster computing speed in simulation software. Future works involve testing the proposed approach with different noise types, overcoming wrong masking predator based on the phase information when high frequency aliasing and achieving online computation in embedded device.

6. ACKNOWLEDGMENT

This work is supported by the research project of Chongqing Educational Commission (KJ130504), and the research project of Chongqing science and technology commission (cstc2015jcyjA40027).

7. REFERENCES

[1] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," The Journal of the Acoustical Society of America, vol. 107, no. 1, pp. 384–391, 2000.

[2] Y. Guo, X. Wang, C. Wu, Q. Fu, N. Ma, and G. J. Brown, "A robust dual-microphone speech source localization algorithm for reverberant environments." in INTERSPEECH, 2016, pp. 3354–3358.

[3] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE transactions on acoustics, speech, and signal processing, vol. 24, no. 4, pp. 320–327, 1976.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in Microphone Arrays. Springer, 2001, pp. 157–180.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE transactions on antennas and propagation, vol. 34, no. 3, pp. 276–280, 1986.

[6] O. Schwartz, Y. Dorfan, E. A. Habets, and S. Gannot, "Multi-speaker doa estimation in reverberation conditions using expectation-maximization," in 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2016, pp. 1–5.

[7] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, and H. Li, "Weighted spatial covariance matrix estimation for music based tdoa estimation of speech source." in INTERSPEECH, 2017, pp. 1894–1898

[8] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 178–188, 2019.

[9] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Realtime multiple sound source localization and counting using a circular microphone array," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2193– 2206, 2013.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 504–511.

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