Scientia Iranica D (2017) 24(6), 3268–3279

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
www.scientiairanica.com

Time difference of arrival estimation of sound source using cross correlation and modified maximum likelihood weighting function

M.S. Hosseini[a,*], A.H. Rezaie[a] and Y. Zanjireh[b]

a. *Automation and Intelligent Monitoring Systems Lab (AIMS Lab), Department of Electrical Engineering, Amirkabir University of Technology, Tehran, P.O. Box 15875-4413, Iran.*
b. *Signal Processing Laboratory (SiPLABoratory), University of Algarve, Portugal.*

Received 7 April 2016; received in revised form 25 June 2016; accepted 25 July 2016

KEYWORDS
Generalized Cross Correlation (GCC); Time Difference Of Arrival (TDOA); Sound Source Localization (SSL); Maximum Likelihood (ML).

Abstract.
The Generalized Cross Correlation (GCC) framework is one of the most widely used methods for Time Difference Of Arrival (TDOA) estimation and Sound Source Localization (SSL). TDOA estimation using cross correlation without any pre-filtering of the received signals has a large number of errors in real environments. Thus, several filters (weighting functions) have been proposed in the literature to improve the performance of TDOA estimation. These functions aim to mitigate TDOA estimation error in noisy and reverberant environments. Most of these methods consider the noise or reverberation, and as one of them increases, TDOA estimation error increases. In this paper, we propose a new weighting function. This function is a combined and modified version of Maximum Likelihood (ML) and PHAT-$\rho\gamma$ functions. We named our proposed function as Modified Maximum Likelihood with Coherence (MMLC). This function has merits of both ML and PHAT-$\rho\gamma$ functions and can work properly in both noisy and reverberant environments. We evaluate our proposed weighting function using real and synthesized datasets. Simulation results show that our proposed filter has better performance in terms of TDOA estimation error and anomalous estimations.

1. Introduction

Sound Source Localization (SSL) has many applications in military and civilian areas such as mixed audio signals separation, robotics, video conferencing, speech enhancement, tracking of acoustic sources, underwater acoustics, and advanced hearing aids [1-7]. Algorithms for localization of an acoustic source are divided into three main categories:

1. Beamforming;
2. High-resolution spectral estimation;
3. Time Difference Of Arrival (TDOA) [8].

In beamforming approaches, a beam pattern is steered; then, the power of this steered response is calculated for candidate space points. The angle at which the power reaches its maximum value is the Direction Of Arrival (DOA) of the sound source. These algorithms have good stability in direction estimation, but their computational costs are very high. Delay-and-Sum-Beamformer (DSB) is the simplest algorithm in this framework. A good review of beamforming

*. *Corresponding author.*
E-mail address: s.hoseini@aut.ac.ir (M.S. Hosseini)

doi: 10.24200/sci.2017.4355

approaches for localization of sound sources is discussed in [9].

High-resolution spectral estimation methods, which are famous in subspace approaches, use modern spatial-filtering methods and are used in narrowband and far-field signal processing. In speaker localization, these methods deal with constraints that limit their effectiveness. These algorithms are significantly less stable than the beamforming approaches due to source and microphones modeling errors. These errors are due to non-ideality in signal propagation, nonlinear properties of microphones, and variations of source position. Like beamforming algorithms, these approaches are based on spatial search and also have high computational cost [10].

TDOA-based approaches rely on relative time difference between pairs of microphones. These algorithms are divided into two main groups; the first group is based on a pair of microphones, such as Cross Correlation (CC) and Adaptive Eigenvalue Decomposition (AED), and the second group is based on an array of microphones such as Multichannel Cross Correlation Coefficient (MCCC), adaptive blind multichannel identification, and multichannel spatial prediction and interpolation [11]. Thanks to the advances in electronics and the development of new algorithms, utilizing the large microphone arrays for SSL is more simple than before, but two microphone-based approaches are still used in advanced hearing aids, humanoid robots, and human hearing system simulation.

Although beamforming and subspace methods can estimate the location of sound sources with higher accuracy and resolution due to low computational complexity and simplicity in implementation, TDOA-based approaches attract more attention than the other algorithms do. Among them, algorithms based on cross correlation are the most popular frameworks [12]. Usually, to improve the performance of the CC approach, each signal is pre-filtered and cross correlation operation is applied to them. This framework is called Generalized Cross Correlation (GCC). There are many different filters proposed in the literature, and each method has its own features. Figure 1 shows the simple block diagram of the GCC algorithm [13].

In this paper, we propose a new weighting function for the GCC framework. This function is a modified and combined version of Modified Maximum Likelihood (MML) [14] and PHAT-$\rho\gamma$ [15] functions. Our proposed function has the merits of these two functions and can properly work in both noisy and reverberant environments. This paper is organized as follows. In Section 2, we briefly introduce the GCC method and different weighting functions. In Section 3, we propose a new weighting function based on two recently proposed functions. In Section 4, we evaluate our proposed function using synthetic and real-world data. Finally, in Section 5, we conclude the paper.

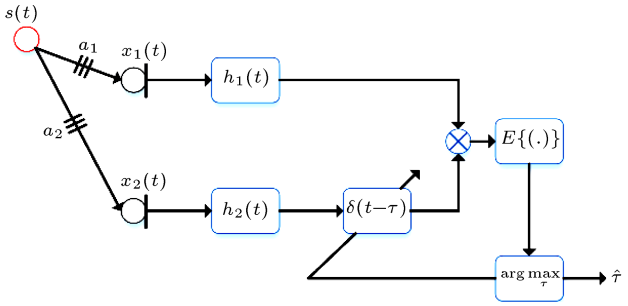

Figure 1. Simple block diagram of GCC framework.

2. TDOA estimation using GCC

The main idea of the cross correlation method is based on single-path propagation of the acoustic plane wave model. Figure 2 shows a simple illustration of this model. In this model, received signals are the delayed and attenuated version of the original acoustic signal, which is emanated from a point source and corrupted by additive white Gaussian noise. This noise is assumed to be uncorrelated with the source signal. Based on this model, the received signals in the microphones are as follows:

$$x_i(t) = \alpha_i s(t - T - \tau_i) + n_i(t), \qquad i = 1, 2, \tag{1}$$

where $s(t)$ is the reference acoustic signal, $x_i(t)$ is the received signal in the ith microphone, α_i is the attenuation factor due to signal propagation ($0 < \alpha_i < 1$), T is the propagation delay between source and the first microphone that captures the signal, $\tau = \tau_2 - \tau_1$ is the relative time delay between two microphones, and $n_i(t)$ is the additive noise of the ith microphone.

