Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 621390, 13 pages
doi:10.1155/2011/621390

Research Article

Performance Improvement of TDOA-Based Speaker Localization in Joint Noisy and Reverberant Conditions

Hamid Reza Abutalebi (EURASIP Member)[1,2] and Hossein Momenzadeh[1]

[1] *Speech Processing Research Lab (SPRL), Electrical and Computer Engineering Department, Yazd University, 89195-741 Yazd, Iran*
[2] *Idiap Research Institute, CH-1920 Martigny, Switzerland*

Correspondence should be addressed to Hamid Reza Abutalebi, habutalebi@yazduni.ac.ir

Received 30 April 2010; Revised 15 October 2010; Accepted 14 January 2011

Academic Editor: Ioannis Psaromiligkos

Copyright © 2011 H. R. Abutalebi and H. Momenzadeh. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

TDOA- (time difference of arrival-) based algorithms are common methods for speech source localization. The generalized cross correlation (GCC) method is the most important approach for estimating TDOA between microphone pairs. The performance of this method significantly degrades in the presence of noise and reverberation. This paper addresses the problem of 3D localization in joint noisy and reverberant conditions and a single-speaker scenario. We first propose a modification to make the GCC-PHAse transform (GCC-PHAT) method robust against environment noise. Then, we use an iterative technique that employs location estimation to improve TDOAs accuracy. Extensive experiments on both simulated and real (practical) data (in a single-source scenario) show the capability of the proposed methods to significantly improve TDOA accuracy and, consequently, source location estimates.

1. Introduction

The ever-increasing communication between humans and machines needs localizing and tracking of acoustic sources. Automatic camera tracking for video-audio applications, microphone array beamforming for suppressing noise and reverberation, distant-talking speech recognition and robot audio systems are sample applications for speech source localization [1–8].

The problem of sound source (speaker) localization has been extensively explored in the last two decades; state-of-the-art methods for sound source localization can be generally classified into four categories [9]: (1) time difference of arrival (TDOA)-based techniques, (2) steered response power- (SRP-) based methods, (3) energy ratio estimation, and (4) subspace characterization. These methods usually employ linear [10, 11] or circular [12–16] microphone arrays to locate the sound source. Considering the practical issues (such as small intermicrophone distance, reverberant environments, etc.), the choices for sound source localization will be actually limited to TDOA- and SRP-based categories [9]. TDOA-based methods have been widely employed in recent years mainly because of their low-complexity. Although SRP-based algorithms, especially the well-known SRP-PHAse transform (SRP-PHAT) method, have shown very good results in sound source localization, the computational complexity is much higher than TDOA-based methods [17].

Also, previous works have been focused mostly on estimating only azimuth, mainly by means of circular arrays (e.g., see [14, 15]). So, the application of a limited number of microphones for 3D localization of the speaker (in either Cartesian or polar coordinates) is still a challenging problem.

In this research, we have focused on 3D localization of (single) speaker in a practical (joint noisy and reverberant) room. Adding the constraint of low complexity, the most appropriate option would be the TDOA-based family. To make 3D localization feasible, we have proposed and implemented a new triangular-shape microphone placement (explained in Section 6).

In the TDOA-based methods, firstly, the TDOA of the signals is estimated for each microphone pair (TDOA estimation stage), then, the source location is estimated based on these TDOAs (location estimation stage) [3].

When only two microphones are available, there are two main approaches for TDOA estimation [18]: the first approach works based on blind estimation of the impulse responses between the source and two microphones [19, 20]. In the other approach, relative delay is directly estimated from the cross correlation of two microphone signals [21–23].

The generalized cross correlation (GCC) method [21] is the most common and the fastest two-channel algorithm for TDOA estimation [18]. The delay is obtained as the time lag that maximizes the cross correlation between (the filtered version of) the received signals [21].

The accuracy of estimated TDOAs is very important, since any error in TDOAs leads to a high error in localization [24]. In real acoustic environments, the accuracy of TDOAs is degraded due to noise and/or reverberation. Several modifications have been proposed to improve the performance of TDOA-based methods in noisy or reverberant situations. While most of these modifications have been proposed to improve the localization accuracy in reverberant environments [24–27], a few of the others deal with noisy conditions [21, 28].

In many practical situations (like meeting rooms), the situation becomes more severe, where the source localization should be done in the presence of both noise and reverberation [28]. This problem has drawn increasing attention in recent years.

One approach would be the use of single-step (direct) methods that preserve and propagate all the intermediate information and use them to estimate the source location at the very last step. A modified version of this class, steered beam (SB) sound source localization has been proposed in [29]. This method has similarities with the SRP-based category and is a good choice when the computational complexity is not the main constraint.

As another solution, a method has been proposed in [30] that employs harmonicity of the speech signal to handle the localization in joint noisy and reverberant situations. This method (and most of the recent works) has high computational complexity and/or fails to provide acceptable performance. So, the topic is still being researched.

This paper aims to improve the performance of the state-of-the-art and simple GCC-based source localization methods in practical joint noisy and reverberant situations. We firstly explain the GCC basics and its variants. Then, noting the defects of these techniques in real (practical) applications, we propose a novel modification of the GCC for TDOA estimation in joint noisy and reverberant situations.

Furthermore, we propose a hybrid localization method to improve the accuracy. In this algorithm, TDOA estimation is iteratively combined with source localization estimation to improve the accuracy of TDOA estimation. This, in turn, makes the source localization more accurate. In the proposed method, TDOA estimation is modified according to the primary estimated location of source (that is estimated by a closed form method such as spherical interpolation (SI) or spherical intersection (SX) [31]). Moreover, we supplement an outlier removal technique to the system that improves the localization accuracy.

By implementing the proposed modifications and evaluating the whole system on simulated and real (practical) data, we have demonstrated the superiority of the proposed methods in accurate speech source localization.

The rest of this paper is organized as follows. In Section 2, the GCC method is described. The modified GCC-PHAT method is presented in Section 3. Section 4 explains closed-form source location estimation methods. In Section 5, hybrid localization method and outlier removal are presented. Sections 6 and 7 explain the setup and the results of the experiments on the simulated and real data, respectively. Finally, some concluding remarks are given in Section 8.

2. Generalized Cross Correlation Method

The GCC algorithm uses time delay information from only one pair of microphones [21]. Due to the use of FFT, the computational complexity of GCC is low; therefore, it is a common choice for real-time applications.

In this method, delay estimation is obtained via [18, 21]

$$\hat{\tau}_{\mathrm{GCC}} = \arg\max_{m} \Psi_{\mathrm{GCC}}[m], \tag{1}$$

where

$$\Psi_{\mathrm{GCC}}[m] = \sum_{K-0}^{K-1} \Phi[k] S_{x_0x_1}[k] e^{j2\pi mk/K}, \tag{2}$$

is the so-called GCC Function (GCCF) and m is the delay index (in samples). $S_{x_0x_1}[k]$ is the cross spectrum and is approximately equal to $X_0[k]\ X_1^*[k]$, where $X_n[k]$ is the DFT of $x_n[n]$ and $*$ is the (complex) conjugate operator. Also, $\Phi[k]$ is a weighting function. Several weighting functions have been proposed in the literature, two of the most important of them will be described in the following.

