EURASIP Journal on Applied Signal Processing 2005:1, 25–36

Performance of GCC- and AMDF-Based Time-Delay Estimation in Practical Reverberant Environments

Jingdong Chen
Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA
Email: jingdong@research.bell-labs.com

Jacob Benesty
INRS-EMT, Université du Québec, 800 de la Gauchetière Ouest, Suite 6900, Montréal, Québec, Canada H5A 1K6
Email: benesty@inrs-emt.uquebec.ca

Yiteng (Arden) Huang
Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA
Email: arden@research.bell-labs.com

Received 8 December 2003; Revised 8 June 2004

Recently, there has been an increased interest in the use of the time-delay estimation (TDE) technique to locate and track acoustic sources in a reverberant environment. Typically, the delay estimate is obtained through identifying the extremum of the generalized cross-correlation (GCC) function or the average magnitude difference function (AMDF). These estimators are well studied and their statistical performance is well understood for single-path propagation situations. However, fewer efforts have been reported to show their performance behavior in real reverberation conditions. This paper reexamines the GCC- and AMDF-based TDE techniques in real room reverberant and noisy environments. Our contribution is threefold. First, we propose a weighted cross-correlation (WCC) estimator in which the GCC function is weighted by the reciprocal of AMDF. This new method can sharpen the peak of the GCC function, which corresponds to the true time delay and thus leads to a better estimation performance as compared to the conventional GCC estimator. Second, we propose a modified version of the AMDF (MAMDF) estimator in which the delay is determined by jointly considering the AMDF and the average magnitude sum function (AMSF). Third, we compare the performance of the GCC, AMDF, WCC, and MAMDF estimators in real reverberant and noisy environments. It is shown that the AMDF estimator can yield better performance in favorable noise conditions and is slightly more resilient to reverberation than the GCC method. The GCC approach, however, is found to outperform the AMDF method in strong noisy environments. Weighting the correlation function by the reciprocal of AMDF can improve the performance of the GCC estimator in reverberation conditions, yet its improvement in noisy environments is limited. The MAMDF algorithm can enhance the AMDF estimator in both reverberant and noisy environments.

Keywords and phrases: time-delay estimation, generalized cross-correlation function, average magnitude difference function, average magnitude sum function.

1. INTRODUCTION

A microphone array, which consists of a set of microphones that are spatially distributed at known locations with reference to a common point, has the ability to reinforce a desired signal from the look direction while suppressing undesired signals such as noise from other directions. This feature impels the increasing use of microphone arrays in such situations as hands-free speech communications where a system operates under strong noise and reverberation conditions. In the microphone array system, the most crucial issue is to measure the time difference of arrival (TDOA) between two microphone signals since such a time difference often serves as the basis for beamforming and the estimation of direction of arrival (DOA).

Extensive work has been reported for determining the TDOA between two signals [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. One typical time-delay estimation (TDE) technique is the generalized cross-correlation (GCC) method [1, 2, 3, 4, 5, 6, 7, 9, 16, 17, 18, 19, 20] in which the delay estimate is obtained as the time lag that maximizes the GCC function between two microphone signals. The measured time delay is an integral multiple of the sampling period. In other words, the time-delay resolution depends on

the sampling period but is not limited to it. A finer resolution can be acquired in light of an interpolation between consecutive samples of the GCC function when necessary [17, 21].

Another widely used traditional TDE technique relies on the identification of the minimum of the average magnitude difference function (AMDF) between two studied signals [7, 17]. Similarly, an interpolation may be employed to refine the delay estimate.

Both the GCC and AMDF methods are formulated based on the ideal propagation model where no multipath effect is taken into account. They perform fairly well in single-path propagation situation, but suffer performance degradation when multipath or reverberation effects are present. Recently, several advanced TDE techniques were proposed [22, 23, 24, 25, 26], which can better deal with reverberation. However, the GCC and AMDF techniques are still preferred by many engineers and are widely used in various systems for their computational efficiency and simplicity to implement.

In single-path propagation situations, the GCC and AMDF algorithms have been extensively investigated and their statistical performance is well understood [8, 17]. However, multipath propagation and reverberation effects are more common in practice. Unfortunately, fewer efforts have been reported to show the performance behavior of these algorithms in practical reverberant environments. An early study [18] examined the effects of the simulated room reverberation on the performance of the GCC approach to TDE. It was shown that the performance of this algorithm severely deteriorated as the reverberation time increased. If the acceptable level of the percentage of anomalous estimates is set to be 10%, the maximum likelihood (ML) GCC method cannot be used reliably when the reverberation time is greater than 0.18 seconds, which is quite common in hands-free communication applications.

In this paper, we reexamine the GCC and AMDF algorithms in real room reverberant and noisy environments. We also show that these estimators can be improved by incorporating some other information. Our contribution is threefold. First of all, inspired by the weighted autocorrelation method, which has been recently proposed for pitch tracking [27], we propose a weighted cross-correlation (WCC) estimator in which the GCC function is weighted by the reciprocal of AMDF. This new method can sharpen the peak of the GCC function, which corresponds to the true time delay and thus leads to a better estimation performance as compared to the conventional GCC estimator. Secondly, we propose a modified version of the AMDF (MAMDF) estimator in which the delay is determined by jointly considering the AMDF and the average magnitude sum function (AMSF). We show that the combination of AMDF and AMSF can enhance the performance of the AMDF estimator. Thirdly, the GCC, AMDF, WCC, and MAMDF estimators are evaluated with data collected in the Varechoic Chamber at Bell Laboratories. On one hand, this evaluation will find which algorithm can produce better TDE in practical situations. On the other hand, such a comparative study can offer insight into the range of TDE techniques that can be employed in practical room reverberation conditions. The experimental results justify that proper manipulation of the GCC function and AMDF can make the TDE techniques more robust with respect to reverberation.

2. THE TDE PROBLEM

2.1. Signal model

A widely used signal model for the TDE problem is given by

$$\begin{aligned} x_1(n) &= s(n-t) + w_1(n), \\ x_2(n) &= \alpha s(n-t-\tau) + w_2(n), \end{aligned} \tag{1}$$

where $x_m(n)$, $m = 1, 2$, denotes the output signal of the mth microphone, α is an attenuation factor due to the propagation effect, t is the propagation time from the unknown source $s(n)$ to Microphone 1, $w_m(n)$ is an additive noise signal at microphone m, and the parameter τ is the true time delay between two microphones. We assume that $w_m(n)$ is a (real) zero-mean stationary random process which is uncorrelated with both $s(n)$ and the noise signal from the other microphone. It is also assumed that $s(n)$ is reasonably broadband. This model reflects an ideal situation in which the signal propagation from the source to each microphone occurs along a single direct path in a nondispersive medium. The TDE problem is to find an estimate $\hat{\tau}$ of the true delay τ, using a finite set of observation samples of $x_1(n)$ and $x_2(n)$. The signal $x_1(n)$ will be also called the *reference* signal.

2.2. TDE principles

The TDE techniques investigated in this paper are based on searching for the extremum of the GCC or some other statistical cost functions of the observed signals. Particularly, we consider the following estimators.

2.2.1. The generalized cross-correlation estimator

The GCC method, proposed by Knapp and Carter in 1976 [1], is the most popular technique for TDE, in which the time-delay estimate is obtained as follows:

$$\hat{\tau}_{\mathrm{GCC}} = \arg\max_n \hat{\Psi}_{\mathrm{GCC}}(n), \tag{2}$$

where

$$\hat{\Psi}_{\mathrm{GCC}}(n) = \sum_{k=0}^{N-1} \Phi(k) S_{x_1x_2}(k) e^{j(2\pi nk/N)} \tag{3}$$

is the GCC function, $S_{x_1x_2}(k) = E\{X_1(k)X_2^*(k)\}$ is the cross spectrum, $E\{\cdot\}$ and $(\cdot)^*$ stand, respectively, for the expectation and complex conjugate operator, $X_m(k)$ is the discrete Fourier transform of the signal $x_m(n)$, $\Phi(k)$ is a weighting function (sometimes called a *prefilter*), and N denotes the number of observation samples during the observation interval.