Using this model, the optimal time delay estimation can be done using the GCC method as follows (Figure 1) [16]:

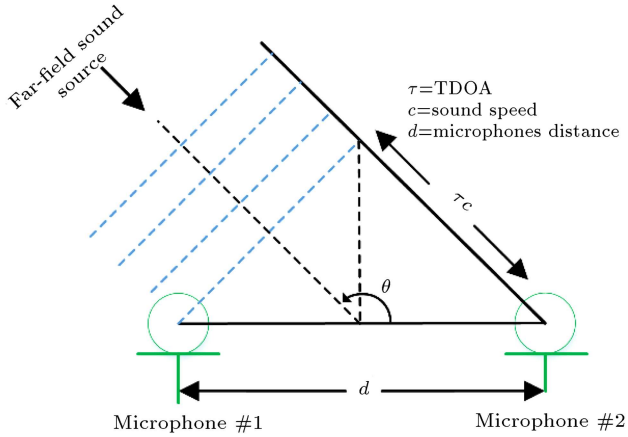

Figure 2. Illustration of DOA of far-field sound source in 2D space.

$$R^g_{x_1x_2}(\tau) = E\{(x_1(t)^* h_1(t)).(x_2(t)^* h_2(t-\tau))\},$$

$$\tau' = \arg\max_{\tau} \left\{ R^g_{x_1x_2}(\tau) \right\}, \quad (2)$$

where $E\{.\}$ is the statistical average (expected value) over time, τ' is an estimation of τ, $h_1(t)$ and $h_2(t)$ are the filters used to improve the estimation accuracy of τ', and * denotes convolution operation. In the frequency domain, GCC, cross Power Spectral Density (PSD), $\varphi_{x_1x_2}(f)$ are related to each other by Eq. (3):

$$R^g_{x_1x_2}(\tau) = \int_{-\infty}^{+\infty} H_1(f)H_2^*(f)\varphi_{x_1x_2}(f)e^{j2\pi f\tau}df. \quad (3)$$

In the GCC framework, the product of $h_1(t)$ and $h_2(t)$ filters in the frequency domain, $H_1(f)H_2^*(f)$, is named generalized frequency weighting function ($\psi_g(f)$). In the next section, we provide a brief review about the most famous weighting functions.

2.1. Weighting functions

Before explaining our proposed method, we are going to summarize the most important weighting function for GCC framework.

2.1.1. Classical cross correlation

The simplest and fastest weighting function is the Classic Cross Correlation (CCC). In this function, $\psi_g(f)$ is equal to 1. But, in noisy and reverberant environments, TDOA estimation error is too high using this function. Thus, researchers develop different weighting functions to mitigate TDOA estimation error in real environments.

2.1.2. Hannan and Tahomson

In 1973, Hannan and Thomson introduced HT (Maximum Likelihood) weighting function [17]. This function has greater weight where the coherence between the two received signals is high. Coherence function is a real valued function which measures correlation between two signals ($0 < \gamma^2_{x_1x_2} < 1$). As correlation between two signals increases, this function tends towards 1; as the correlation between two signals decreases, this function tends to zero. Coherence between signals $x_1(t)$ and $x_2(t)$ in the frequency domain is defined as in Eq. (4):

$$\gamma^2_{x_1x_2}(f) = \frac{|\varphi_{x_1x_2}(f)|^2}{\varphi_{x_1x_1}(f)\varphi_{x_2x_2}(f)}, \quad (4)$$

where $\varphi_{x_ix_j}(f)$ is defined as Eq. (5):

$$\varphi_{x_ix_j}(f) = E\{X_i(f)X_j^*(f)\}. \quad (5)$$

Under low SNR conditions, the HT weighting function is equal to CCC, but under usual conditions (high SNR and low reverberation), it is shown that this weighting function is the maximum likelihood estimator for time delay estimation in the CC framework:

$$\psi^{\mathrm{ML}}_g(f) = \frac{1}{|\varphi_{x_1x_2}(f)|} \frac{\gamma^2_{x_1x_2}(f)}{1-\gamma^2_{x_1x_2}(f)}. \quad (6)$$

2.1.3. PHAT

In 1973, Carter et al. proposed PHAT or the Cross-power Spectrum Phase (CSP) function [18]. It was an intuitive solution for time delay estimation which extracts time delay information from the cross spectrum phase. This weighting function has good results in high SNR and moderate (or high) reverberant conditions, but if SNR or energy of the received signals is low, $|\varphi_{x_1x_2}(f)|$ tends to zero and TDOA estimation error increases:

$$\psi^{\mathrm{PHAT}}_g(f) = \frac{1}{|\varphi_{x_1x_2}(f)|^2}. \quad (7)$$

2.1.4. PHAT-$\rho\gamma$

In 2010, PHAT-$\rho\gamma$ function was developed [15,19]. As mentioned in [20], most of the acoustical noises in untreated enclosure are at frequencies below 200 Hz. So, Rabinkin proposed the ρ-CSP function which uses ρ tuning parameter in the power of CSP function to discard the non-speech portion of CSP (frequencies below 200 Hz). The value of ρ is determined by room acoustical characteristics, but as mentioned by Rabinkin, the optimal value for ρ in different enclosures is about 0.75. In addition to ρ-CSP, in PHAT-$\rho\gamma$ function, the minimum of coherence is added to the weighting function for further error reduction due to low-energy signals:

$$\psi^{\mathrm{PHAT}-\rho\gamma}_g(f) = \frac{1}{|\varphi_{x_1x_2}(f)|^\rho + \min(\gamma^2_{x_1x_2}(f))}. \quad (8)$$

3. The proposed weighting function

Most of the proposed weighting functions in the literature have been designed under usual conditions (high SNR and low reverberation) or just by considering one of the difficulties in real environments. Thus, their performance is degraded in adverse environments. In this paper, we propose the new Modified Maximum Likelihood with Coherence (MMLC) weighting function. It aims to achieve accurate results in noisy and reverberant environments.

As mentioned in [16], the ML weighting function can be written as a function of phase variance:

$$\psi^{\mathrm{ML}}_g(f) \approx \frac{1}{|X_1(f)X_2^*(f)|\mathrm{var}[\theta(f)]}, \quad (9)$$

where $X_1(f)$ and $X_2^*(f)$ are discrete Fourier transforms of $x_1(t)$ and $x_2(t)$, respectively. $\mathrm{var}[\theta(f)]$ is the variance of the cross spectrum phase, and $\theta(f)$ is defined as:

$$X_1(f)X_2^*(f) = A(f)e^{j\theta(f)}. \quad (10)$$

For the ML weighting function, approximation of $\text{var}[\theta(f)]$ is as:

$$\text{var}[\theta(f)]^{\text{ML}} = \frac{1 - |\gamma^2_{x_1x_2}(f)|}{|\gamma^2_{x_1x_2}(f)|} = \mu_{x_1x_2}(f). \tag{11}$$

Then, Maximum Likelihood (ML) weighting function is as:

$$\psi_g^{\text{ML}}(f) \approx \frac{1}{|X_1(f)X_2^*(f)|} \frac{|\gamma^2_{x_1x_2}(f)|}{1 - |\gamma^2_{x_1x_2}(f)|}. \tag{12}$$

3.1. Phase variance estimation

By using the joint complex Gaussian model in the frequency domain for the received signals, we can write Eq. (13) [14] (argument f is omitted for simplicity):

$$p(\mathbf{X}|\boldsymbol{\phi}) = \frac{1}{\boldsymbol{\pi}^2\boldsymbol{\phi}} e^{(-X^H\boldsymbol{\phi}^{-1}X)}, \tag{13}$$

where matrix $\boldsymbol{\phi}$ is the cross covariance of $x_1(t)$ and $x_2(t)$ signals, $\mathbf{X} = [X_1, X_2]^T$, and $\mathbf{X}^H$ is the Hermitian Transpose of $\mathbf{X}$:

$$\boldsymbol{\phi} = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{12}^* & \varphi_{22} \end{bmatrix},$$

and φ_{ij} is computed as Eq. (5).