2.1. GCC-PHAT Algorithm. In this method, the weighting function is applied by a PHAse Transform (PHAT) function defined as [21]:

$$\Phi_{\mathrm{PHAT}}[k] = \frac{1}{|S_{x_0x_1}[k]|}. \tag{3}$$

Neglecting noise effects in (2), we can deduce that the weighted cross correlation spectrum is free from the source signal and depends only on the channel response. More precisely, it can be shown [16] that the PHAT is a special case of the maximum likelihood (ML) approach for sound localization under low noise conditions. Moreover, PHAT remains an optimal solution in ML sense regardless of the amount of reverberation [16]. This way, we can justify good performance of the method in reverberant situations.

2.2. GCC-ML Algorithm. In this case, the weighting function is a maximum likelihood (ML) filter defined as [21]

$$\Phi_{\mathrm{ML}}[k] = \frac{|X_0[k]||X_1[k]|}{|N_1[k]|^2|X_0[k]|^2 + |N_0[k]|^2|X_1[k]|^2}, \tag{4}$$

where $N_n[k]$ is the noise power spectrum in the nth microphone and is estimated during silent frames [3]. In the ML filter, signal and noise are assumed independent and stationary. So, in reverberant environments where these conditions are not satisfied, the performance of the GCC-ML method will drastically degrade.

3. Modified GCC-PHAT Algorithm

The most important problem with GCC-PHAT method is its low robustness in noisy situations. This problem can be justified by the identical contribution of different frequency bins in the PHAT weighting function. In other words, even the frequency components with dominant noise have the same effect in the PHAT function calculation.

To de-emphasize the effect of noisy frequency components, we propose a method based on the idea of generalized spectral subtraction method (a well-known technique in speech enhancement [32]). We call this new method GCC-Modified PHAT (or briefly, GCC-MPHAT).

The proposed method works as follows: First, for each microphone signal, the normalized quantity $w'[k]$ is obtained according to signal spectrum and the estimation of the noise spectrum in each frame via

$$w'[k] = \frac{|X[k]|^{\alpha} - \beta|N[k]|^{\alpha}}{|X[k]|^{\alpha}}, \tag{5}$$

where α and β are spectral subtraction parameters that are determined according to the environment situations. $N[k]$ is the noise power spectrum in the microphone and is estimated in a way similar to that in GCC-ML algorithm. Then, we define $w[k]$ as

$$w[k] = \begin{cases} 1, & w'[k] > R, \\ \gamma, & w'[k] < R, \end{cases} \tag{6}$$

where R is a threshold value ($0 \le R \le 1$) and $0 \le \gamma < 1$ is a floor value for noisy frequency components. Finally, the PHAT filter (3) is modified as

$$\Phi_{\mathrm{MPHAT}}[k] = \frac{w_0[k]w_1[k]}{|X_0[k]X_1[k]|}. \tag{7}$$

$w_0[k]$ and $w_1[k]$ are computed through (6) for the first and second microphones, respectively.

4. Closed-Form Source Location Estimation

In a constant sound velocity environment, the TDOAs are proportional to differences in source-sensor ranges, called range-differences (RDs). The source location is conventionally found as a weighted intersection of the set of constant-RD hyperboloids. This results in a nonlinear set of equations with high computational complexity. Although several optimal solutions have been proposed for this problem in the literature, suboptimal closed-form solutions (like SI and SX) are of much interest due to the tremendous computational savings. SX and SI localization methods can be briefly explained as follows [31].

Considering $\underline{x}_s = (x_s, y_s, z_s)$ as the source position and $\underline{x}_i = (x_i, y_i, z_i)$ as the position of ith microphone, the source-microphone distance, source-origin distance, and microphone-origin distance are determined via $D_i = \|\underline{x}_i - \underline{x}_s\|$, $R_s = \|\underline{x}_s\|$ and $R_i = \|\underline{x}_i\|$, respectively. Hence, the RD between the ith and jth microphone will be $d_{ij} = c \cdot \tau_{ij} = D_i - D_j$, $(i = 1, \ldots, N,\ j = 1, \ldots, N)$. In the SI or SX methods, $\underline{x}_s$ is determined such that matches with d_{ij}'s, in a suboptimal manner.

Defining the error vector as $\underline{\varepsilon} = \underline{\delta} - 2R_s\underline{d} - 2S\underline{x}_s$, where

$$\underline{\delta} = \begin{bmatrix} R_2^2 - d_{21}^2 \\ R_3^2 - d_{31}^2 \\ \vdots \\ R_N^2 - d_{N1}^2 \end{bmatrix}, \quad d = \begin{bmatrix} d_{21} \\ d_{31} \\ \vdots \\ d_{N1} \end{bmatrix}, \quad S = \begin{bmatrix} x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{bmatrix}, \tag{8}$$

and considering W as the error weighting matrix, by minimizing $\underline{\varepsilon}^T W \underline{\varepsilon}$, the least squares (LS) solution for the source location is obtained as

$$\underline{x}_s = \frac{1}{2} S_w^* (\underline{\delta} - 2R_s\underline{d}), \tag{9}$$

where

$$S_w^* = \left(S^T W S\right)^{-1} S^T W. \tag{10}$$

The SI and SX methods are suboptimal solutions that approximate the above nonlinear problem. In the SI method, the source location is estimated as

$$\underline{x}_s = \frac{1}{2} S_w^* \left(\underline{\delta} - 2\widetilde{R}_s\underline{d}\right), \tag{11}$$

where

$$\widetilde{R}_s = \frac{\underline{d}^T P_s^0 V P_s^0 \underline{\delta}}{2\underline{d}^T P_s^0 V P_s^0 \underline{d}}, \quad P_s = S\left(S^T W S\right)^{-1} S^T W, \quad P_s^0 = 1 - P_s. \tag{12}$$

The SX solution is obtained by substituting the LS solution (9) for $\underline{x}_s$ given R_s into the quadratic equation $R_s^2 = \underline{x}_s^T \underline{x}_s$:

$$R_s^2 = \left[\frac{1}{2} S_w^* (\underline{\delta} - 2R_s\underline{d})\right]^T \left[\frac{1}{2} S_w^* (\underline{\delta} - 2R_s\underline{d})\right]. \tag{13}$$

After expansion, the above equation yields the standard form $aR_s^2 + bR_s + c = 0$, where

$$\alpha = 4 - 4\underline{d}^T S_w^{*T} S_w^* \underline{d}, \quad b = 4\underline{d}^T S_w^{*T} S_w^* \underline{\delta}, \quad c = -\underline{\delta}^T S_w^{*T} S_w^* \underline{\delta}. \tag{14}$$

This quadratic equation has two solutions of the form

$$R_s = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{15}$$

where the positive one is taken as an estimate of the source-to-origin distance. Substituting this value in (9), the source location, $\underline{x}_s$, is estimated.