The weighting function $\Phi(k)$ plays an important role in controlling the TDE performance. It is chosen according to some criterion. Commonly used weighting functions include unit weighting (the classical cross-correlation method),

the smoothed coherence transform (SCOT) [3], the Roth processor [4], the Echart filter, the phase transform (PHAT), the ML processor [1], the Hassab-Boucher transform [5], and so forth. Some of these are optimal in the sense that the estimation variance can achieve the Cramèr-Rao lower bound (CRLB). Others are suboptimal but possess special properties, as for example the PHAT algorithm [1, 12], where the weighting function is chosen as $\Phi(k) = 1/|S_{x_1x_2}(k)|$. Substituting $\Phi_{\mathrm{PHAT}}(k)$ into (3) and neglecting the noise effects, one can readily derive that the weighted cross spectrum is free from the source signal and depends only on the channel responses. Consequently, the PHAT algorithm performs more consistently than many other GCC members when the characteristics of the source signal vary in time. Hence, this weighting function is adopted in this research.

2.2.2. The AMDF estimator

The AMDF between two studied signals is described by

$$\hat{\Psi}_{\mathrm{AMDF}}(n) = \frac{1}{N}\sum_{i=0}^{N-1} |x_1(i) - x_2(i+n)|. \tag{4}$$

The delay estimate, based on AMDF, is given by

$$\hat{\tau}_{\mathrm{AMDF}} = \arg\min_n \hat{\Psi}_{\mathrm{AMDF}}(n). \tag{5}$$

The AMDF approach has been used for TDE and pitch tracking for decades [7, 28]. The preference of employing AMDF over GCC for TDE is mainly due to the following facts. First, the performance of the AMDF estimator in favorable noise conditions is better than that of the GCC method as reported in [17]. Second, the AMDF technique has relatively low computational cost as no multiplications are involved in the estimation of AMDF although the computational burden may not be a big concern with today's computers.

Assuming that the signal $s(t)$ can be modeled as a zero-mean Gaussian process, from the invariance technique [29, 30], we can derive the expectation of the AMDF as follows (see the appendix):

$$\begin{aligned} E\{\hat{\Psi}_{\mathrm{AMDF}}(n)\} &= E\left\{\frac{1}{N}\sum_{i=0}^{N-1} |x_1(i) - x_2(i+n)|\right\} \\ &= \sqrt{\frac{2}{\pi}[e_{x_1} + e_{x_2} - 2R_{x_1x_2}(n)]}, \end{aligned} \tag{6}$$

where $e_{x_m} = E\{x_m^2(n)\}$ represents the energy of the observation signal $x_m(n)$, and $R_{x_1x_2}(n) = E\{x_1(i)x_2(i+n)\}$ is the direct cross-correlation function between $x_1(n)$ and $x_2(n)$. Inspection of (6) shows that the magnitude of the principle minimum of the AMDF is essentially influenced by the intensity variation and the background noise of the observation signal. This indicates that the AMDF method may be sensitive to the background noise. As a matter of fact, many reported experiments have confirmed that the AMDF estimator is less robust with respect to noise than the GCC method [31]. Equation (6) also suggests that the performance of AMDF can be affected by the source signal, like in the conventional cross-correlation approach. This problem, however, can be alleviated by prewhitening the observation signal before the estimation of AMDF.

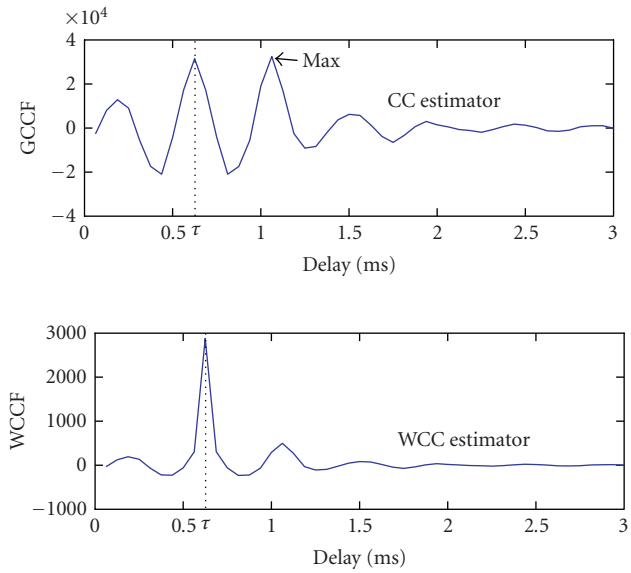


Figure 1: GCC function and WCC function in a moderately reverberant but noise-free condition.

2.2.3. The weighted cross-correlation estimator

The maximum of the GCC function does not necessarily occur at the true time delay as also pointed out in [6, 17]. This is mainly due to the delayed version of the signal containing new samples for different time lags. Figure 1 shows one estimated GCC function between two microphone signals in a moderately reverberant but noise-free condition. As can be seen, the GCC function has two large peaks. One appears at 0.625 milliseconds which corresponds to the true time delay, and another one appears at 1.125 milliseconds. Unfortunately, the maximum peak appears at 1.125 milliseconds which leads to an estimation failure. In comparison, AMDF generally produces more accurate estimates. However, as mentioned before, the primary disadvantage of the AMDF approach is the lack of robustness with respect to noise.

To achieve a good compromise between the robustness of the GCC method and the accuracy of the AMDF approach, we propose a heuristic method by weighting the GCC function with the reciprocal of AMDF, which may not necessarily be the optimum way to combine both, but will certainly improve TDE performance, as shown in Section 3. The resulting estimator is described by

$$\hat{\tau}_{\mathrm{WCC}} = \arg\max_n \hat{\Psi}_{\mathrm{WCC}}(n), \tag{7}$$

where

$$\hat{\Psi}_{\mathrm{WCC}}(n) = \frac{\hat{\Psi}_{\mathrm{GCC}}(n)}{\hat{\Psi}_{\mathrm{AMDF}}(n) + \varepsilon}, \tag{8}$$

and ε is a small positive number to prevent division overflow. Figure 1 also shows the weighted GCC function (WCCF). In this case, picking the maximum of the WCCF will lead to a correct estimate.

2.2.4. The modified AMDF estimator

The AMDF specifies the synchrony between the reference signal and a delayed version of this signal. In the noise-free condition, the AMDF yields its minimum when the two signals are synchronized. Synchrony can also be described by the AMSF defined as follows:

$$\hat{\Psi}_{\mathrm{AMSF}}(n) = \frac{1}{N}\sum_{i=0}^{N-1} \left| x_1(i) + x_2(i+n) \right|. \tag{9}$$

If both signal and noise are assumed to be uncorrelated Gaussian processes, when two microphone signals are synchronized, the AMDF will reach its minimum while the AMSF will approach its maximum. We can further show that the correlation coefficients between AMDF and AMSF are approximately zero (see the appendix). This suggests that AMDF and AMSF contain supplementary information though both of them measure the same synchrony between two studied signals. Hence, we can expect to improve the TDE performance by combining AMDF and AMSF. The resulting new estimator is called the MAMDF method described as follows:

$$\hat{\tau}_{\mathrm{MAMDF}} = \arg\min_{n} \hat{\Psi}_{\mathrm{MAMDF}}(n), \tag{10}$$

where

$$\hat{\Psi}_{\mathrm{MAMDF}}(n) = \frac{\hat{\Psi}_{\mathrm{AMDF}}(n)}{\hat{\Psi}_{\mathrm{AMSF}}(n) + \epsilon}, \tag{11}$$

and again ϵ is a fixed positive number similar to ε in (8) to prevent division overflow.