If we assume that $s(t)$, $n_1(t)$, and $n_2(t)$ are correlated with each other and the signals of Eq. (12) in polar form are written, we reach Eq. (14):

$$p(X_1, X_2, \theta) = \frac{|X_1X_2|}{\pi^2|\boldsymbol{\phi}|} \exp \left\{ \frac{2|X_1X_2|\text{Re}\{\varphi_{12}^*\} - \varphi_{22}|X_1|^2 - \varphi_{11}|X_2|^2}{|\boldsymbol{\phi}|} \right\}, \tag{14}$$

where $[X_1, X_2] = [|X_1|e^{j\theta_1}, |X_2|e^{j\theta_2}]$ and $\theta = \theta_1 - \theta_2$. Using marginalization and changing of the variables, we can write:

$$p_\gamma(\theta) = \frac{1 - \gamma^2_{x_1x_2}}{2\pi[1 - \gamma^2_{x_1x_2}\cos^2\theta]} \left[1 + \frac{\arccos(\gamma_{x_1x_2})\gamma_{x_1x_2}\cos\theta}{\sqrt{1 - \gamma^2_{x_1x_2}\cos^2\theta}} \right]. \tag{15}$$

Then, we use Eq. (16) to obtain $\text{var}[\theta]$:

$$\text{var}[\theta] = \int_{-\pi}^{+\pi} \theta^2 p_\gamma(\theta) d\theta. \tag{16}$$

We use MATLAB® [21] based simulation for finding $\text{var}[\theta(f)]$. Figure 3 shows this simulation result. As we can see, the approximation proposed by [14] is better than the ML approximation. This approximation is as in Eq. (17):

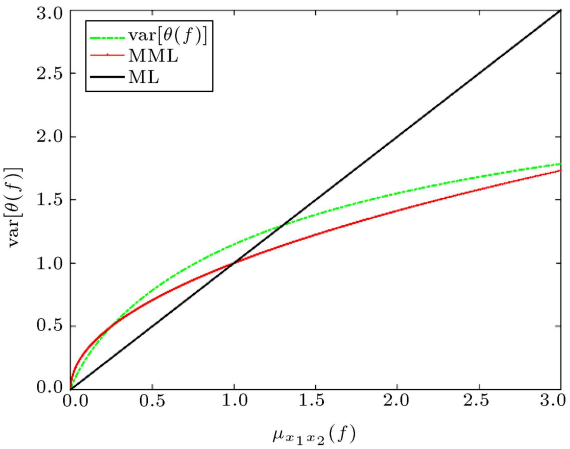

Figure 3. True value of $\text{var}[\theta(f)]$ and its approximations in ML and MML weighting functions.

$$\text{var}[\theta]^{\text{MML}} = \sqrt{\frac{|1 - \gamma^2_{x_1x_2}|}{|\gamma^2_{x_1x_2}|}} = \sqrt{\mu_{x_1x_2}}. \tag{17}$$

Then, Modified Maximum Likelihood (MML) weighting is as follows:

$$\psi_g^{\text{MML}} \approx \frac{1}{|X_1X_2^*|} \sqrt{\frac{|\gamma^2_{x_1x_2}|}{|1 - \gamma^2_{x_1x_2}|}}. \tag{18}$$

But, as mentioned before, as the energy of the signal decreases, $|\varphi_{x_1x_2}(f)|$ tends to zero. To solve this problem, we used the solution proposed by Liu and Shen [15]. Liu proposed using the minimum value of coherence in the denominator of weighting function. This causes better results, because in the situations in which the microphones capture low energy signals, the dominator of the weighting function tends to $\min(|\gamma^2_{x_1x_2}(f)|)$ instead of zero.

Also, for suppressing the non-speech portion of CSP, tuning parameter ρ is used as a power of $|\varphi_{x_1x_2}(f)|$. Then, the frequencies below 200 Hz of the CSP are discarded. By using these two modifications on the MML function, we proposed the Modified Maximum Likelihood with Coherence (MMLC) function:

$$\psi_g^{\text{MMLC}} \approx \frac{1}{|\varphi_{x_1x_2}|^\rho + \min(|\gamma^2_{x_1x_2}|)} \sqrt{\frac{|\gamma^2_{x_1x_2}|}{|1 - \gamma^2_{x_1x_2}|}}, \tag{19}$$

where we set ρ tuning parameter to 0.75.

4. Simulation

To evaluate the proposed weighting function, we used two different simulations using MATLAB®. The first simulation is based on synthetic data and Monte-Carlo simulation, and the second one is based on real-world data. For both simulations, we used the block diagram proposed in [22]. As seen in Figure 4, at

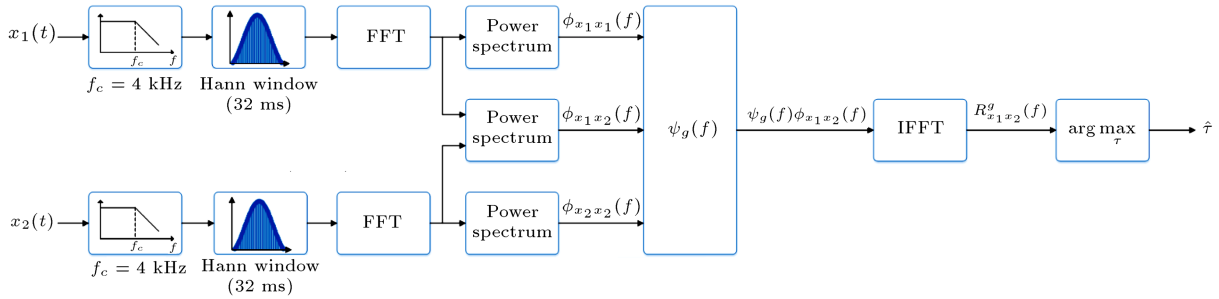

Figure 4. Block diagram of simulation method [22].

first, each received signal is filtered by a low-pass filter with cutoff frequency of 4 kHz (speech signals in the dataset have been sampled with 16 kHz), then 32 ms of signals are separated using the Hann window with 25% overlap. The power spectrum and cross power spectrum are computed using 512-point Fast Fourier Transform (FFT). The generalized cross correlation between received signals is computed by Inverse Fast Fourier Transform (IFFT) of function $\psi_g(f)\varphi_{x_1x_2}(f)$. Finally, using an interpolation stage, time delay between the received signals is estimated. In this paper, we choose Cubic Spline interpolation.