5. Hybrid Method for Source Localization

5.1. Problem Definition. TDOA estimation algorithms can be employed in two- or multi-microphone forms. Although two-microphone algorithms are fast, in real-life applications, they fail in estimation of accurate TDOA. On the other hand, multimicrophone algorithms use redundant information of several microphone-pairs and have better performance in TDOA estimation. An example method that uses this redundancy for the disambiguation of TDOA estimations in multipath multisource environments is the DATEMM that is proposed in [33].

Many source localization techniques do the TDOA estimation and location estimation as two separate stages; however, these two stages are obviously related. In the conventional algorithms, if the estimate of TDOA is erroneous (due to noise and/or reverberation), there will be no way to correct it. Actually, if the estimated TDOA of only one microphone pair is erroneous, the source location estimation (the second stage of the whole localization process) will be biased.

5.2. Proposed Hybrid Method. As in the typical example shown in Figure 1, in the case of incorrect estimation of TDOA, the GCC-PHAT function usually has a local maximum in the correct delay sample; however, this maximum is not a global one. The idea we have used in this research can be explained as follows. By employing information about primary estimation of source location and microphone positions, we find a (more) correct local maximum for the GCC function (or a more correct TDOA estimation). In turn, a more accurate estimation of source location will be available. The process is iterated until a convergence in estimated location is reached. This idea has been employed in the proposed hybrid localization method as explained in the following.

Assuming the (true) source location is known, exact TDOA estimation in ith microphone pair can be written as

$$\tau_{\text{exact}} = \frac{|s - m_{i1}| - |s - m_{i0}|}{c}, \tag{16}$$

where s is source location, $c = 341$ m/s is the velocity of sound, and m_{i0} and m_{i1} are the microphone positions.

In the proposed hybrid method, a primary estimation of source location ($\hat{s}_P$) is first calculated using the primary TDOA (τ_P) values of all microphone pairs; this is done using the SI or SX methods (such as were explained in Section 4). Due to erroneous TDOAs in the input of the SX or SI method, $\hat{s}_P$ will be biased. Then, by substitution of s with $\hat{s}_P$ in (16), we obtain a new TDOA value (called "Intermediate TDOA" or τ_I). For the microphone pairs with correct TDOA, intermediate TDOA (τ_I) and primary TDOA (τ_P) are expected to be about the same, but this is not the case for microphone pairs with incorrect TDOA. Although $\hat{s}_P$ is a biased estimation of the source location, it can be shown [22] that the primary estimation of direction of arrival (DOA) is not so affected by erroneous TDOAs. This justifies the iterative use of (16). In practice, this update process is run iteratively only for the microphone pairs which have an intermediate TDOA in a predefined range of the primary one (such that has been explained in Section 5.3).

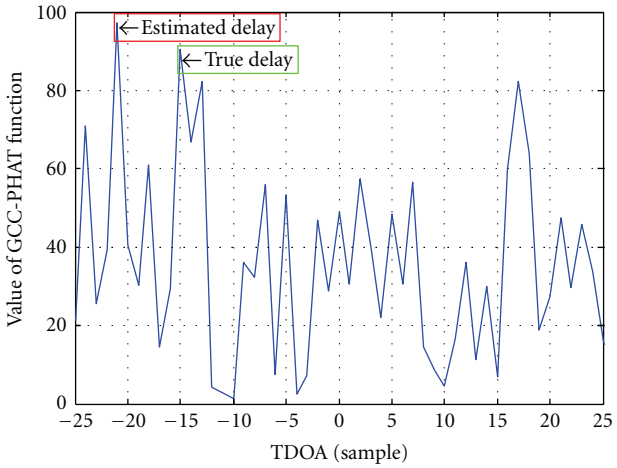

Figure 1: Typical curve of GCC function.

According to the above explanation, it is expected that the true value of the TDOA will be in the neighboring interval of τ_I. By searching the GCC function around the intermediate TDOA, we find the correct local maximum. In turn, this determines the accurate TDOA (called "Final TDOA" or τ_F). τ_F is calculated through

$$\tau_F = \arg \max_{\tau_I - \Delta \le m \le \tau_I + \Delta} \Psi_{\text{GCC}}[m], \tag{17}$$

where 2Δ determines the search interval among the delay index, m. Δ should be small enough that an incorrect global maximum does not lie in the search interval and also should be large enough that the correct local maximum lies in the search interval.

5.3. TDOA Outlier Removal. To improve the accuracy of source localization, we have also proposed the elimination of outlier TDOA estimates.

It is known that for 3D source localization, four microphones are necessary. If we employ more than four microphones (or equivalently, more than three independent TDOAs), we will have some degrees of freedom to remove outlier TDOAs. Removing outlier TDOAs leads to more accurate location estimation [34]. In the proposed hybrid system, if the difference between τ_I and τ_P is more than a predefined threshold (T), we deduce that the TDOA estimate is incorrect and that it should be removed. The outlier elimination process is explained mathematically as follows:

$$|\tau_I - \tau_P| \begin{array}{c} \text{Remove} \\ > \\ < \\ \text{Not Remove} \end{array} T. \tag{18}$$

There is an obvious tradeoff between keeping as many correct TDOAs as possible and removing erroneous ones. Thus, the optimal value of T is determined experimentally. In our case, we use $T = 5$.

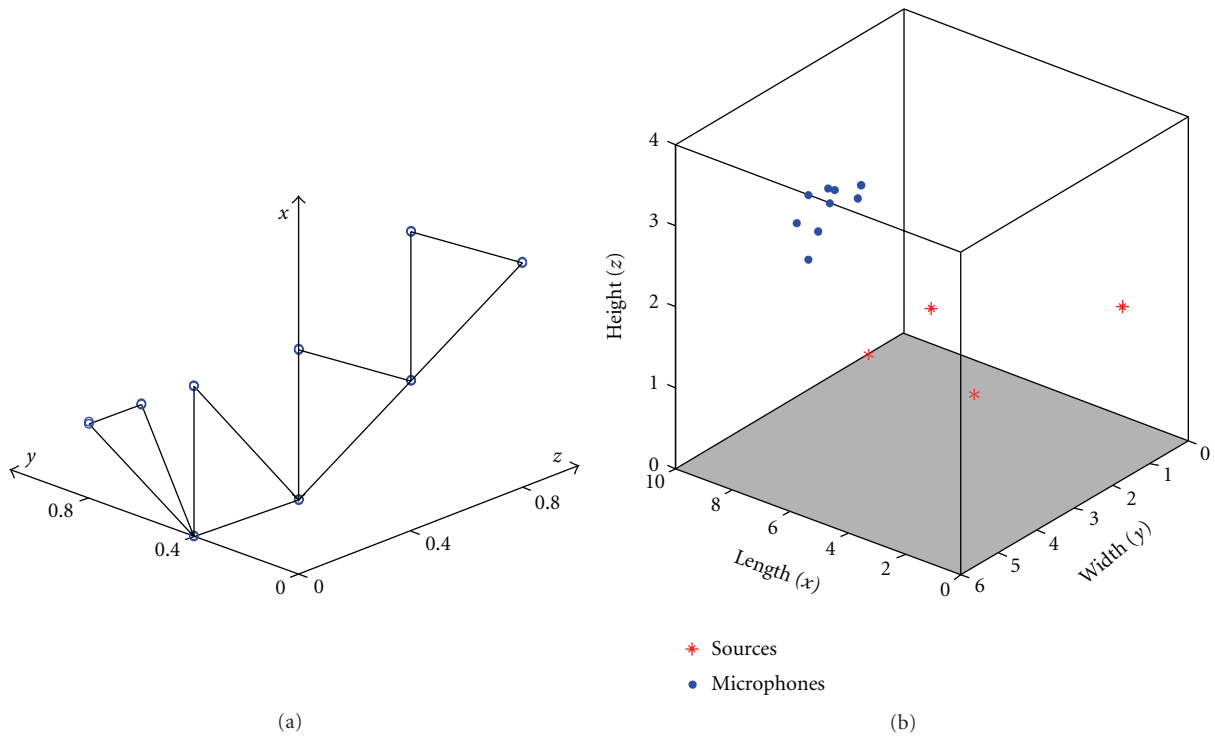

Figure 2: (a) Triangular-shape microphone array. (b) Schematic representation of the simulated room.