2.3. Implementation

From (2), (5), (7), and (10), one can readily see that the estimated time delay is an integral multiple of the sampling period. This resolution is usually not sufficient for many microphone array applications. Much effort has been devoted to solving this problem [17, 21, 32]. Among these, interpolation around the detected peaks of the cost function is the simplest yet most effective way to refine the TDE. Here we employ the 3-point Lagrange's method to improve the resolution such that the estimated time delay can be a fraction of the sampling period. The implementation procedure of the developed estimators is summarized below.

(i) Partition the observation signal sequences $x_1(n)$ and $x_2(n)$ into nonoverlapping frames with a frame width of 128 milliseconds. For all experiments, microphone signal is digitized with a sampling frequency of 16 kHz. A Hamming window of length 128 milliseconds is applied for a better spectral estimate.

(ii) To reduce the dependence of the TDE on the structure of the source signal, we prewhiten signals $x_m(n)$ before starting the TDE. The prewhitening process is performed in the frequency domain and the FFT algorithm is used for efficiency, that is, $\mathrm{IFFT}\{\mathrm{FFT}[x_m(n)]/|\mathrm{FFT}[x_m(n)]|\}$.

(iii) Compute the cost function defined in (2), (4), (8), and (11).

(iv) Search for the extremum of the cost function and the corresponding lag time is denoted as $\hat{n}_{\mathrm{ext}}$.

(v) Interpolate 4 points between $\hat{n}_{\mathrm{ext}} - 1$ and $\hat{n}_{\mathrm{ext}}$ and another 4 points between $\hat{n}_{\mathrm{ext}}$ and $\hat{n}_{\mathrm{ext}} + 1$, using the 3-point Lagrange's method (the AMDF-based cost functions are squared before interpolating [17]). Then search the extremum of the interpolated cost function. The corresponding peak (valley) position relative to $\hat{n}_{\mathrm{ext}}$ is denoted as $\hat{\Delta}$ ($\hat{\Delta}$ is negative when the extremum is located in the left-hand side of $\hat{n}_{\mathrm{ext}}$, and is positive when the extremum is located in the right-hand side of $\hat{n}_{\mathrm{ext}}$).

(vi) The time-delay estimate is obtained as $\hat{\tau} = \hat{n}_{\mathrm{ext}} + \hat{\Delta}$.

3. PERFORMANCE EVALUATION

In general, the performance of the GCC, AMDF, WCC, and MAMDF techniques is affected by the interpolation and finite width of the estimation window. Apart from these systematic factors, the accuracy of the estimates is substantially impaired by noise and reverberation. In this section, we present the results of the experiments to investigate the statistical performance of TDE in real reverberant and noisy environments.

Following [6, 18], we distinguish an estimate as either an *anomaly* or a *nonanomaly* according to its absolute error. If the absolute error $|\hat{\tau}_i - \tau| > T_c/2$, the estimate is identified as an anomaly; otherwise, it is declared as a nonanomaly, where τ and $\hat{\tau}_i$ are the true delay and ith delay estimate, respectively, and T_c is the signal correlation time. To compute T_c, we divide the source signal into short frames with a frame size of 128 milliseconds. A short-time autocorrelation function is estimated from each frame of data. The long-term average autocorrelation function is then computed as the arithmetic average of the short-time autocorrelation functions. T_c is computed as the 3 dB width of the main lobe of the long-time average autocorrelation function (in our experiment, the calculated T_c is equal to 4.3 samples). We evaluate the TDE performance in terms of the percentage of anomalous estimates over the total estimates, the bias, and the standard deviation of the nonanomalous estimates.

3.1. Experimental setup

Experiments were carried out in the Varechoic Chamber which is a unique facility at Bell Laboratories. The chamber is a $6.7 \times 6.1 \times 2.9$ m room whose surfaces are covered by a total of 369 active panels which can be controlled digitally. Each panel consists of two perforated sheets. When the holes in the

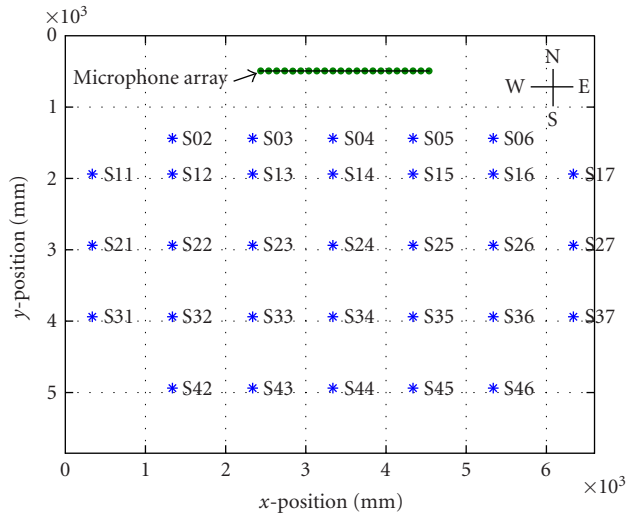


FIGURE 2: Layout of the microphone array and source positions in the varechoic chamber.

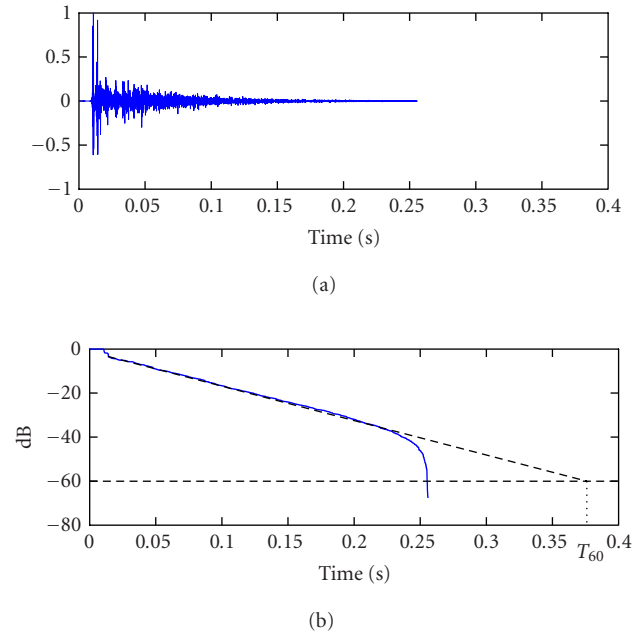


FIGURE 3: (a) Measured impulse response when 30% of the panels are closed. (b) Backward integrated impulse response.

sheets are aligned, absorbing material behind the sheets will be exposed to the sound field, whereas a highly reflective surface can be formed if the holes are shifted to misalignment. Combination of open and closed panels can produce 2^{369} different acoustic environments where the 60 dB reverberation time T_{60} can change from 0.2 to almost 1 second (refer to [33, 34] for more details).

A linear microphone array which consists of 22 omnidirectional Panasonic WM-61A microphones was mounted at the distance of 500 mm from the north wall of the chamber and approximately at the center of the wall. The 22 microphones are uniformly distributed along an aluminum rod whose diameter is 1 cm. The spacing between adjacent microphones is 10 cm. The source signal is played by a Cabasse Baltic Murale loudspeaker in 46 different positions. An illustration of this setup is shown in Figure 2.

In order to reduce the reflections from the north wall, the wall behind the array is covered by a 3-inch-thick fiber class pillow which has a rectangle shape of 3230×750 mm. Its lower edge is 90 mm above the floor and the left edge 1950 mm from the west wall of the chamber. During the experiment, the chamber was not completely empty; objects such as chairs, loudspeakers, and unused equipments were left in the room. Also the inner door of the room in the east corner of the south wall was kept open during the course of the experiment.