4.1. Simulation using synthetic data

In the first step, we evaluate our proposed weighting function using synthetic data and Monte-Carlo simulation. We simulate four different rooms using the image method [23]. Dimensions of these rooms are 8 m×5 m×3.5 m (x, y, z) and reflection coefficients vary from 0 to 1, so that reverberation times are 0.2 s, 0.4 s, 0.6 s, and 0.8 s. In these rooms, the source is located at 45° and distance between microphone pairs and the source is 1.5 m. The distance between microphones is 10 cm. A clean speech file was selected from the SiSEC 2010 dataset [24]. For each file, we add white Gaussian noise with SNR –20 dB to 60 dB by 5dB increasing step, and we conduct a simulation process with 500 iterations for each step. The performance of weighting functions is measured with two metrics: RMSE and Anomaly. Anomaly measures the ratio of the outlier TDOAs to all of the estimated TDOAs [25]:

$$\Gamma(\tau') = \begin{cases} 1, & \text{if } |\tau - \tau'| > \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Anomaly}(\tau') = E\{\Gamma(\tau')\}. \tag{20}$$

We assume that ε equals 1, which means that if the difference between the estimated TDOA and true TDOA is greater than 1 sample, this estimation is assumed to be anomaly.

Figure 5 shows the simulation results for synthetic data. This figure contains two rows and four columns. The first row of this figure shows the RMSE versus SNR and the second row shows the Anomaly versus SNR. Each column of this figure shows RMSE and Anomaly estimations of different reverberation times. As it can be seen from the first row of this figure in –20 dB SNR, CCC and PHAT weighting functions have the highest error rate among the other functions. As SNR increases, RMSE of all methods decreases, but this reduction is greater than the others for MMLC. For example, in 0.2 s reverberation time and SNRs higher than 50 dB, RMSE of the proposed method is 0.34 samples less than MML and PHAT methods. In 0.4 s, 0.6 s, and 0.8 s reverberation times and SNRs higher than 50 dB, RMSE of MMLC is 0.272, 0.175, and 0.277 samples less than the MML function, respectively.

As was mentioned, the second row of Figure 5 shows the percentage of anomaly for TDOA estimation. In –20 dB SNR, the anomaly of MMLC and PHAT-$\rho\gamma$ is less than the others. As SNR increases, our proposed function shows better performance in terms of anomaly; for example, in 0.2 s reverberation time and in SNRs higher than 40 dB, our proposed function has 4.35% less anomalous TDOA estimation than PHAT. In 0.4 s, 0.6 s, and 0.8 s reverberation times and in SNRs higher than 50 dB, our proposed method has 5.27%, 5.17%, and 1.91% less anomalous estimations in comparison to the MML function, respectively.

Simulations based on synthetic data indicate that MMLC and PHAT-$\rho\gamma$ functions have better performance than the other functions in terms of RMSE and anomalous estimations in low SNR conditions, because these two functions have parameter ρ and the minimum value of coherence in their denominators. As mentioned earlier, using parameter ρ in the weighting function discards the non-speech portion of CSP. On the other hand, using the minimum value of the coherence weighting function prevents the denominator from tending to zero in low energy signal conditions. In higher SNRs, the MMLC and MML functions outperform the other functions, since the weighting factor in Eq. (16) has greater weight where the coherence between the two received signals is high. However, since the approximation of var$[\theta]$ in Eq. (17) is a better approximation than Eq. (11), these two functions outperform the ML function.

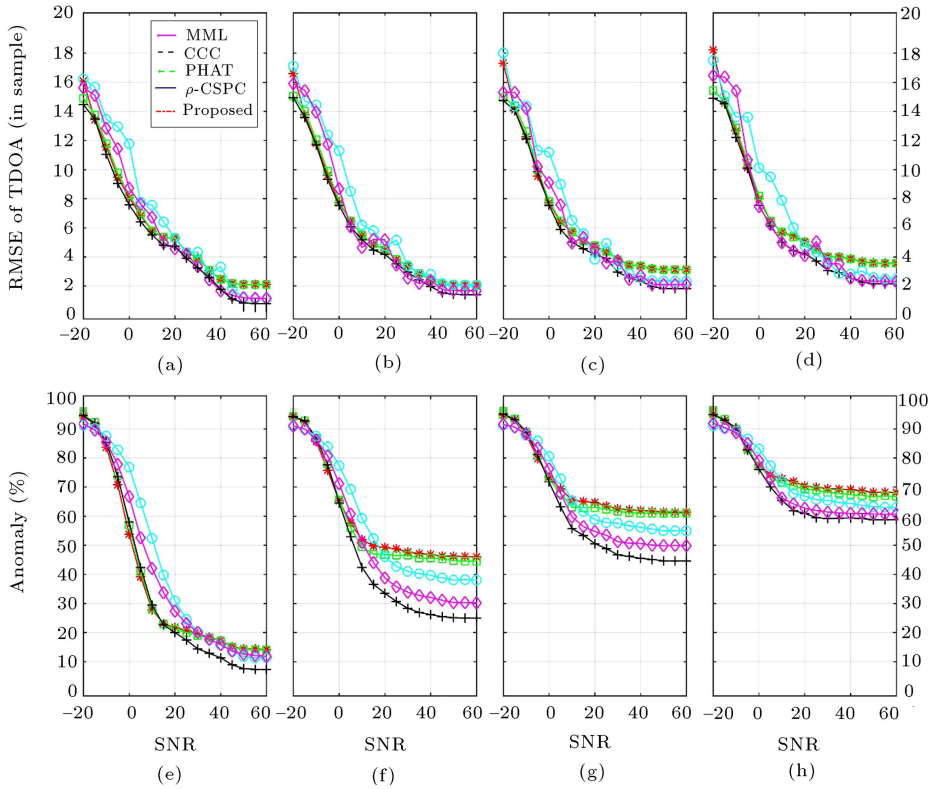

Figure 5. Simulation results of synthetic data using Monte-Carlo method. The RMSE and anomaly of estimated TDOAs for 0.2 s, 0.4 s, 0.6 s, and 0.8 s reverberation times, respectively (a, b), (c, d), (e, f) and (g, h) sub-figures.

4.2. Simulation based on real-world data

In this section, we used the SiSEC 2010 dataset. This dataset contains 30 stereo WAV files. Each file is a mixture of a speech signal with real background noise with a 16 KHz sampling frequency. Noises are recorded at three different environments: Cafeteria, Subway, and Square. Speech signals and recorded noises in the cafeteria and subway are played again in an office room (reverberant room), and then a mixture of them is collected using omnidirectional microphones. In square environment, speech signals and noises are mixed together anechoicly using computer simulation. The distance between the microphones is 8.6 cm (Given the microphones distance of 8.6 cm and sampling frequency of 16 kHz, it can be concluded that maximum resolution of TDOA estimation is 4 samples and final maximum resolution of DOA estimation is 22.5°.). DOA of the main source is different for each file, and the SNR level is randomly chosen between −17dB and +12dB [24].