Table 1: Comparison between 3D RMSE of various speech source localization methods on artificially generated data.

Method	GCC method	RMSE (m) for near source	RMSE (m) for middle-center source	RMSE (m) for middle-corner source	RMSE (m) for far source	Average RMSE (m)
SI	PHAT	2.094	1.190	1.488	2.344	1.779
	MPHAT	1.964	1.116	1.339	1.877	1.574
SX	PHAT	1.692	0.941	1.157	1.894	1.421
	MPHAT	1.653	0.897	1.058	1.560	1.292
Hybrid SI	PHAT	1.271	0.739	0.934	1.520	1.116
	MPHAT	1.260	0.712	0.865	1.151	0.997
Hybrid SX	PHAT	1.069	0.597	0.758	1.112	0.884
	MPHAT	0.971	0.561	0.668	0.901	0.775
SI + outlier remove	PHAT	1.442	0.809	1.027	1.610	1.222
	MPHAT	1.281	0.762	0.929	1.280	1.063
SX + outlier remove	PHAT	1.297	0.647	0.802	1.242	0.997
	MPHAT	1.116	0.631	0.776	1.073	0.899
Hybrid SI + outlier Remove	PHAT	1.247	0.649	0.814	1.266	0.994
	MPHAT	1.184	0.645	0.754	1.065	0.912
Hybrid SX + outlier remove	PHAT	0.924	0.546	0.709	1.089	0.817
	MPHAT	0.919	0.530	0.641	0.882	**0.743**
SRP-PHAT		0.980	0.655	0.605	0.950	0.798

We note that the outlier removal procedure has been practically implemented in the body of hybrid localization method (explained in Section 5.2).

6. Experiments on Simulated Data

To evaluate the effect of the proposed modifications, we first simulated a practical room. The parameters of this simulation are explained as follows. More details about the selection of these parameters is available in [35].

(a) Dimensions of the simulated room: $10 \times 6 \times 4$ m ($x \times y \times z$).

(b) Array structure and position: we have considered a novel 3D (multi-) triangular-shape microphone array (already proposed by the authors in [35])

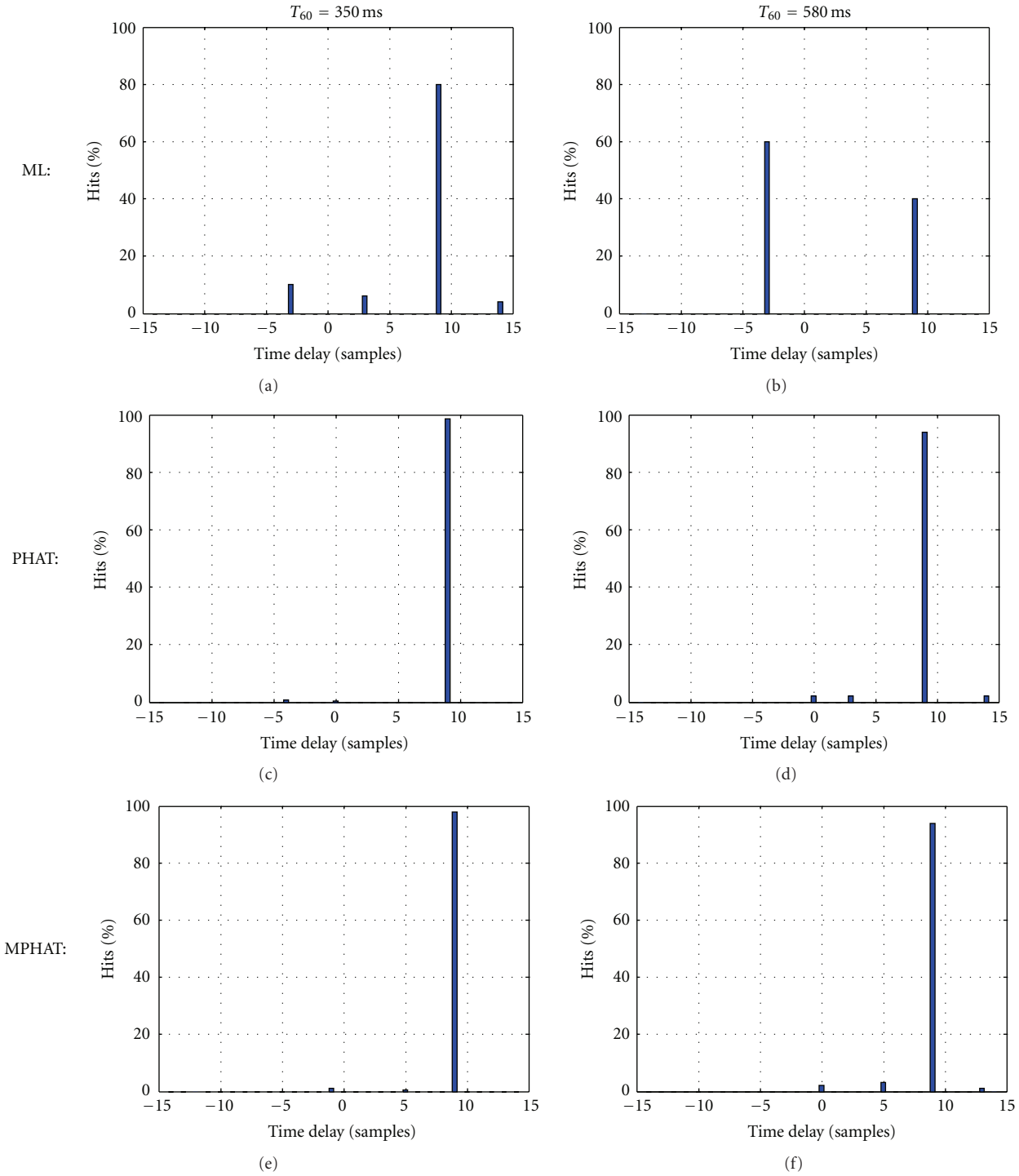

Figure 3: TDE performance in reverberant situations (exact value of TDOA is 9 samples).

that is depicted in Figure 2(a). The array consists 9 point microphones with a spacing of 40 cm. Superior performance of the triangular-shape array has been demonstrated in comparison with rectangular- and L-shape arrays. This can be justified by proper coverage of all dimensions yielded by the proposed triangular-shape array. The location of the array in the room is shown in Figure 2(b) (note to the coordinates). As shown, the reference microphone is located at (5, 6, 4).