For the purpose of data reusability, the impulse response from each source location to each microphone was measured [34]. The measurement of the impulse responses was performed using the built-in measurement tool of the Huron Lake system [34]. A 65536-point long logarithmic sweep signal digitized at a sampling rate of 48 kHz was used as the excitation signal. From each source location to each microphone, the excitation is played and recorded. An estimate of the transfer function is obtained by a spectral division between the original source excitation and the recorded microphone signal. We show in Figure 3 an impulse response measured from Microphone 22 when 30% of the panels are closed and the loudspeaker is placed at the position S21 (shown in Figure 2). Also shown in Figure 3 is the backward integrated decay curve of the measured impulse response. One can see from this decay curve that the reverberation time T_{60} is approximately 0.37 second.

The observation signal is obtained by convolution of the recorded speech with the measured impulse response, and then adding noise to the results. Two types of noise have been used in the experiments: the computer generated pseudo Gaussian noise and a noise signal recorded from a New York Stock Exchanging (NYSE) room. The NYSE noise consists of sounds from various sources such as speakers, telephone rings, electric fans, and so forth. Figure 4 plots the first two seconds of the NYSE noise and its spectrogram, from which we can see the changing characteristics of such noise.

3.2. Experimental results

As pointed out before, the microphone output signal is computed by convolving a 4-minute speech from a female speaker with the corresponding measured impulse response and then adding zero-mean noise to the results for a given signal-to-noise ratio (SNR). This output signal is then segmented into nonoverlapping frames with a frame width of 128 milliseconds. For each frame, a time-delay estimate is obtained by estimators described in (2), (5), (7), and (10). The array consists of 22 microphones in total, so we have $C_{22}^2 = 231$ microphone pairs. In our experiment, however, we choose Microphone 1 as a reference and only measure the time delay of each microphone signal relative

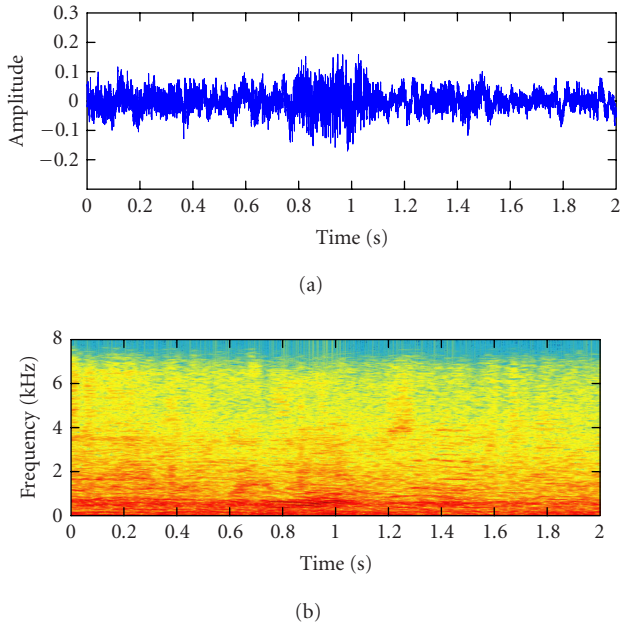


Figure 4: The NYSE noise: (a) waveform of the first two seconds of the noise; (b) the spectrogram of the signal shown in (a).

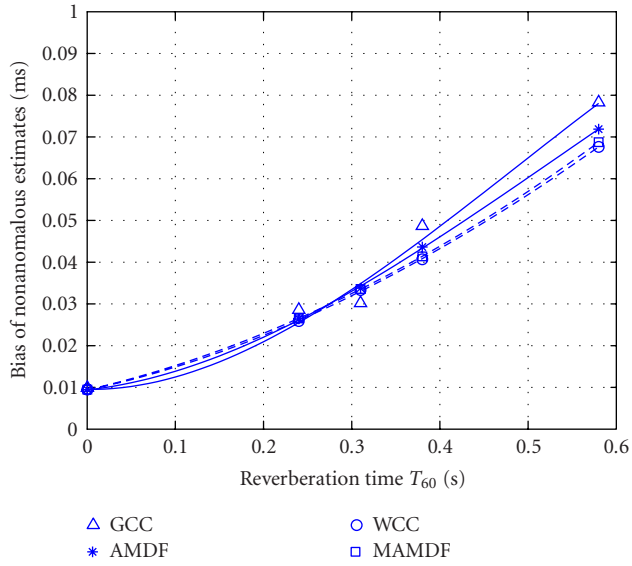


Figure 6: Bias of nonanomalous time-delay estimates versus T_{60} among the GCC, AMDF, WCC, and MAMDF algorithms at SNR = 25 dB. Microphones 1 and 3 are used in the experiment. The fitting curve is a third-order polynomial.

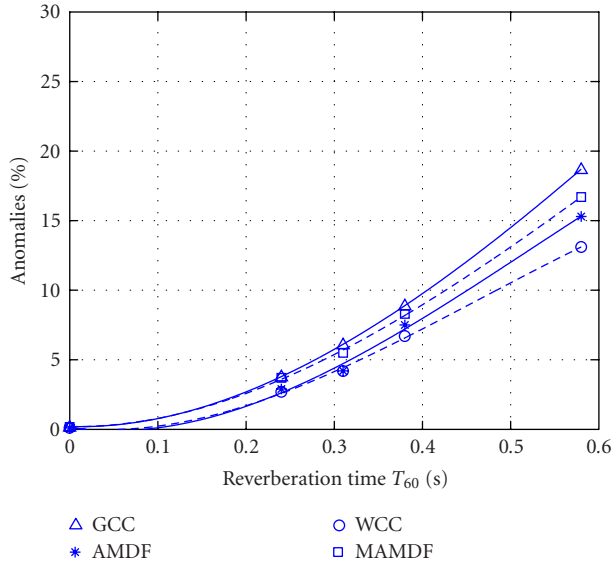


Figure 5: Percentage of anomalous time-delay estimates versus T_{60} among the GCC, AMDF, WCC, and MAMDF algorithms at SNR = 25 dB. Microphones 1 and 3 are used in the experiment. The fitting curve is a third-order polynomial.

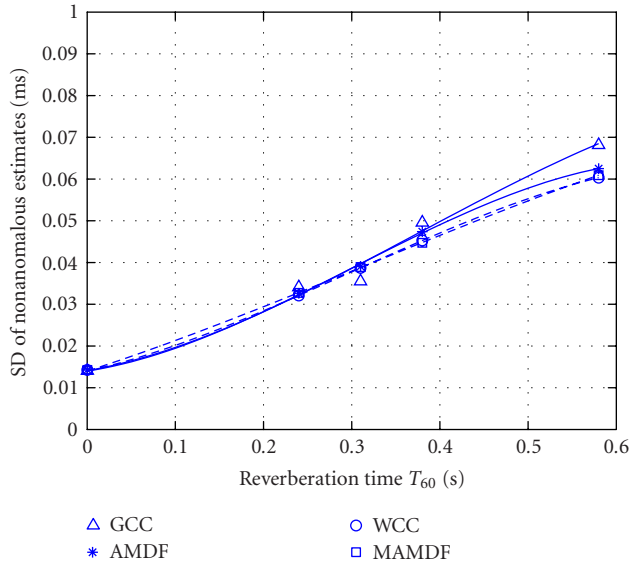


Figure 7: Standard deviation (SD) of the nonanomalous time-delay estimates versus T_{60} among the GCC, AMDF, WCC, and MAMDF algorithms at SNR = 25 dB. Microphones 1 and 3 are used in the experiment. The fitting curve is a third-order polynomial.

to the signal of Microphone 1. For a specific reverberation and noise condition, we have 21 (microphone pairs) × 46 (source positions) × 240 (seconds)/0.128(frame length) ≈ 1.8 million time-delay estimates. If taking into account the different reverberation and SNR conditions, we have in total 5 (different reverberation time) × 10 (SNR) × 1.8 M ≈ 90 million delay estimates. For the sake of brevity, we selected some representative results to present here.