In this section, we use the simulation method proposed in [22]. For each file in this dataset, we plot the TDOA histogram that shows the probability distribution of estimated TDOAs. In addition, we compute four statistical parameters for better comparison. These parameters are:

1. Mode (most probable TDOA): index of maximum in histogram;
2. Mean: mean value of estimated TDOAs;
3. Mode frequency: relative frequency of mode occurrence. Comparison of Mode and Mean is useful to detect if the distribution is multi-modal;
4. RMSE: RMSE shows the overall accuracy of method.

In this section, we briefly explain some results of the simulations (first sub-figure of each figure). Figure 6(a) and Table 1(a) show a case where the main source is at 143° and cafeteria noise is located at 90°. Figure 6(a) indicates that TDOA estimation using the MMLC function has the highest concentration around TDOA of the main source. This can be seen from the Mode frequency value in Table 1(a). The Mode frequency of MMLC is 18% better than the best weighting function (PHAT). As seen in Table 1(a), the proposed weighting function has the lowest RMSE between the other weighting functions, and this error is 28.5% lower than the best weighting function (PHAT-$\rho\gamma$). After MMLC, PHAT has the best results in terms of Mode frequency, but this function and MML have the highest RMSE. This shows that most of the estimated TDOAs using PHAT are concentrated around the TDOA of the main

Table 1. The statistical values (mean, mode, frequency of mode and RMSE) for first 12 files of SiSEC 2010 dataset.

Approach	(a) Source at 143° and noise at 90° Noise: Cafeteria				(b) Source at 80° and noise at 90° Noise: Cafeteria				(c) Source at 33° and noise at 90° Noise: Square			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	−2.56	0.22	−2.19	1.34	0.35	0.36	0.11	1.02	2.81	0.26	1.59	3.99
PHAT	−3.05	0.41	−2.35	1.59	0.03	0.63	−0.19	1.62	3.20	0.18	−0.42	4.95
MML	−3.01	0.29	−2.23	1.59	0.01	0.77	−0.13	1.37	−3.04	0.14	0.02	4.59
PAHT-$\rho\gamma$	−2.86	0.31	−2.40	1.19	0.35	0.39	−0.05	1.20	2.99	0.24	0.91	4.35
MMLC	−3.04	0.59	−2.76	0.85	0.12	0.62	−0.20	1.47	3.27	0.33	0.29	4.45
Approach	**(d) Source at 143° and noise at 90° Noise: Square**				**(e) Source at 143° and noise at 60° Noise: Square**				**(f) Source at 143° and noise at 60° Noise: Square**			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	−3.23	0.29	−2.42	4.76	−3.22	0.21	−1.06	2.88	−2.96	0.31	−1.72	2.41
PHAT	−3.10	0.36	−2.86	1.28	−3.25	0.29	−2.10	2.78	−3.58	0.18	−2.17	2.57
MML	−3.13	0.43	−2.96	0.88	−3.19	0.32	−1.21	3.03	−3.36	0.14	−1.65	2.63
PAHT-$\rho\gamma$	−3.23	0.37	−2.51	2.62	−3.21	0.35	−1.36	2.87	−3.14	0.37	−2.08	2.15
MMLC	−3.14	0.49	−3.10	0.65	−3.23	0.38	-2.01	2.69	−3.19	0.51	−2.66	1.97
Approach	**(g) Source at 80° and noise at 90° Noise: Subway**				**(h) Source at 80° and noise at 90° Noise: Subway**				**(i) Source at 120° and noise at 90° Noise: Cafeteria**			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	0.84	0.18	0.71	1.75	0.39	0.14	0.53	2.71	−1.78	0.13	−0.94	1.65
PHAT	0.02	0.45	0.01	1.52	0.03	0.55	−0.07	1.43	−2.09	0.23	−1.48	1.75
MML	0.01	0.84	0.06	1.03	−0.01	0.84	0.04	0.91	−2.23	0.35	−1.36	1.71
PAHT-$\rho\gamma$	0.32	0.10	0.15	2.46	0.17	0.13	0.25	2.60	−2.09	0.17	−1.19	1.56
MMLC	0.00	0.39	0.08	1.46	−0.02	0.59	0.12	1.19	−2.28	0.51	−1.59	1.58
Approach	**(j) Source at 120° and noise at 90° Noise: Cafeteria**				**(k) Source at 115° and noise at 60° Noise: Cafeteria**				**(l) Source at 120° and noise at 60° Noise: Cafeteria**			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	−1.82	0.16	−1.13	1.43	−0.92	0.26	−1.00	1.00	−2.12	0.25	−1.71	0.81
PHAT	−2.09	0.29	−1.57	1.67	−2.07	0.58	−1.56	1.23	−2.15	0.44	−2.10	1.03
MML	−2.22	0.45	−1.64	1.37	−2.06	0.68	−1.61	0.96	−2.27	0.67	−2.20	0.63
PAHT-$\rho\gamma$	−2.01	0.19	−1.25	1.56	−1.68	0.30	−1.22	0.87	−2.27	0.35	−1.85	0.77
MMLC	−2.33	0.60	−1.75	1.27	−1.92	0.83	−1.79	0.68	−2.33	0.77	−2.20	0.70

source, but anomalous estimations are far from the true TDOA.

By comparing the mode values of different weighting functions in this table, we found that three weighting functions PHAT, MML and MMLC have the nearest values to the TDOA of the main source (TDOA of the main source is –3.20), but by comparing the mean and mode values of these three functions, we can conclude that the MMLC function has the lowest difference between mean and mode values. This indicates that the anomalous estimations using MMLC have shorter distance to the TDOA of the main source.

Figure 7(a) and Table 2(a) show a case in which the main source is at 120° and square noise is located at 90°. As seen from Table 2(a), PHAT-$\rho\gamma$ and MMLC functions have the lowest RMSE, but MMLC has higher Mode frequency than PHAT-$\rho\gamma$ function. This indicates that anomalous estimations using MMLC