(c) Source location: we focus solely on single-speaker localization. To examine the effect of speaker position (relative to the array), the experiments were repeated for four different source positions; these are: (5, 5, and 1.8) (near to and in front of the array), (5, 3, and

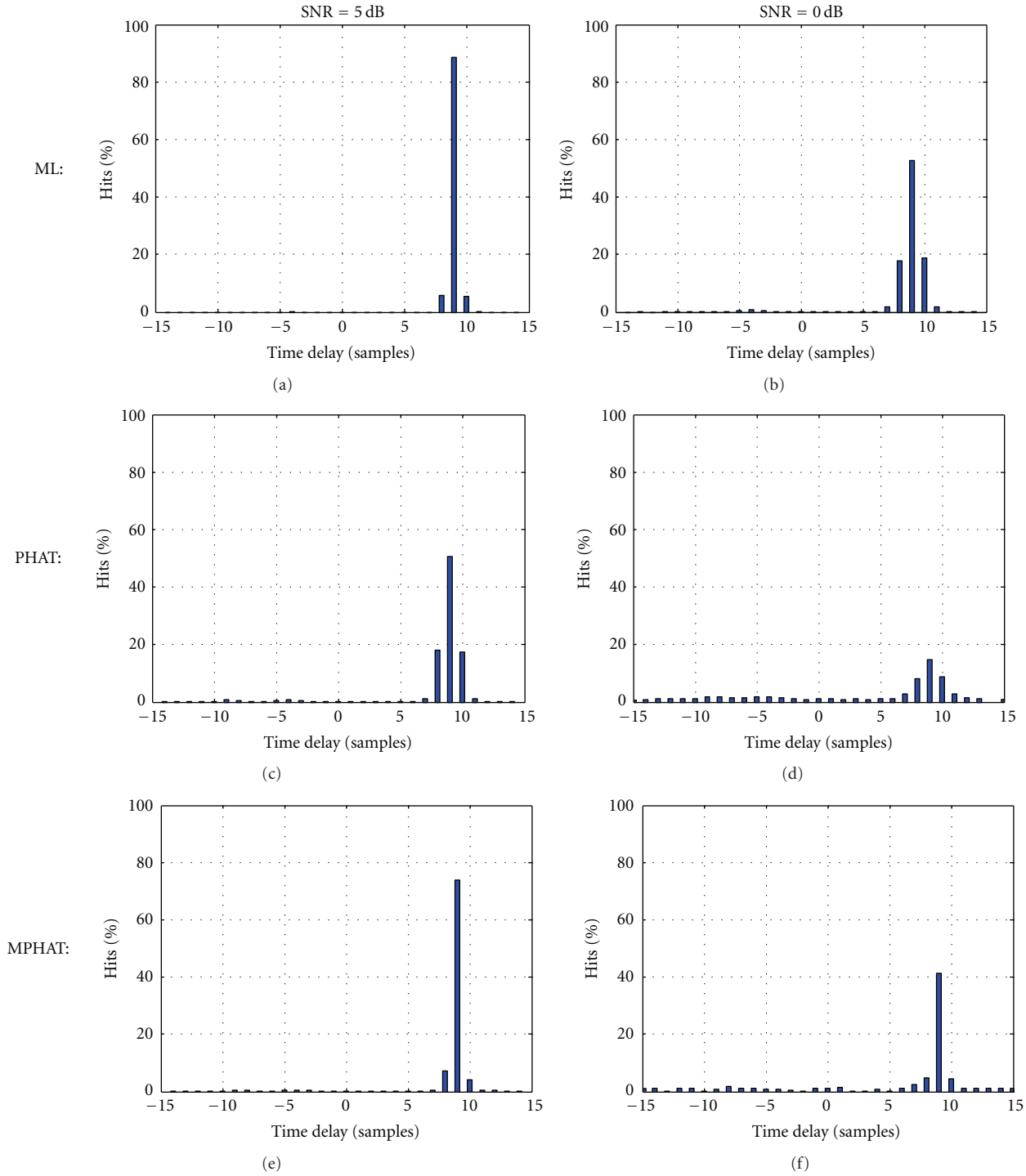

Figure 4: TDE performance in noisy situations (exact value of TDOA is 9 samples).

1.8) (middle-center the room, in front of the array), (3, 4, and 1.8) (middle-corner of the room), and (1, 1, and 1.8) (far from the array).

(d) Reverberation and noise modeling: for reverberation modeling, we have used the image method [36]. The reverberation time has been assumed $T_{60} = 350$ ms and $T_{60} = 350$ ms to model moderate- and high-reverberant rooms, respectively. Once the impulse responses from the source to each microphone were determined, the speech signal was convolved with the synthetic impulse responses. The original speech signal was from a male speaker, digitized at 16-bit resolution at $F_S = 16$ kHz. The original signal was from the TIMIT database [37] and had

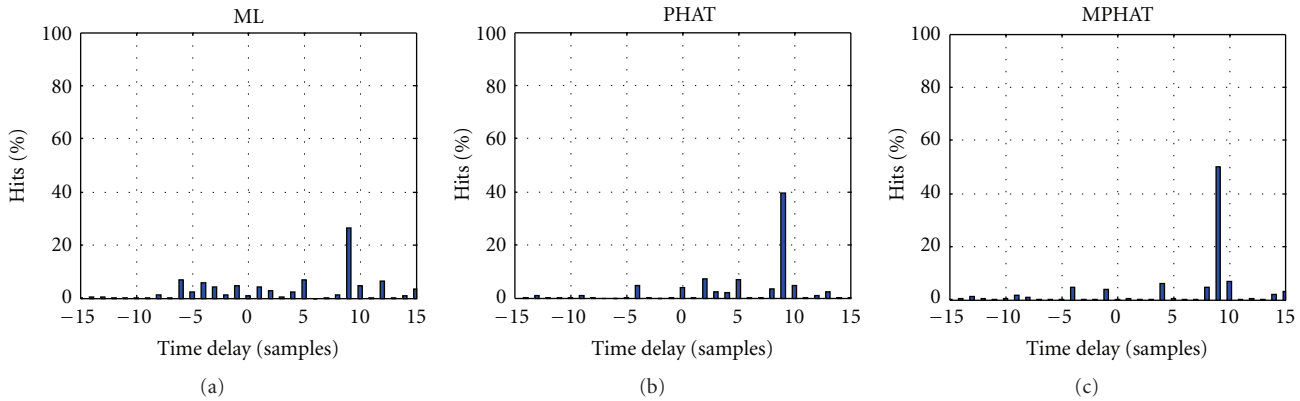

Figure 5: TDE performance in joint noisy and reverberant situations (exact value of TDOA is 9 samples).

about 30 s time length. Finally, mutually independent white Gaussian noise was scaled and added to each microphone signal to set the SNR at two levels (0 and 5 dB).

In the implementation of TDOA estimation and localization algorithms, the values of the parameters were selected as follows. Ideally, optimal values for most of these parameters should be determined adaptively (in different environments and even different frames of signal).

(a) The processes have been done in a frame-by-frame basis. The reported results are the average over all active (speech) frames of 30 s input microphone signals. Speech presence was detected using a voice activity detector (VAD). In all the experiments, a 64 ms (or $K = 1024$ at $F_s = 16$ kHz) nonoverlapping Kaiser window was applied to the frames.

(b) The parameters of (5) and (6) (i.e., α, β, R, and γ) are practically dependent on the frame SNR. To determine the optimal values for each of these parameters, we fixed the other three parameters and examined the effect of several different values for the intended parameter on the accuracy of TDOA estimation. Extensive trials were done on the all simulated microphone signals gathered for above-mentioned four source positions. A detailed report on these examinations is available at [35]. It was shown that optimal values for α are in the range of $0.8 \leq \alpha \leq 1$. Also, examining three different values for β (0.4, 0.7, and 1), it was shown that much better results could be achieved in the case of $\beta = 0.7$, while smaller values for β make the performance of the MPHAT very similar to that of PHAT, the larger values of β remove many informative frequency bins. In the tradeoff between noise reduction and signal distortion, the optimal value for R was found to be $R = 0.2$. Also, optimal value for the noise floor level was determined via extensive trials on different values of γ, while large values for γ make the MPHAT similar to the PHAT, small (or near zero) values for γ degrade the performance of TDOA estimator in reverberant situations. Briefly, the following values for the algorithm parameters were used in our experiments

$$\alpha = 1, \quad \beta = 0.7, \quad R = 0.2, \quad \gamma = 0.1. \tag{19}$$

(c) Using extensive trials on the outlier removal algorithm, proper values for the parameters were found to be $T = 5$ and $\Delta = 5$. Experiments show that these values lead to acceptable results in almost all cases. It is noted that both sampling frequency and microphone spacing have a direct effect on TDOA, and consequently, on the value of T. Also, the sampling frequency directly affects the search interval (Δ).

6.1. Performance of GCC-MPHAT. To compare TDOA estimation methods, we evaluated their performance in reverberant and noisy situations, separately, in Figures 3 and 4. As a sample, we report TDOAs of the third microphone pair in the case of second source position (middle-center the room, in front of the array). Similar comparative results are obtained for other microphone pairs and different source locations. The histograms of TDOA estimates of GCC-ML, -PHAT, and -MPHAT functions in different reverberant situations are depicted in Figure 2, while those for different noisy situations are shown in Figure 3.

As illustrated in Figure 3, in a moderately reverberant situation ($T_{60} = 350$ ms), all algorithms result in approximately accurate TDOAs. However, when reverberation becomes high ($T_{60} = 580$ ms), ML performance decreases significantly, while PHAT and MPHAT retain very good performance.

Figure 4 shows that all algorithms have acceptable performance in moderately noisy situations (SNR = 5 dB). However, when the noise level is increased (SNR = 0 dB), PHAT performance will decrease drastically, while ML and MPHAT method have significantly better performance.

Also, we compared the performance of these methods in joint noisy and reverberant situations (SNR = 5 dB, $T_{60} = 350$ ms) in Figure 5. As seen, the performance of MPHAT method is much improved over the others. This

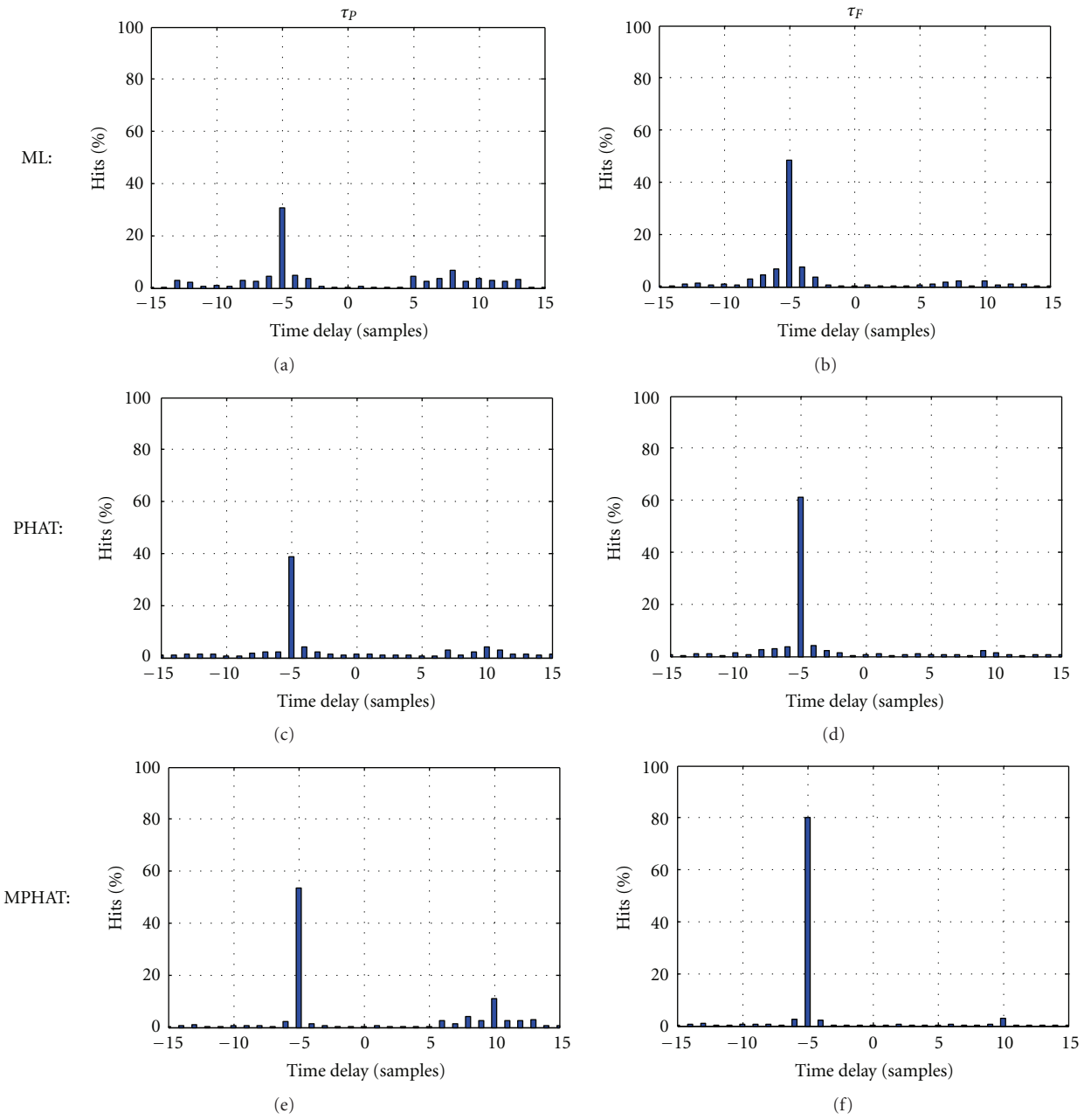

Figure 6: Comparison between (a, c, e) primary TDOA (τ_P) and (b, d, f) final TDOA (τ_F) values estimated using the hybrid algorithm (exact value of TDOA is (-5) samples.).

demonstrates MPHAT robustness against both noise and reverberation.

6.2. Performance of Hybrid Localization Method. As a primary evaluation of the proposed hybrid system, we applied the hybrid TDOA estimation on artificially generated microphone signals and compared the histograms of τ_P and τ_F. To examine the effect of different TDOA estimation techniques, we repeated this evaluation for GCC-ML, -PHAT, and -MPHAT methods. It is noted that the comparisons of this part were done in a reverberant and moderate noisy condition (SNR = 5 dB, T_{60} = 350 ms). The results were drawn in Figure 6 for the case of first microphone pair and the second source position (as a sample case). In each row, the left histogram is for τ_P and the right one is for τ_F. As shown, in all cases, τ_F is more accurate (robust) compared to τ_P. This demonstrates the superiority of the proposed hybrid localization method. The improvement is more obvious in the case of MPHAT (and PHAT).

In the next experiment on the artificially generated data, we performed sound source localization using the SX and SI methods and compared 3D RMSE (root mean square error) values. Table 1 summarizes the comparative results for

different localization methods. As a reference, we also include the RMSE values for the well-known SRP-PHAT method [17] (with a grid size of $0.1 \times 0.1 \times 0.1$ m $(x \times y \times z)$). As shown, we have the following.

(i) MPHAT weighting function outperforms PHAT in all cases.

(ii) While the hybrid localization and outlier removal techniques have improved the accuracy of the sound localizer, the best results were reached by joint hybrid localization and outlier removal.

(iii) The highest localization accuracy was obtained for the middle-center source. As the source-array distance increases, the reverberation increases; this, in turn, degrades the localization accuracy.

(iv) In the case of near source (at (5, 5, and 1.8)), the far-field assumption is clearly violated; this explains poor performance of the TDOA-based localization methods for the near source.

(v) The localization accuracy of the proposed method is of the order of SRP-PHAT accuracy, while requiring lower computational complexity.

7. Experiments on Real Data

We also evaluated the performance of the whole speech source localization system (and proposed modifications) on real data recorded in a sample practical room. Figure 7 shows a schematic representation of the real-data recording room. The room dimension is $5.65 \times 7.34 \times 3.23$ m $(x \times y \times z)$. Considering different environmental noise sources (from fans, PCs, lights, babble noise from outside, etc.), the noise field can be approximated as a diffuse one. The hard surfaces and walls made the environment highly reverberant. Reverberation time of the room is estimated $T_{60} \cong 650$ ms. Data recording was done by means of a microphone array setup that consists of 16 microphones with a spacing of 35 cm. The microphones were attached to the edges of a table.

The speech data was recorded from a male speaker, digitized at 16-bit resolution at $F_s = 16$ kHz. Three marked positions were considered as the speaker standing point; these positions were (1.78, 2.78, and 1.6) (near the microphones), (3.02, 4.38, and 1.6) (middle of room), and (1.78, 5.28, and 1.6) (far from the microphones). In these locations, the average SNR in the reference microphone was about 12.7 dB, 7.1 dB, and 3.2 dB, respectively. At each position, the speaker uttered a predefined text with a time length of about 20 s. The details of recording setup and the microphone placements have been explained in [35].

In Figure 8, we compare performance of GCC methods in the real acoustic environment. This comparison has been done for the data from a near speaker and a far speaker. The advantage of the MPHAT method over the ML and PHAT methods is evident in both near and far cases. As expected, by increasing the distance between speaker and microphones, the reverberation becomes more challenging; consequently, the performance of the ML method is highly degraded. Furthermore, as the distance increases, the SNR at the input decreases. This results in the degradation of PHAT performance. So, the MPHAT superiority is more obvious in the case of a far speaker, where the input signal is highly noisy and reverberant.

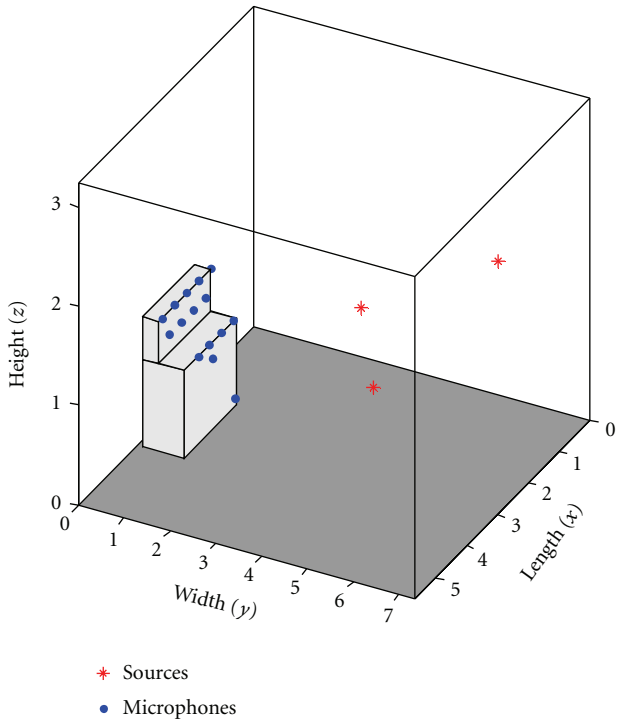

Figure 7: Schematic representation of real-data recording room.

As a final evaluation, we have evaluated the effect of the proposed modifications on the real (practical) data. Table 2 compares the 3D RMSE of the proposed hybrid method with conventional SX and SI methods. Both PHAT and MPHAT methods for TDOA estimation are considered in comparative evaluations. Again, we have also included the RMSE values for SRP-PHAT for reference. As it is shown, we have the following.

(i) Applying the MPHAT technique for TDOA estimation results in more accurate estimation of source location.

(ii) The hybrid localization method improves the performance of both SI and SX methods.

(iii) TDOA outlier removal increases the localization accuracy.

(iv) By applying all proposed modifications (i.e., hybrid SI + outlier remove with MPHAT), we get the best results.

(v) The highest localization accuracy is achieved in the case of the second source position, where the speaker is in the middle of the room and in front of the array.

8. Conclusions

In this paper, we presented and evaluated three novel modifications to improve the performance of TDOA-based

Table 2: Comparison between 3D RMSE of various speech source localization methods on real (practical) data.