3.2.1. *TDE performance versus reverberation time*

In the first experiment, we analyze the TDE performance versus reverberation time. To do so, we assume that the background noise is white Gaussian noise, and SNR is relatively high, say SNR = 25 dB. The source position varies from S02 to S46, as shown in Figure 2, whereas the microphone pair is a fixed one (Microphones 1 and 3 are used in this experiment). Figures 5, 6, and 7 plot, respectively,

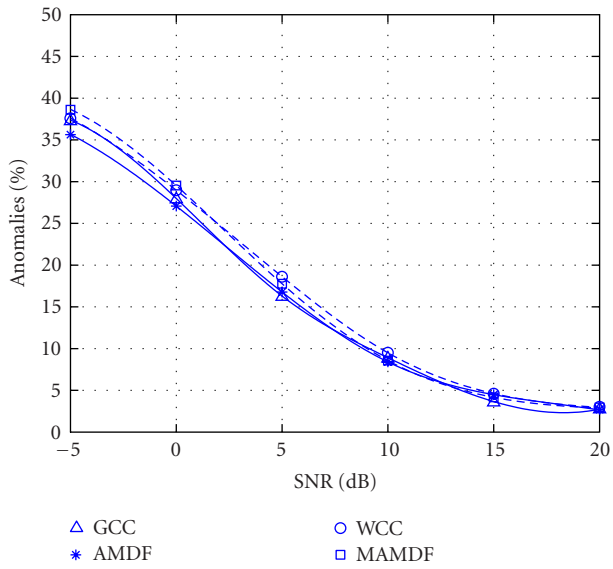


Figure 8: Percentage of anomalous time-delay estimates versus SNR in white Gaussian noise when $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31.

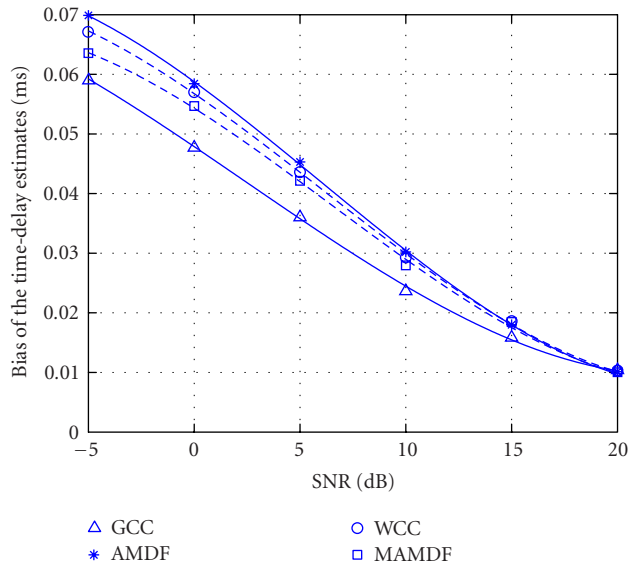


Figure 9: Bias of nonanomalous time-delay estimates versus SNR in white Gaussian noise when $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31. The fitting curve is a third-order polynomial.

the average percentage of anomalies, and the bias and standard deviation of the nonanomalous estimates, all as a function of the reverberation time T_{60}.

Obviously, the percentage of the anomalous estimates of all estimators increases with the reverberation time. As the reverberation time increases, more reflected signals with different delay will reach the microphone sensor and as a result, the erroneous peaks (for GCC and WCC) or valleys (for AMDF and MAMDF) of the cost function increase, which leads to more mistakes in extremum searching, and eventually leads to more anomalous time-delay estimates. Compared with the GCC approach, the AMDF estimator exhibits less anomalies when the reverberation time increases. This shows the advantage of AMDF for TDE. It is interesting to note from Figure 5 that weighting the GCC function by the reciprocal of AMDF can reduce the probability of anomalous estimates. However, if the acceptable level of anomalous estimates is set to 10%, according to Figure 5, all the studied methods cannot be used reliably when $T_{60} >$ 0.5 second.

From Figures 6 and 7, it can be seen that both the bias and standard deviation of the nonanomalous estimates degrades severely as the reverberation time increases. The four estimators exhibit a similar bias and standard deviation in light reverberation conditions. In highly reverberant environments, the GCC estimator has a slightly worse performance.

3.2.2. TDE performance versus SNR

In the above experiment, we show the impact of reverberation time on TDE performance, where a very high SNR is assumed. In a practical situation, however, the TDE has to deal with both reverberation and noise. The second experiment is to evaluate the TDE performance in both simulated

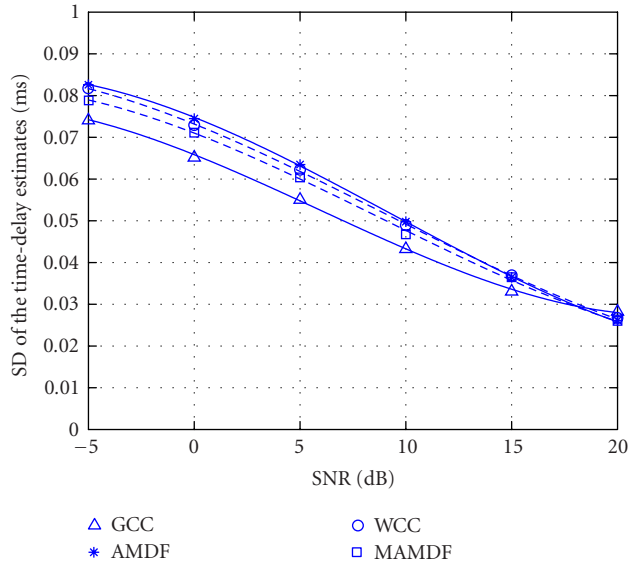


Figure 10: Standard deviation (SD) of nonanomalous time-delay estimates versus SNR in white Gaussian noise when $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31. The fitting curve is a third-order polynomial.

(white Gaussian) and real (NYSE) noisy environments, where we assume a moderate reverberation, say $T_{60} = 0.31$ second. The results in Gaussian noise are presented in Figures 8, 9, and 10, and the results in NYSE noise are portrayed in Figures 11, 12, and 13. We found that the probability of anomalies reduces as SNR increases, and the four TDE methods have a similar percentage of anomalies in both noise conditions.

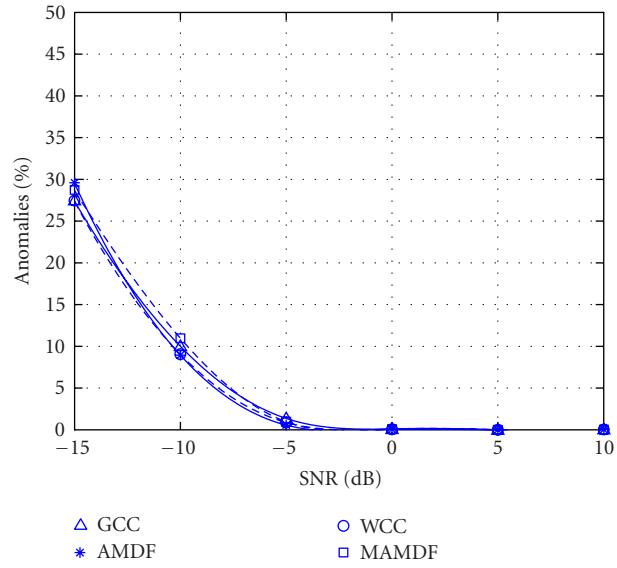


Figure 11: Percentage of anomalous time-delay estimates versus SNR in the NYSE noise when $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31. The fitting curve is a third-order polynomial.