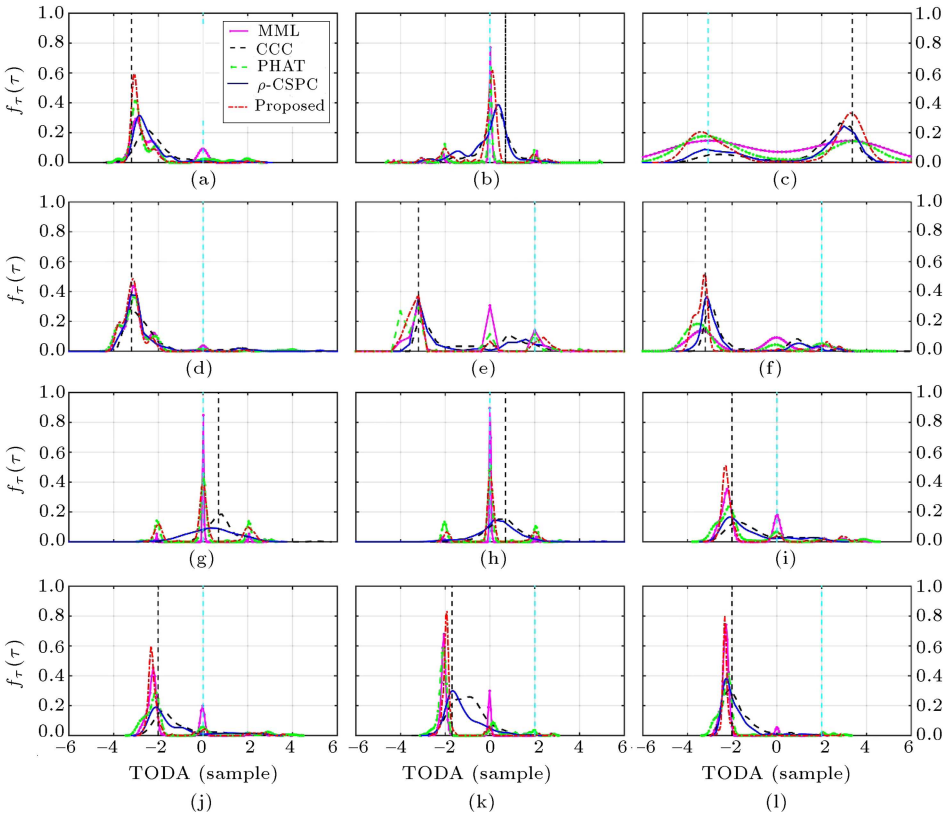

Figure 6. (a) to (l) show the histogram of the estimated TDOA for Table 1. Black dashed vertical line shows the TDOA of the main sound source and cyan line shows the TDOA of the noise source.

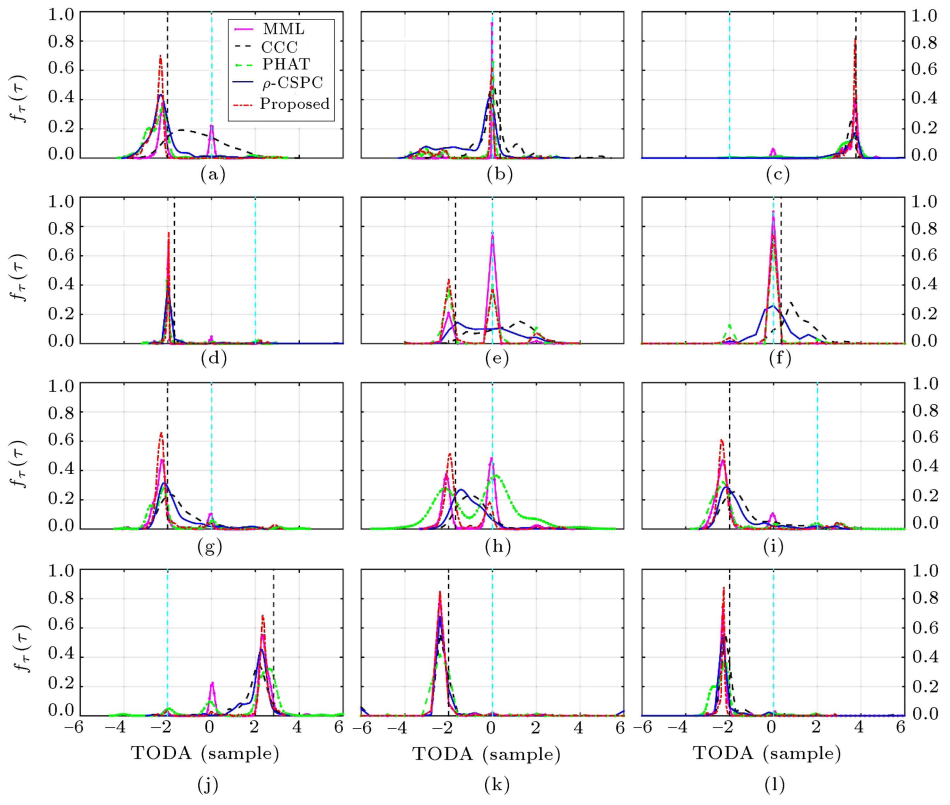

Figure 7. (a) to (l) show the histogram of the estimated TDOA for Table 2. Black dashed vertical line shows the TDOA of the main sound source and cyan vertical line shows the TDOA of the noise source.

have a greater distance from the TDOA of the main source.

By comparing the mean and mode values in PHAT-$\rho\gamma$ and MMLC functions, we found that the difference between mean and mode values of PHAT-$\rho\gamma$ is less than the MMLC function. This confirms that anomalous estimations using PHAT-$\rho\gamma$ function have a shorter distance to TDOA of the main source.

Figure 8(a) and Table 3(a) show a case in which the main source is located at 20° and the square noise source is at 120°. In this case, the MMLC function has the lowest RMSE. By comparing Mode frequency values of different functions, it can be seen that in MMLC, 72% of TDOA estimations are concentrated around TDOA of the main source and this value is 5% better than the PHAT-$\rho\gamma$ function. In this case, TDOA of the main source is 4.38 and as we can see in Table 3(a), the mode value of the PHAT function is the nearest value to the TDOA of the main source. But, mode and mean values in this function have the maximum distance to each other (in comparison to the other functions). In this case, the MMLC function has

Table 2. The statistical values (mean, mode, frequency of mode and RMSE for second 12 files of SiSEC 2010 dataset.

Approach	(a) Source at 120° and noise at 90° Noise: Square				(b) Source at 85° and noise at 90° Noise: Square				(c) Source at 20° and noise at 120° Noise: Square			
	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE
CC	−1.38	0.13	−0.69	1.70	0.02	0.52	0.16	0.98	3.55	0.27	3.16	2.94
PHAT	−2.24	0.34	−2.30	1.12	0.02	0.65	−0.76	1.81	3.70	0.14	3.19	1.26
MML	−2.25	0.38	−1.58	1.22	−0.02	0.92	−0.16	0.84	3.76	0.40	3.33	1.11
PAHT-$\rho\gamma$	−2.29	0.30	−2.30	0.81	−0.11	0.41	−0.93	1.77	3.77	0.18	3.21	2.87
MMLC	−2.33	0.70	−2.45	0.81	−0.02	0.62	−0.96	1.93	3.73	0.82	3.62	0.25
Approach	**(d) Source at 115° and noise at 60° Noise: Square**				**(e) Source at 115° and noise at 90° Noise: Subway**				**(f) Source at 85° and noise at 90° Noise: Subway**			
	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE
CC	−1.94	0.33	−1.34	2.52	1.21	0.17	0.71	2.63	0.82	0.26	1.04	1.00
PHAT	−2.04	0.43	−1.72	1.05	0.00	0.43	−0.69	1.70	0.03	0.80	−0.16	0.98
MML	−1.96	0.69	−1.83	0.84	0.00	0.85	−0.31	1.58	0.00	0.90	0.00	0.35
PAHT-$\rho\gamma$	−1.99	0.40	−1.43	2.44	−0.42	0.13	−0.34	1.80	−0.02	0.27	0.09	0.98
MMLC	−1.96	0.76	−1.87	0.80	−2.05	0.46	−0.81	1.53	0.00	0.91	−0.13	0.68
Approach	**(g) Source at 120° and noise at 90° Noise: Cafeteria**				**(h) Source at 115° and noise at 90° Noise: Cafeteria**				**(i) Source at 120° and noise at 60° Noise: Cafeteria**			
	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE
CC	−1.88	0.23	−1.43	1.06	−1.03	0.23	−0.88	1.04	−1.81	0.25	−1.32	1.24
PHAT	−2.17	0.27	−1.87	1.36	0.19	0.36	−0.58	1.88	−2.30	0.32	−1.61	1.98
MML	−2.29	0.47	−1.99	1.08	−0.05	0.48	−0.86	1.47	−2.32	0.47	−1.77	1.52
PAHT-$\rho\gamma$	−2.15	0.31	−1.63	1.04	−1.42	0.27	−1.08	0.95	−2.16	0.28	−1.49	1.32
MMLC	−2.29	0.66	−2.01	1.04	−1.93	0.51	−1.40	1.11	−2.38	0.61	−1.84	1.50
Approach	**(j) Source at 45° and noise at 120° Noise: Cafeteria**				**(k) Source at 120° and noise at 90° Noise: Square**				**(l) Source at 120° and noise at 90° Noise: Square**			
	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE	Mode	Freq.	Mean	RMSE
CC	2.14	0.35	2.03	0.96	−2.11	0.56	−2.21	4.89	−2.18	0.56	−1.38	5.06
PHAT	2.60	0.32	1.86	1.96	−2.22	0.41	−2.24	0.82	−2.20	0.37	-2.27	0.78
MML	2.34	0.54	1.89	1.49	−2.28	0.82	−2.13	0.77	−2.31	0.74	−2.04	1.76
PAHT-$\rho\gamma$	2.30	0.45	2.00	1.09	−2.08	0.68	−2.17	4.60	−2.33	0.65	−1.80	3.51
MMLC	2.33	0.68	2.16	1.15	−2.27	0.85	−2.21	0.58	−2.26	0.88	−2.23	0.63

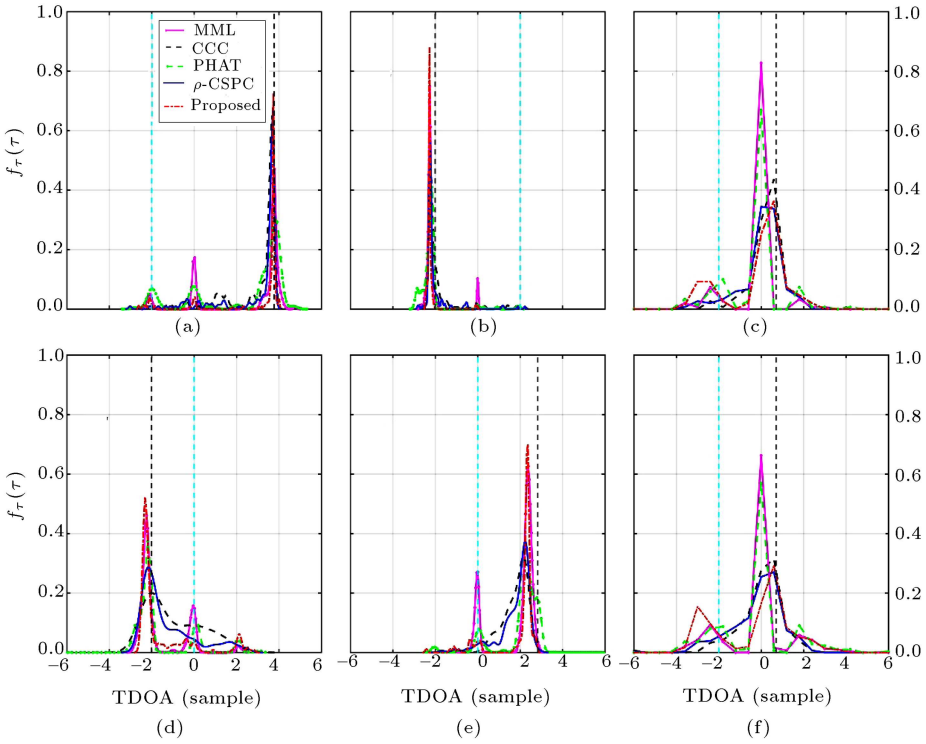

Figure 8. (a) to (f) show the histogram of the estimated TDOA for Table 3. Black dashed vertical line shows the TDOA of the main sound source and cyan vertical line shows the TDOA of the noise source.

Table 3. The statistical values (mean, mode, frequency of mode and RMSE) for the last 4 files of SiSEC 2010 dataset.

Approach	(a) Source at 20° and noise at 120° Noise: Square				(b) Source at 120° and noise at 60° Noise: Square				(c) Source at 80° and noise at 90° Noise: Cafeteria			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	3.66	0.65	2.90	1.48	−2.23	0.45	−1.82	0.92	0.32	0.43	0.27	0.87
PHAT	3.81	0.30	2.40	2.54	−2.15	0.41	−2.18	0.70	−0.01	0.67	−0.32	1.64
MML	3.78	0.47	2.64	2.24	−2.25	0.74	−2.11	0.58	0.04	0.83	−0.18	1.33
PAHT-$\rho\gamma$	3.71	0.67	2.86	1.78	−2.23	0.56	−2.02	0.70	0.30	0.34	0.04	1.16
MMLC	3.73	0.72	3.00	1.42	−2.27	0.88	−2.23	0.39	0.39	0.36	−0.15	1.68
Approach	**(d) Source at 120° and noise at 90° Noise: Subway**				**(e) Source at 45° and noise at 90° Noise: Subway**				**(f) Source at 80° and noise at 90° Noise: Cafeteria**			
	Mode	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**	**Mode**	**Freq.**	**Mean**	**RMSE**
CC	−2.00	0.20	−0.96	1.67	2.19	0.32	1.77	1.21	0.20	0.31	−0.12	2.52
PHAT	−2.20	0.37	−1.71	1.82	2.20	0.36	2.05	1.33	0.04	0.57	−0.21	1.88
MML	−2.24	0.46	−1.66	1.17	2.40	0.62	1.84	1.43	0.01	0.66	−0.14	1.64
PAHT-$\rho\gamma$	−2.14	0.28	−1.28	1.47	2.26	0.37	1.75	1.34	0.34	0.27	−0.21	2.06
MMLC	−2.32	0.52	−1.51	1.48	2.36	0.70	1.95	1.33	0.45	0.29	−0.31	2.01

the minimum difference between mode and mean values; after PHAT and MML functions, MMLC has the nearest value to TDOA of the main source. As a result, estimated TDOAs using the MMLC function have the maximum concentration on true TDOA on the one hand and anomalous estimations have minimum distance to TDOA of the main source on the other hand.