Method	GCC method	RMSE (m) for near source	RMSE (m) for middle source	RMSE (m) for far source	Average RMSE (m)
SI	PHAT	2.765	1.864	2.861	2.497
	MPHAT	2.412	1.651	2.437	2.167
SX	PHAT	2.803	1.976	2.915	2.565
	MPHAT	2.519	1.765	2.608	2.297
Hybrid SI	PHAT	1.486	0.867	1.847	1.400
	MPHAT	1.245	0.764	1.608	1.206
Hybrid SX	PHAT	1.515	0.964	1.867	1.449
	MPHAT	1.327	0.881	1.688	1.299
SI + outlier remove	PHAT	1.841	1.216	2.139	1.732
	MPHAT	1.529	0.976	1.962	1.489
SX + outlier remove	PHAT	1.870	1.416	2.224	1.837
	MPHAT	1.651	1.237	2.060	1.649
Hybrid SI + outlier remove	PHAT	1.300	0.851	1.726	1.292
	MPHAT	1.195	0.751	1.589	**1.178**
Hybrid SX + outlier remove	PHAT	1.326	0.902	1.745	1.324
	MPHAT	1.266	0.784	1.652	1.234
SRP-PHAT		1.227	0.742	1.701	1.223

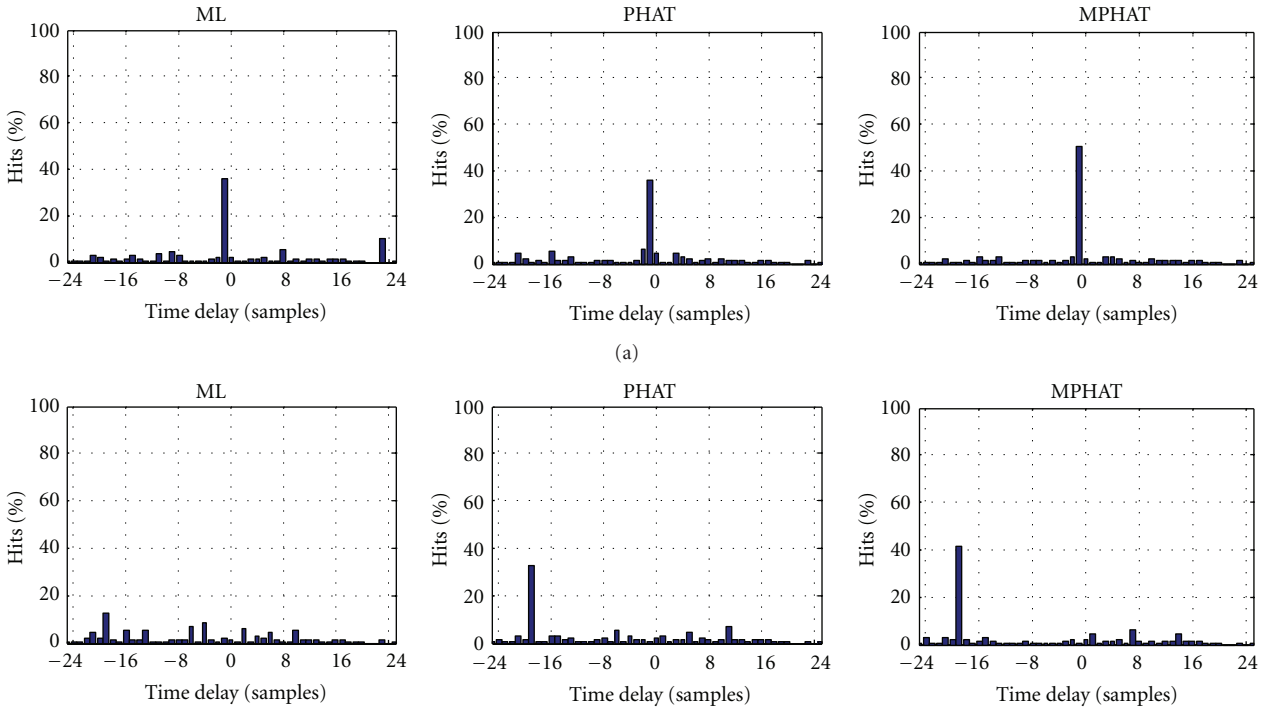

Figure 8: TDE performance in real acoustic environment (a) near speaker (exact value of TDOA is (−1) samples), (b) far speaker (exact value of TDOA is (−19) samples).

3D localization system in a single-speaker scenario. The proposed modifications were MPHAT (instead of PHAT), a hybrid localization method, and TDOA outlier removal.

The GCC-MPHAT method modifies the PHAT weighting function based on an idea borrowed from the generalized spectral subtraction method. The GCC-MPHAT has the advantages of the PHAT method, while it is also robust against noise. In the hybrid algorithm, we use the primary estimation of the source location to modify erroneous TDOA estimates and find true delays. Consequently, a more accurate

estimate of source location is achieved. At the TDOA outlier removal stage, we find erroneous TDOAs and remove them from the source localization process.

Our extensive experiments on both simulated and real (practical) data have demonstrated the capability of the proposed modifications in improvement of a speech source localization system.

Acknowledgment

The authors would like to sincerely thank Philip N. Garner (senior researcher at Idiap) for his constructive comments and corrections that helped to improve the paper.

References

[1] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843–1854, 2002.

[2] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[3] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, no. 2, pp. 91–126, 1997.

[4] J. E. Adcock, M. S. Brandstein, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.

[5] C. Wang and M. S. Brandstein, "Multi-source face tracking with audio and visual data," in *Proceedings of the IEEE 3rd Workshop on Multimedia Signal Process*, pp. 169–174, Copenhagen, Denmark, 1999.

[6] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[7] A. Pentland, "Smart rooms," *Scientific American*, vol. 274, pp. 68–76, 1996.

[8] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 2, no. 3, pp. 209–223, 2001.

[9] F. Ribeiro, C. Zhang, D. A. Florêncio, and D. E. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1781–1792, 2010.

[10] J. Kleban, *Combined acoustic and visual processing for videoconferencing systems*, M.S. thesis, Rutgers University, 2000.

[11] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 187–190, Munich, Germany, 1997.

[12] R. Cutler, Y. Rui, A. Gupta et al., "Distributed meetings: a meeting capture and broadcasting system," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 503–512, Juan-les-Pins, France, 2002.

[13] Y. Rui, D. Florêncio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, pp. 93–96, March 2005.

[14] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. 125–128, April 2007.

[15] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[16] C. Zhang, D. Florêlncio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 2565–2568, April 2008.

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

[18] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 26503, 19 pages, 2006.

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

[20] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[21] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[22] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 71–74, October 2001.

[23] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "Practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[24] J. H. DiBiase, H. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., pp. 131–154, Springer, New York, NY, USA, 2001.

[25] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[26] M. S. Brandstein and H. F. Silverman, "Robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 375–378, Munich, Germany, 1997.

[27] S. Valaee and P. Kabal, "Wideband array processing using a two-sided correlation transformation," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 160–172, 1995.

[28] Y. Rui and D. Florêncio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. 133–136, May 2004.

[29] Y. Rui and D. Florêncio, "New direct approaches to robust sound source localization," in *Proceedings of IEEE International Conference on Multimedia & Expo*, 2003.

[30] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2914–2919, 1999.

[31] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[32] J. R. Deller, H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2000.

[33] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[34] E. E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, pp. 1321–1324, Philadelphia, PA, USA, October 1996.

[35] H. Momenzadeh, *Speech source localization using microphone arrays*, M.S. thesis, Electrical Engineering Department, Yazd University, Yazd, Iran, 2008.

[36] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[37] J. H. Garofolo, L. F. Lamel, W. M. Fisher et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, PA, USA, 1993.