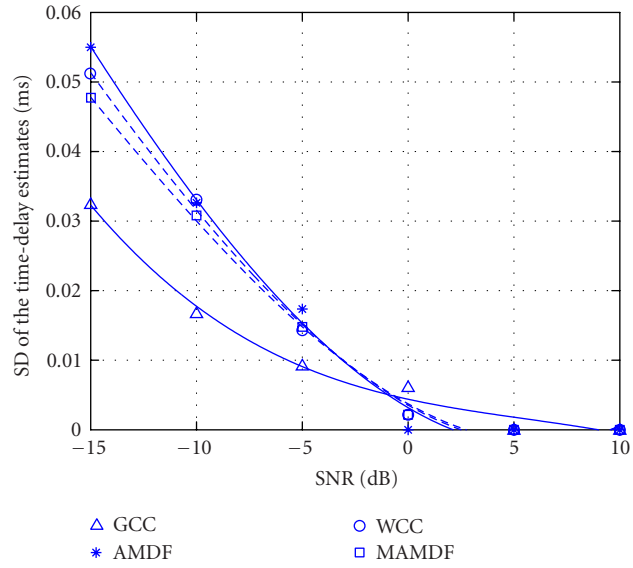


Figure 13: Standard deviation (SD) of nonanomalous time-delay estimates versus SNR in the NYSE noisy environments. $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31.

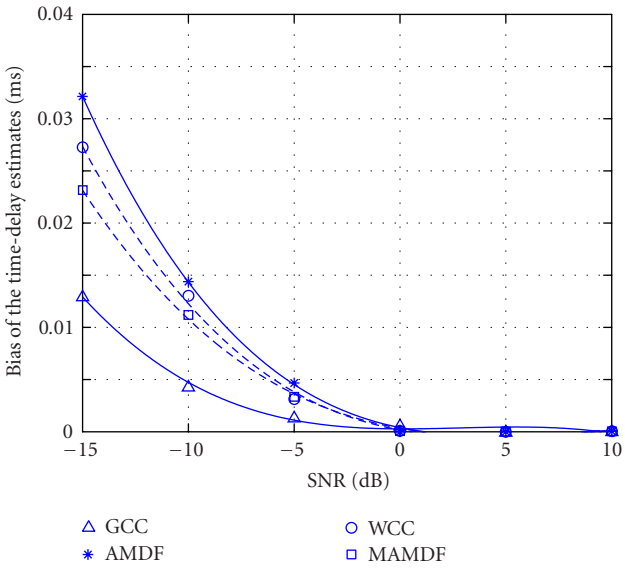


Figure 12: Bias of nonanomalous time-delay estimates versus SNR in the NYSE noisy environments when $T_{60} = 0.31$ second. Microphones 1 and 3 are used in the experiment. The source is in S31. The fitting curve is a third-order polynomial.

From Figures 9 and 12, we see that in high SNR conditions, four estimators have almost identical estimation bias. In lower SNR situations, the AMDF estimator has a higher bias. This is inconsistent with the result reported in [17], in which the AMDF is shown to have much lower bias than the GCC method in high SNR conditions and has almost the same bias as GCC in strong noise environments. We attribute the difference to three factors. Firstly, the experiment in [17] was performed only in high noise conditions where reverberation is absent. Secondly, the GCC estimator tested in [17] is based on the direct cross-correlation function rather than the GCC function. Finally, the results reported in [17] did not distinguish between estimates as anomalies and nonanomalies.

In Figures 10 and 13, it can be seen that in high SNR conditions, the GCC method has a lightly higher deviation than the other three estimator. When SNR becomes lower, however, the GCC estimator shows a smaller deviation, indicating the robustness of the GCC method with respect to noise. Weighting the AMDF function with the reciprocal of the AMSF function can enhance the performance of the AMDF estimator. However, the performance of the WCC approach in noisy conditions is basically a tradeoff between the AMDF and GCC methods.

3.2.3. *Other experiments*

Additional experiments were performed including changing the source locations and using different microphone pairs. Figures 14, 15, and 16 plot the statistical performance as a function of loudspeaker position shown in Figure 2. One can see that percentage of anomalies does not vary much when the source is moved from one position to another. However, the bias and standard deviation fluctuate a lot as the source location varies. According to our experience, though in general the change of the reverberation time T_{60} is negligible, the echo structure varies appreciably as the source position moves. This will eventually lead to fluctuation of the bias and standard deviation of the time-delay estimate.

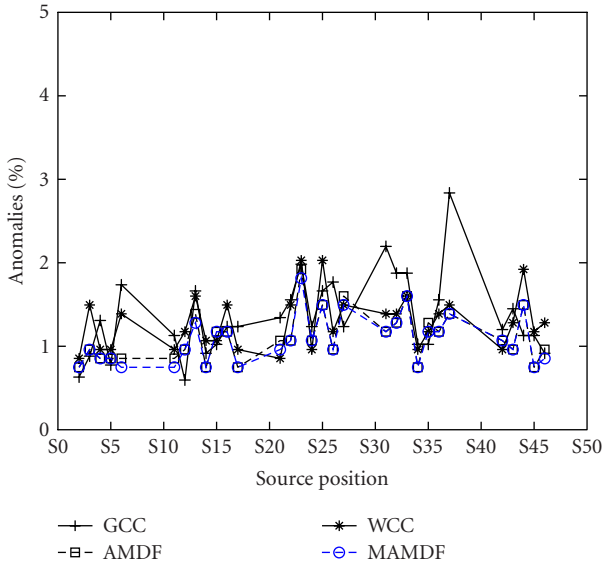


Figure 14: Percentage of anomalous time-delay estimates versus source location among the GCC, AMDF, WCC, and MAMDF algorithms when $T_{60} = 0.31$ second and SNR = 30 dB. Microphones 1 and 2 are used in the experiment.

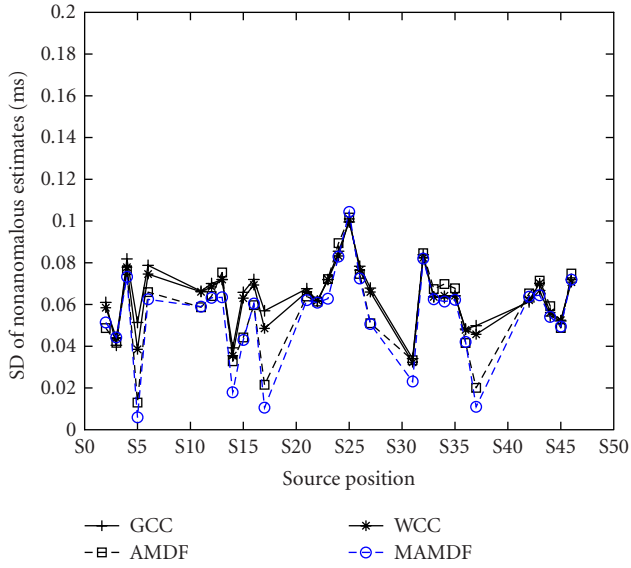


Figure 16: Standard deviation (SD) of the nonanomalous time-delay estimates versus source location among the GCC, AMDF, WCC, and MAMDF algorithms when $T_{60} = 0.31$ second and SNR = 30 dB. Microphones 1 and 2 are used in the experiment.