5. Conclusion

In this paper, we proposed a new weighting function for the GCC framework using a combination of modified ML and PHAT-$\rho\gamma$ functions. This function has the merits of both of them. This function uses parameter ρ and minimum value of coherence in its denominator

to improve the TDOA estimation in low SNRs and high reverberations. On the other hand, this function uses new approximation of phase variance that is used in the MML function to improve the TDOA estimation in high SNRs and low reverberation time. We evaluate our proposed function using real and synthesized datasets. In the first step, we evaluate our proposed weighting function using synthetic data and Monte-Carlo simulation against SNR and reverberation time variations. Simulation results show that in terms of RMSE and in low SNRs, PHAT-$\rho\gamma$ and MMLC have the best results due to using parameter ρ and the minimum value of coherence in the denominator of the functions. As SNR increases, our proposed function shows better results due to better approximation of phase variance. For example, in 0.8 s reverberation time and in SNRs higher than 50 dB, RMSE of MMLC is 0.277 samples less than the MML function. In the second step, we evaluate our proposed function using a real-world dataset, and we compare our proposed weighting function with CCC, PHAT, MML, and PHAT-$\rho\gamma$ functions. We used the SiSEC 2010 dataset as a real-world dataset for comparing weighting functions. For each file in this dataset, we plot a histogram of the estimated TDOAs, and also for a better comparison of weighting functions, we calculate four statistical parameters: mean, mode, frequency of mode, and RMSE.

References

1. Nikunen, J. and Virtanen, T. "Direction of arrival based spatial covariance model for blind sound source separation", *IEEE/ACM Trans. Audio, Speech, Language Process*, **22**(3), pp. 727-739 (2014).
2. Kim, U.-H., Nakadai, K. and Okuno, H.G. "Improved sound source localization in horizontal plane for binaural robot audition", *Appl. Intell.*, **42**(1), pp. 63-74 (2015).
3. Marti, A., Cobos, M. and Lopez, J.J. "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments", *36th Int. Conf. on Acoust. Speech and Signal Process*, pp. 2592-2595 (2011).
4. Hu, J.-S., Lee, M.-T. and Yang, C.-H. "Robust adaptive beamformer for speech enhancement using the second-order extended $H\infty$ filter", *IEEE Trans. Audio, Speech, and Language Process*, **21**(1), pp. 39-50 (2013).
5. Zhang, Q., Chen, Z. and Yin, F. "Microphone clustering and BP network based acoustic source localization in distributed microphone arrays", *Adv. in Electr. Comp. Eng.*, **13**(4), pp. 33-40 (2013).
6. Liu, C., Zakharov, Y.V. and Chen, T. "Broadband underwater localization of multiple sources using basis pursuit de-noising", *IEEE Trans. Signal Process*, **60**(4), pp. 1708-1717 (2012).
7. Van den Bogaert, T., Carette, E. and Wouters, J. "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna", *Int. J. of Audiology*, **50**(3), pp. 164-176 (2011).
8. Chen, H. and Ser, W. "Sound source DOA estimation and localization in noisy reverberant environments using least-squares support vector machines", *J. of Signal Process. Syst.*, **63**(3), pp. 287-300 (2011).
9. Cohen, I., Benesty, J. and Gannot, S. "Speech processing in modern communication: challenges and perspectives", **3** ,*Springer Science & Business Media*, pp. 307-337 (2009).
10. Brandstein, M. and Ward, D., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Science & Business Media, pp. 181-201 (2001).
11. Benesty, J., *Springer Handbook of Speech Processing*, Springer Science & Business Media, pp. 1043-1063 (2008).
12. Benesty, J. and Huang, Y., *Adaptive Signal Processing: Applications to Real-World Problems*, Springer Science & Business Media, pp. 227-247 (2013).
13. Chen, J., Huang, Y.A. and Benesty, J. "A comparative study on time delay estimation in reverberant and noisy environments", In *IEEE Workshop Applicat. of Signal Process. Audio and Acoust.*, pp. 21-24 (2005).
14. Lee, B., Said, A., Kalker, T. and Schafer, R.W. "Maximum likelihood time delay estimation with phase domain analysis in the generalized cross correlation framework", In *Hands-Free Speech Commun. and Microphone Arrays, HSCMA*, pp. 89-92 (2008).
15. Liu, H. and Shen, M. "Continuous sound source localization based on microphone array for mobile robots", In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 4332-4339 (2010).
16. Knapp, C.H. and Carter, G.C. "The generalized correlation method for estimation of time delay", *IEEE Trans. Acoust. Speech Signal Process*, **24** (4), pp. 320-327 (1976).
17. Hannan, E.J. and Thomson, P. "Estimating group delay", *Biometrika*, **60**(2), pp. 241-253 (1973).
18. Carter, G.C., Nuttall, A.H. and Cable, P.G. "The smoothed coherence transform", *Proceedings of the IEEE*, **61**(10), pp. 1497-1498 (1973).
19. Argentieri, S., Danes, P. and Soueres, P. "A survey on sound source localization in robotics: From binaural to array processing methods", *Computer Speech & Language*, **34**(1), pp. 87-112 (2015).

20. Rabinkin, D.V., Renomeron, R.J., Dahl, A., French, J.C., Flanagan, J.L., and Bianchi, M. "DSP implementation of source location using microphone arrays", in *SPIE's Int. Symp. Optical Sci., Eng. Instrum.*, pp. 88-99 (1996).

21. Guide, M.U.s. "The mathworks", Inc., Natick, MA, **5**, p. 333 (1998).

22. Perez-Lorenzo, J., Viciana-Abad, R., Reche-Lopez, P., Rivas, F. and Escolano, J. "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments", *Appl. Acoust.*, **73**(8), pp. 698-712 (2012).

23. Allen, J.B. and Berkley, D.A. "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.*, **65**(4), pp. 943-950 (1979).

24. "Signal Separation Evaluation Campaign (SiSEC)", URL: http://sisec.inria.fr (2010).

25. Dvorkind, T.G. and Gannot, S. "Time difference of arrival estimation of speech source in a noisy and reverberant environment", *Signal Process.*, **85**(1), pp. 177-204 (2005).

Biographies

Mir Saber Hosseini received his BSc degree in Electrical Engineering from Urmia University, Iran, in 2012 and received his MSc degree from Amirkabir University of Technology, Tehran, Iran in 2014. His main interests include acoustic signal processing, speech processing, blind source separation, cocktail party effect, and pattern recognition.

Amir Hossein Rezaie graduated with a degree in Electrical Engineering from the Amirkabir University of Technology, Iran, in 1983 and he received his PhD in Engineering from Bristol University, England, in 1988. Since 1988, he has been an Associated Professor in the Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran. His main interests include the fields of automation, digital design, and wireless sensor networks.

Yousef Zanjireh received MSc and PhD degrees from Technical University of Amirkabir (in Electrical Engineering), Tehran, Iran in 2003 and 2014, respectively. From 2003-2007, he was employed at Research Center of Intelligent Systems, working on acoustic signal processing. From Dec. 2015-July 2016, he was a Senior Scientist at CINTAL Lab, University of Algarve, Faro, Portugal, where he worked on geoacoustic and seismic inversion techniques. His research interests include optimization, modeling, recognition, classification and inversion of acoustic, seismic and biomedical signals.