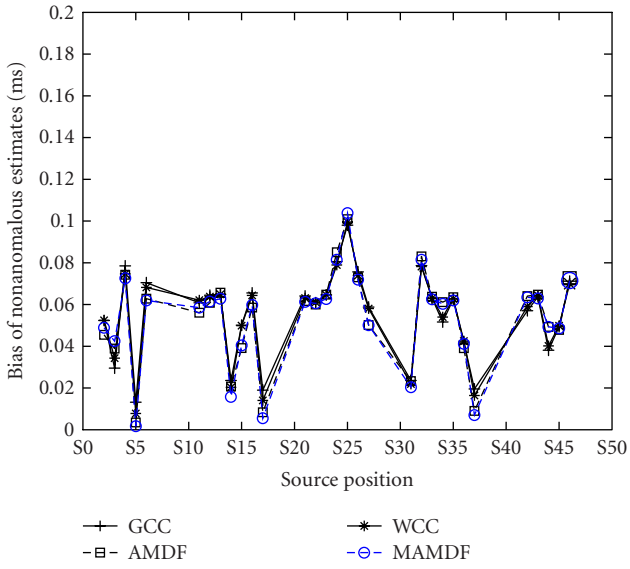


Figure 15: Bias of the nonanomalous time-delay estimates versus source location among the GCC, AMDF, WCC, and MAMDF algorithms when $T_{60} = 0.31$ second and SNR = 30 dB. Microphones 1 and 2 are used in the experiment.

It is interesting to note that biases of the four investigated estimators are almost identical, whereas the percentage of anomalies and standard deviation of nonanomalous estimates of the GCC method is slightly higher than the AMDF-based methods. This is consistent with the observation from the previous experiments since this experiment is carried out at very high SNR and moderate reverberation environments.

We also varied the microphone pairs while keeping the source position fixed. Quite similar qualitative behavior as above was observed.

4. CONCLUSION

This paper addressed the TDE problem in real reverberant and noisy environments. We have proposed two new time-delay estimators. One is the weighted cross-correlation method in which the GCC function is weighted by the reciprocal of AMDF. This weighting process can sharpen the desired peak and suppress the other peaks in the GCC function, hence leading to more accurate time-delay estimates. The other proposed estimator is the modified version of the AMDF method in which the AMDF is weighted by the inverse AMSF—another function that can measure the synchrony between two signals. This approach is seen to exhibit a superior performance to the AMDF method in both high reverberation and high noise conditions.

We have evaluated the GCC, WCC, AMDF, and MAMDF approaches in both room reverberant and noisy environments. In general, it is observed that the cross-correlation-based method exhibits a slightly higher percentage of anomalous estimates than the AMDF-based estimator in favorable noise conditions. However, the GCC-based approaches are more resilient to strong noise.

APPENDIX

The expectation of the AMDF defined in (4) can be written as follows:

$$E\{\hat{\Psi}_{\mathrm{AMDF}}(n)\} = E\left\{\frac{1}{N}\sum_{i=0}^{N-1}|x_1(i)-x_2(i+n)|\right\} = \frac{1}{N}\sum_{i=0}^{N-1}E\{|x_1(i)-x_2(i+n)|\}. \tag{A.1}$$

Assuming that both signal and noise can be modeled as zero-mean Gaussian processes, we know from [29] that

$$E\{|x_1(i)-x_2(i+n)|\} = \sqrt{\frac{2}{\pi}}\sqrt{E\left\{[x_1(i)-x_2(i+n)]^2\right\}}. \tag{A.2}$$

Then it is trivial to derive

$$E\{\hat{\Psi}_{\mathrm{AMDF}}(n)\} = \sqrt{\frac{2}{\pi}[e_{x_1}+e_{x_2}-2R_{x_1x_2}(n)]}, \tag{A.3}$$

where $e_{x_1} = E\{x_1^2(n)\}$ and $e_{x_2} = E\{x_2^2(n)\}$ represent the energies of the signals $x_1(n)$ and $x_2(n)$, and $R_{x_1x_2}(n) = E\{x_1(i)x_2(i+n)\}$ is the direct cross-correlation function. Similarly, one can derive the expectation of the AMSF as follows:

$$E\{\hat{\Psi}_{\mathrm{AMSF}}(n)\} = \sqrt{\frac{2}{\pi}[e_{x_1}+e_{x_2}+2R_{x_1x_2}(n)]}. \tag{A.4}$$

The covariance between AMDF and AMSF is written as follows:

$$\mathrm{cov}[\hat{\Psi}_{\mathrm{AMDF}}(n),\hat{\Psi}_{\mathrm{AMSF}}(n)] = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{k=0}^{N-1}E\{|x_1(i)-x_2(i+n)|\,|x_1(k)+x_2(k+n)|\} - E\{\hat{\Psi}_{\mathrm{AMDF}}(n)\}E\{\hat{\Psi}_{\mathrm{AMSF}}(n)\}. \tag{A.5}$$

For two random Gaussian variables θ and ϑ, it can be derived from [29] that

$$E\{|\theta|\cdot|\vartheta|\} = \frac{2}{\pi}\left\{E[\theta\cdot\vartheta]\sin^{-1}\frac{E[\theta\cdot\vartheta]}{\sqrt{E[\theta^2]E[\vartheta^2]}} + \sqrt{E[\theta^2]\cdot E[\vartheta^2]-(E[\theta\cdot\vartheta])^2}\right\}. \tag{A.6}$$

Therefore, the covariance between AMDF and AMSF can be expressed as follows:

$$\mathrm{cov}[\hat{\Psi}_{\mathrm{AMDF}}(n),\hat{\Psi}_{\mathrm{AMSF}}(n)] = \xi_1(n)+\xi_2(n)-\xi_3(n), \tag{A.7}$$

where

$$\xi_1(n) = \frac{2}{\pi N^2}\sum_{i=0}^{N-1}\sum_{k=0}^{N-1}\mathcal{R}\cdot\sin^{-1}\frac{\mathcal{R}}{\sqrt{[R_{x_1x_1}(0)+R_{x_2x_2}(0)]^2-4[R_{x_1x_2}(n)]^2}},$$
$$\xi_2(n) = \frac{2}{\pi N^2}\times\sum_{i=0}^{N-1}\sum_{k=0}^{N-1}\sqrt{[R_{x_1x_1}(0)+R_{x_2x_2}(0)]^2-4[R_{x_1x_2}(n)]^2-\mathcal{R}^2},$$
$$\xi_3(n) = \frac{2}{\pi}\sqrt{[R_{x_1x_1}(0)+R_{x_2x_2}(0)]^2-4[R_{x_1x_2}(n)]^2},$$
$$\mathcal{R} = R_{x_1x_1}(k-i)+R_{x_1x_2}(k+n-i)-R_{x_1x_2}(i+n-k)-R_{x_2x_2}(k-i). \tag{A.8}$$

A similar derivation can be used to derive the variance of $\hat{\Psi}_{\mathrm{AMDF}}(n)$ and $\hat{\Psi}_{\mathrm{AMSF}}(n)$. We can then calculate the correlation coefficient $\rho(n)$ defined as

$$\rho(n) = \frac{\mathrm{cov}[\hat{\Psi}_{\mathrm{AMDF}}(n),\hat{\Psi}_{\mathrm{AMSF}}(n)]}{\sqrt{\mathrm{var}[\hat{\Psi}_{\mathrm{AMDF}}(n)]\,\mathrm{var}[\hat{\Psi}_{\mathrm{AMSF}}(n)]}}. \tag{A.9}$$

For simplicity of analysis, besides the assumptions made in Section 2, we further assume that the signal is also a Gaussian process with zero-mean and variance σ_s^2, the noise observed from different microphones has the same variance denoted by σ_w^2, and that the relative propagation attenuation between two microphones is negligible, that is, $\alpha = 1$. After making the above assumptions, we have

$$R_{x_1x_1}(n) = \begin{cases}\sigma_s^2+\sigma_w^2, & \text{for } n=0,\\ 0, & \text{otherwise,}\end{cases}$$
$$R_{x_2x_2}(n) = \begin{cases}\sigma_s^2+\sigma_w^2, & \text{for } n=0,\\ 0, & \text{otherwise,}\end{cases}$$
$$R_{x_1x_2}(n) = \begin{cases}\sigma_s^2, & \text{for } n=\tau,\\ 0, & \text{otherwise.}\end{cases} \tag{A.10}$$

We now consider to estimate $\rho(n)$ in two conditions, that is, $n=\tau$ and $n\neq\tau$.

(i) When $n=\tau$. Substituting $R_{x_1x_1}(n)$, $R_{x_2x_2}(n)$, and $R_{x_1x_2}(n)$ into (A.7), we have

$$\xi_1(n) = 0,$$
$$\xi_2(n) = \xi_3(n) = \frac{4}{\pi}\sqrt{\sigma_w^4+2\sigma_s^2\sigma_w^2}. \tag{A.11}$$

Therefore, in this case, $\rho(n) = 0$.

(ii) When $n \neq \tau$, we have

$$\begin{aligned}
\xi_1(n) &= \frac{4}{\pi N}\sigma_s^2 \sin^{-1}\frac{\sigma_s^2}{2(\sigma_s^2+\sigma_w^2)},\\
\xi_2(n) &= \frac{4}{\pi}(\sigma_s^2+\sigma_w^2)\\
&\quad + \frac{8}{\pi N}(\sigma_s^2+\sigma_w^2)\left[\sqrt{1-\frac{\sigma_s^4}{4(\sigma_s^2+\sigma_w^2)^2}}-1\right],\\
\xi_3(n) &= \frac{4}{\pi}(\sigma_s^2+\sigma_w^2).
\end{aligned} \tag{A.12}$$

In the context of TDE, N is often taken between a few hundreds and a few thousands. With such a large N, it can be verified that $\xi_3(n) \approx \xi_2(n) \gg \xi_1(n)$. It is trivial then to show that $\rho(n) \approx 0$.

ACKNOWLEDGMENTS

The authors are very grateful to Dr. Aki Harma and Dr. Gary W. Elko for providing the measured impulse responses. They also would like to thank Dr. Dennis R. Morgan for carefully reading a draft and providing many constructive comments and suggestions that have improved the clarity of this paper.

REFERENCES

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

[2] G. C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 463–470, 1981.

[3] G. C. Carter, A. H. Nutall, and P. G. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.

[4] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, pp. 62–70, 1971.

[5] J. C. Hassab and R. E. Boucher, "Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 549–555, 1981.

[6] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[7] R. Cusani, "Performance of fast time delay estimators," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 5, pp. 757–759, 1989.

[8] A. Kumar and Y. Bar-Shalom, "Time-domain analysis of cross correlation for time delay estimation with an autocorrelated signal," *IEEE Trans. Signal Processing*, vol. 41, no. 4, pp. 1664–1668, 1993.

[9] D. Hertz, "Time delay estimation by combining efficient algorithms and generalized cross-correlation methods," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 1–7, 1986.

[10] J. A. Stuller and N. Hubing, "New perspectives for maximum likelihood time-delay estimation," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 513–525, 1997.

[11] C. During, "Recursive versus nonrecursive correlation for real-time peak detection and tracking," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 781–785, 1997.

[12] P. G. Georgiou, P. Tsakalides, and C. Kyriakakis, "Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 291–301, 1999.

[13] Z. Wang, J. Y. Cheung, Y. S. Xia, and J. D. Z. Chen, "Neural implementation of unconstrained minimum L_1-norm optimization-least absolute deviation model and its application to time delay estimation," *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 11, pp. 1214–1226, 2000.

[14] A. Stephenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 5, pp. 3055–3058, Detroit, Mich, USA, May 1995.

[15] X. Sun and S. C. Douglas, "Adaptive time delay estimation with allpass constraints," in *Proc. 33rd Asilomar Conference on Signals, Systems, and Computers*, vol. 2, pp. 898–902, Pacific Grove, Calif, USA, October 1999.

[16] G. C. Carter, "Coherence and time delay estimation," in *Signal Processing Handbook*, C. H. Chen, Ed., pp. 443–482, Marcel Dekker, New York, NY, USA, 1988.

[17] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 525–533, 1993.

[18] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[19] Y. Bar-Shalom, F. Palimieri, A. Kumar, and H. M. Shertukde, "Analysis of wide-band cross correlation for time-delay estimation," *IEEE Trans. Signal Processing*, vol. 41, no. 1, pp. 385–387, 1993.

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

[21] R. E. Boucher and J. C. Hassab, "Analysis of discrete implementation of generalized cross correlator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 609–611, 1981.

[22] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 384–391, 2000.

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

[24] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech, and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.

[25] T. G. Dvorkind and S. Gannot, "Approaches for time difference of arrival estimation in a noisy and reverberant environment," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 215–218, Kyoto, Japan, September 2003.

[26] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

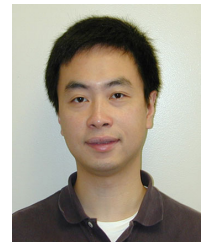
[27] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 7, pp. 727–730, 2001.

[28] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[29] G. Jacovitti, A. Neri, and R. Cusani, "On a fast digital method of estimating the autocorrelation of a gaussian stationary process," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 5, pp. 968–976, 1984.
[30] G. Jacovitti and R. Cusani, "An efficient technique for high correlation estimation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 35, no. 5, pp. 654–660, 1987.
[31] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker, New York, NY, USA, 1992.
[32] D. L. Maskell and G. S. Woods, "The estimation of subsample time delay of arrival in the discrete-time measurement of phase delay," *IEEE Trans. Instrumentation and Measurement*, vol. 48, no. 6, pp. 1227–1231, 1999.
[33] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centennial Symposium*, pp. 343–346, Woodbury, NY, USA, June 1994.
[34] A. Harma, "Acoustic measurement data from the varechoic chamber," Tech. Memorandum 110101, Agere Systems, Allentown, Pa, USA, 2001.

Jingdong Chen received the B.S. degree in electrical engineering and the M.S. degree in array signal processing from the Northwestern Polytechnic University in 1993 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998. His Ph.D. research focused on speech recognition in noisy environments. From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, as a Research Fellow, where he engaged in research in robust speech recognition, signal processing, and discriminative feature representation. From 2000 to 2001, he was with ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories as a member of the technical staff in July 2001. His current research interests include adaptive signal processing, speech enhancement, adaptive noise/echo cancellation, microphone array signal processing, signal separation, and source localization. He is the recipient of 1998-1999 research grant from the Japan Key Technology Center, and the 1996-1998 President's Award from the Chinese Academy of Sciences.

Jacob Benesty was born in 1963. He received the M.S. degree in microwaves from Pierre & Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecommunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was with Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined INRS-EMT, University of Quebec, Montreal, Quebec, Canada, as an Associate Professor. His research interests are in acoustic signal processing and multimedia communications. He was the Cochair of the 1999 International Workshop on Acoustic Echo and Noise Control. He is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Electroacoustics. He is a Member of the Editorial Board of the EURASIP Journal on Applied Signal Processing. He is the recipient of the IEEE Signal Processing Society's 2001 Best Paper Award. He coauthored the book *Advances in Network and Acoustic Echo Cancellation* (Springer-Verlag, Berlin, 2001) and coedited/coauthored three more books.



Yiteng (Arden) Huang received the B.S. degree from the Tsinghua University in 1994, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech) in 1998 and 2001, respectively, all in electrical and computer engineering. Upon graduation, he joined Bell Laboratories as a member of the technical staff in March 2001. His current research interests are in adaptive filtering, multichannel signal processing, source localization, microphone array for hands-free telecommunication, statistical signal processing, and wireless communications. Dr. Huang is currently an Associate Editor of the IEEE Signal Processing Letters. He is a Coeditor/Coauthor of the book *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, Berlin, 2003). He received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000-2001 Outstanding Graduate Teaching Assistant Award from the School of Electrical and Computer Engineering, Georgia Tech, the 2000 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Tech, and the 1997-1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech.