Robust Time Delay Estimation Exploiting Redundancy Among Multiple Microphones

Jingdong Chen, *Member, IEEE*, Jacob Benesty, *Member, IEEE*, and Yiteng (Arden) Huang, *Member, IEEE*

***Abstract*—To find the position of an acoustic source in a room, typically, a set of relative delays among different microphone pairs needs to be determined. The generalized cross-correlation (GCC) method is the most popular to do so and is well explained in a landmark paper by Knapp and Carter. In this paper, the idea of cross-correlation coefficient between two random signals is generalized to the multichannel case by using the notion of spatial prediction. The multichannel spatial correlation matrix is then deduced and its properties are discussed. We then propose a new method based on the multichannel spatial correlation matrix for time delay estimation. It is shown that this new approach can take advantage of the redundancy when more than two microphones are available and this redundancy can help the estimator to better cope with noise and reverberation.**

***Index Terms*—Cross-correlation coefficient, linear prediction, spatial correlation, time delay estimation.**

I. Introduction

Traditionally, time delay estimation (TDE), from measurements provided by an array of sensors, has played an important role in radar, sonar, and seismology for localizing radiating sources. Nowadays, with the increased development of communications among humans and human–machine interfaces, the need for localizing and tracking acoustic sources in a room has become essential. Two specific examples are automatic camera tracking for video-conferencing and microphone array beam steering for suppressing noise and reverberation in all types of communication and voice processing systems. The time delay estimation-based locator has become the technique of choice in these applications, especially in recent digital systems [1]–[7].

The generalized cross-correlation (GCC) method, proposed by Knapp and Carter in 1976 [8], is the most popular technique for TDE. The delay estimate is obtained as the time-lag that maximizes the cross-correlation between filtered versions of the received signals. Since then, many new ideas have been proposed to deal better with noise and reverberation; see [9]–[15]. However, reverberation remains a problem and in a highly reverberant room, all known methods fail.

There are mainly two approaches to deal more efficiently with reverberation. The first one is to blindly estimate the impulse responses from the source to the two microphones [14], [15]. The better this estimate is, the better the relative delay between these two microphone signals can be estimated; but this is a difficult problem and the resulting time delay estimates are sensitive to noise. The second approach is to use more than two microphones and take advantage of the redundancy. This is the choice that we have taken here.

In this paper, the idea of cross-correlation coefficient between two random signals is generalized to the multichannel case by using the notion of spatial prediction. The multichannel spatial correlation matrix is then deduced and its properties are discussed. We then propose a new time delay estimator, which can take advantage of the redundancy among multiple microphones. It is believed that this redundancy can help to better deal with both noise and reverberation. Numerical studies have been performed and the results show that the effect of noise and reverberation is dramatically reduced when multiple microphones are used. The relative delay estimation accuracy increases with the number of microphones.

II. Signal Model

Suppose that we have an array as shown in Fig. 1, which consists of $L+1$ microphones whose outputs are denoted as $x_l[n], l=0,1,\ldots,L$. Without loss of generality, we assume that the wave is in-phase at microphone 0. We consider the following propagation model:

$$\begin{bmatrix} x_0[n] \\ x_1[n] \\ x_2[n] \\ \vdots \\ x_L[n] \end{bmatrix} = \begin{bmatrix} \alpha_0 & 0 & 0 & \cdots & 0 \\ 0 & \alpha_1 & 0 & \cdots & 0 \\ 0 & 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \alpha_L \end{bmatrix} \times \begin{bmatrix} s[n-t] \\ s[n-t-\tau] \\ s[n-t-f_2(\tau)] \\ \vdots \\ s[n-t-f_L(\tau)] \end{bmatrix} + \begin{bmatrix} w_0[n] \\ w_1[n] \\ w_2[n] \\ \vdots \\ w_L[n] \end{bmatrix} \quad (1)$$

where $\alpha_l, l=0,1,2,\ldots,L$, are the attenuation factors due to propagation effects, t is the propagation time from the unknown source $s[n]$ to microphone 0, $w_l[n]$ is an additive noise signal at the lth microphone, τ is the relative delay between microphones 0 and 1, and $f_l(\tau)$ is the relative delay between microphones 0 and l. The function f_l depends of τ but also of the microphone

Manuscript received November 20, 2002; revised June 9, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

J. Chen and Y. Huang are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com; arden@research.bell-labs.com).

J. Benesty is with University of Quebec-INRS, Montreal, QC, H5A 1K6, Canada (e-mail: benesty@inrs-emt.uquebec.ca).

Digital Object Identifier 10.1109/TSA.2003.818025

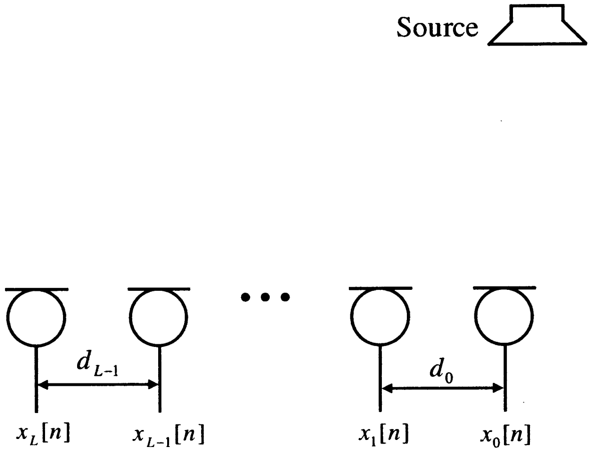


Fig. 1. Linear microphone array.

array geometry. For example, in the far-field case (plane wave propagation), for a linear equispaced array, we have

$$f_l(\tau) = l\tau \tag{2}$$

and for a linear nonequispaced array, we have

$$f_l(\tau) = \frac{\sum_{i=0}^{l-1} d_i}{d_0}\tau \tag{3}$$

where d_i is the distance between microphones i and $i+1, i = 0, 1, 2, \ldots, L-1$. In the near-field case, f_l depends also on the position of the source. In general τ is not known, but the geometry of the antenna is known such that the exact mathematical relation of the relative delay between microphones 0 and l is well defined and given. It is further assumed that $s[n]$ and $w_l[n], l = 0, 1, 2, \ldots, L$, are zero-mean, mutually uncorrelated, stationary Gaussian random processes.

III. Spatial Prediction and Interpolation

The notion of spatial prediction was presented in [16] but in the simple case that makes the spatial prediction equivalent to the classical linear prediction. In this section, we generalize this idea in a way that the geometry of the array is taken into account as well as the relative delay among the elements of this array. As a result, the spatial correlation matrix has a much more general form.

A. Linear Forward Spatial Prediction

Considering the microphone 0, we would like to align successive time samples of this microphone signal with spatial samples from the L other microphone signals. It is clear that $x_0[n - f_L(\tau)]$ is in-phase with the signals $x_l[n - f_L(\tau) + f_l(\tau)], l = 1, 2, \ldots, L$. From these observations, we define the following forward spatial prediction error signal:

$$e_0[n - f_L(m)] = x_0[n - f_L(m)] - \mathbf{x}_{1:L}^T[n - f_L(m)]\mathbf{a}_m \tag{4}$$

where m is any guessed relative delay, superscript ^{T} denotes transpose of a vector or a matrix and the first equation at the bottom of the page is the linear forward spatial predictor. Consider the criterion

$$J_{m,0} = E\left\{e_0^2[n - f_L(m)]\right\} \tag{5}$$

where $E\{\cdot\}$ denotes mathematical expectation. Minimization of (5) leads to the equation

$$\mathbf{R}_{m,1:L}\mathbf{a}_m = \mathbf{r}_{m,1:L} \tag{6}$$

where the second equation at the bottom of the page is the spatial correlation matrix, and

$$\begin{aligned}\mathbf{r}_{m,1:L} &= E\{\mathbf{x}_{1:L}[n - f_L(m)]x_0[n - f_L(m)]\}\\ &= \begin{bmatrix} E\{x_1[n - f_L(m) + f_1(m)]x_0[n - f_L(m)]\} \\ E\{x_2[n - f_L(m) + f_2(m)]x_0[n - f_L(m)]\} \\ \vdots \\ E\{x_L[n]x_0[n - f_L(m)]\} \end{bmatrix}\\ &= \begin{bmatrix} E\{x_1[n]x_0[n - f_1(m)]\} \\ E\{x_2[n]x_0[n - f_2(m)]\} \\ \vdots \\ E\{x_L[n]x_0[n - f_L(m)]\} \end{bmatrix}\end{aligned}$$

is the spatial correlation vector. Note that the spatial correlation matrix is not Toeplitz in general, except for some particular cases.

For $m = \tau$ and for the noise free case where $w_l[n] = 0, l = 1, 2, \ldots, L$, it can easily be checked that with our signal model, the rank of matrix $\mathbf{R}_{\tau,1:L}$ is equal to 1. This means that the samples $x_0[n-\tau]$ can be perfectly predicted from any of one other microphone samples. However, the noise is never zero in practice and is in general isotropic. The energy of the different noise

$$\mathbf{x}_{1:L}[n - f_L(m)] = [x_1[n - f_L(m) + f_1(m)] \quad x_2[n - f_L(m) + f_2(m)] \quad \cdots \quad x_L[n]]^T$$

and

$$\mathbf{a}_m = [a_{m,1} \quad a_{m,2} \quad \cdots \quad a_{m,L}]^T$$

$$\begin{aligned}\mathbf{R}_{m,1:L} &= E\left\{\mathbf{x}_{1:L}[n - f_L(m)]\mathbf{x}_{1:L}^T[n - f_L(m)]\right\}\\ &= \begin{bmatrix} E\left\{x_1^2[n]\right\} & E\{x_1[n - f_2(m)]x_2[n - f_1(m)]\} & \cdots & E\{x_1[n - f_L(m)]x_L[n - f_1(m)]\} \\ E\{x_2[n - f_1(m)]x_1[n - f_2(m)]\} & E\left\{x_2^2[n]\right\} & \cdots & E\{x_2[n - f_L(m)]x_L[n - f_2(m)]\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_L[n - f_1(m)]x_1[n - f_L(m)]\} & E\{x_L[n - f_2(m)]x_2[n - f_L(m)]\} & \cdots & E\left\{x_L^2[n]\right\} \end{bmatrix}\end{aligned}$$

signals at the microphones will be added at the main diagonal of the correlation matrix $\mathbf{R}_{\tau,1:L}$, will regularize it, and this matrix will become positive definite (which we suppose in the rest of this paper). A unique solution to (6) is then guaranteed whatever the number of microphones is. This solution is optimal from a Wiener theory point of view.

B. Linear Backward Spatial Prediction

Considering the microphone L, we would like to align successive time samples of this microphone signal with spatial samples from the L other microphone signals. It is clear that $x_L[n]$ is in-phase with the signals $x_l[n-f_L(\tau)+f_l(\tau)], l=0,1,\ldots,L-1$. From these observations, we define the following backward spatial prediction error signal:

$$e_L[n-f_L(m)] = x_L[n] - \mathbf{x}_{0:L-1}^T[n-f_L(m)]\mathbf{b}_m, \tag{7}$$

where the first equation at the bottom of the page is the linear backward spatial predictor. Minimization of the criterion

$$J_{m,L} = E\left\{e_L^2[n-f_L(m)]\right\} \tag{8}$$

leads to the equation

$$\mathbf{R}_{m,0:L-1}\mathbf{b}_m = \mathbf{r}_{m,0:L-1} \tag{9}$$

where

$$\mathbf{R}_{m,0:L-1} = E\left\{\mathbf{x}_{0:L-1}[n-f_L(m)]\mathbf{x}_{0:L-1}^T[n-f_L(m)]\right\}$$

and

$$\mathbf{r}_{m,0:L-1} = E\{\mathbf{x}_{0:L-1}[n-f_L(m)]x_L[n]\}.$$

C. Linear Spatial Interpolation

The ideas presented for spatial prediction can easily be extended to spatial interpolation, where we consider any microphone element $l, l=0,1,2,\ldots,L$. The spatial interpolation error signal is defined as

$$e_l[n-f_L(m)] = -\mathbf{x}_{0:L}^T[n-f_L(m)]\mathbf{c}_{m,l} \tag{10}$$

where the second equation at the bottom of the page with $c_{m,l,l}=-1$, is the spatial interpolator. The criterion associated with (10) is

$$J_{m,l} = E\left\{e_l^2[n-f_L(m)]\right\}. \tag{11}$$

The rest flows immediately from the previous sections on prediction.

IV. Application to Time Delay Estimation

In this section, we only use the forward spatial prediction idea but of course backward spatial prediction and spatial interpolation can also be used. So we consider the minimization of criterion $J_{m,0}$ for different m.

Let $J_{m,0;\min}$ denote the minimum mean-squared error, for the value m, defined by

$$J_{m,0;\min} = E\left\{e_{0;\min}^2[n-f_L(m)]\right\}. \tag{12}$$

If we replace $\mathbf{a}_m$ by $\mathbf{R}_{m,1:L}^{-1}\mathbf{r}_{m,1:L}$ in (4), we get

$$\begin{aligned} e_{0;\min}[n-f_L(m)] &= x_0[n-f_L(m)] \\ &\quad - \mathbf{x}_{1:L}^T[n-f_L(m)]\mathbf{R}_{m,1:L}^{-1}\mathbf{r}_{m,1:L}. \end{aligned} \tag{13}$$

We deduce that

$$J_{m,0;\min} = E\left\{x_0^2[n-f_L(m)]\right\} - \mathbf{r}_{m,1:L}^T\mathbf{R}_{m,1:L}^{-1}\mathbf{r}_{m,1:L}. \tag{14}$$

The value of m that gives the minimum $J_{m,0;\min}$, for different m, corresponds to the time delay between microphone 0 and 1. Mathematically, the solution to our problem is then given by

$$\hat{\tau} = \arg\min_m J_{m,0;\min}, \tag{15}$$

where $\hat{\tau}$ is an estimate of τ.

Particular case: Two microphones $(L=1)$. In this case, the solution is

$$\begin{aligned} \hat{\tau} &= \arg\min_m \left\{E\left\{x_0^2[n-m]\right\}\right. \\ &\qquad \left.\times \left[1 - \frac{E^2\{x_0[n-m]x_1[n]\}}{E\left\{x_0^2[n-m]\right\}E\left\{x_1^2[n]\right\}}\right]\right\} \\ &= \arg\min_m \left\{E\left\{x_0^2[n-m]\right\}\left[1-\rho_{m,01}^2\right]\right\} \\ &= \arg\min_m \left\{1-\rho_{m,01}^2\right\} \\ &= \arg\max_m \left(\rho_{m,01}^2\right) \end{aligned} \tag{16}$$

where $\rho_{m,01}(\rho_{m,01}^2 \leq 1)$ is the cross-correlation coefficient between $x_0[n-m]$ and $x_1[n]$. When the cross-correlation coefficient is close to 1, this means that the two signals that we compare are highly correlated which happens when the signals are

$$\begin{aligned} &\mathbf{x}_{0:L-1}[n-f_L(m)] \\ &\quad = [x_0[n-f_L(m)+f_0(m)] \quad x_1[n-f_L(m)+f_1(m)] \quad \cdots \quad x_{L-1}[n-f_L(m)+f_{L-1}(m)]]^T \end{aligned}$$

and

$$\mathbf{b}_m = [b_{m,1} \quad b_{m,2} \quad \cdots \quad b_{m,L}]^T$$

$$\mathbf{x}_{0:L}[n-f_L(m)] = [\,x_0[n-f_L(m)+f_0(m)] \quad x_1[n-f_L(m)+f_1(m)] \quad \cdots \quad x_L[n]\,]^T$$

and

$$\mathbf{c}_{m,l} = [c_{m,l,0} \quad c_{m,l,1} \quad \cdots \quad c_{m,l,L}]^T$$

in-phase, i.e., $m \approx \tau$ and this implies that $J_{\tau,0;\min} \approx 0$. This approach is similar to the generalized cross-correlation method proposed by Knapp and Carter [8]. Note that in the general case with any number of microphones, the proposed approach can be seen as a cross-correlation method, but we take advantage of the knowledge of the microphone array to estimate only one time delay (instead of estimating multiple time delays independently) in an optimal way in a least mean square sense.

V. Other Information From the Spatial Correlation Matrix

Consider the $L+1$ microphone signals $x_l, l = 0, 1, \ldots, L$, the corresponding spatial correlation matrix is

$$\begin{aligned}\mathbf{R}_{m,0:L} &= \mathbf{R}_m \\ &= E\left\{\mathbf{x}_{0:L}[n - f_L(m)]\mathbf{x}_{0:L}^T[n - f_L(m)]\right\}. \end{aligned} \tag{17}$$

It can be shown that $\mathbf{R}_m$ can be factored as

$$\mathbf{R}_m = \mathbf{D}\tilde{\mathbf{R}}_m\mathbf{D} \tag{18}$$

where

$$\mathbf{D} = \begin{bmatrix} \sqrt{E\left\{x_0^2[n]\right\}} & 0 & \cdots & 0 \\ 0 & \sqrt{E\left\{x_1^2[n]\right\}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{E\left\{x_L^2[n]\right\}} \end{bmatrix} \tag{19}$$

is a diagonal matrix

$$\tilde{\mathbf{R}}_m = \begin{bmatrix} 1 & \rho_{m,01} & \cdots & \rho_{m,0L} \\ \rho_{m,01} & 1 & \cdots & \rho_{m,1L} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{m,0L} & \cdots & \rho_{m,L-1L} & 1 \end{bmatrix} \tag{20}$$

is a symmetric matrix, and

$$\rho_{m,kl} = \frac{E\{x_k[n - f_l(m)]x_l[n - f_k(m)]\}}{\sqrt{E\left\{x_k^2[n]\right\}E\left\{x_l^2[n]\right\}}}, \quad k, l = 0, 1, \ldots, L, \tag{21}$$

is the cross-correlation coefficient between $x_k[n - f_l(m)]$ and $x_l[n - f_k(m)]$.

We now give two propositions that will be useful for TDE.

Proposition 1: We have

$$0 < \det(\tilde{\mathbf{R}}_m) \leq 1 \tag{22}$$

where "det" stands for *determinant*.

Proof: Since $\mathbf{R}_m$ is symmetric and is supposed to be positive definite, it is clear that $\det(\mathbf{R}_m) > 0$ which implies that $\det(\tilde{\mathbf{R}}_m) > 0$. To show that $\det(\tilde{\mathbf{R}}_m) \leq 1$, we can use the *Cholesky factorization* [17]. Since $\tilde{\mathbf{R}}_m$ is symmetric and positive definite, there exists a unique lower triangular matrix $\mathbf{Q}_m$ with positive diagonal entries such that $\tilde{\mathbf{R}}_m = \mathbf{Q}_m\mathbf{Q}_m^T$, where

$$\mathbf{Q}_m = \begin{bmatrix} q_{m,00} & 0 & \cdots & 0 \\ q_{m,10} & q_{m,11} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ q_{m,L0} & \cdots & q_{m,LL-1} & q_{m,LL} \end{bmatrix}. \tag{23}$$

It can be shown that the elements of the main diagonal of matrix $\mathbf{Q}_m$ can be computed as follows:

$$q_{m,ll} = \sqrt{1 - \sum_{k=0}^{l-1} q_{m,lk}^2}, \quad l = 0, 1, \ldots, L. \tag{24}$$

It follows immediately from (24) that $0 < q_{m,ll} \leq 1, \forall\ l$. Furthermore, since $\mathbf{Q}_m$ is a triangular matrix, we have

$$\det(\tilde{\mathbf{R}}_m) = \prod_{l=0}^{L} q_{m,ll}^2 \leq 1.$$

That completes the proof.

Another way to show this proposition is by induction, i.e.,

$$\det(\tilde{\mathbf{R}}_m) = \det(\tilde{\mathbf{R}}_{m,0:L}) \leq \det(\tilde{\mathbf{R}}_{m,1:L}) \leq \cdots \leq 1. \tag{25}$$

Proposition 2: We have

$$\det(\tilde{\mathbf{R}}_m) \leq \frac{J_{m,0;\min}}{E\left\{x_0^2[n]\right\}} \leq 1. \tag{26}$$

Proof: The forward prediction error signal defined in (4) can be rewritten as

$$e_0[n - f_L(m)] = -\mathbf{x}_{0:L}^T\underline{\mathbf{a}}_m \tag{27}$$

where $\underline{\mathbf{a}}_m = [-1, \mathbf{a}_m^T]^T$. Then the criterion shown in (5) can be expressed as

$$J_{m,0} = E\left\{e_0^2[n - f_L(m)]\right\} + \mu(\boldsymbol{\delta}^T\underline{\mathbf{a}}_m + 1) \tag{28}$$

where $\boldsymbol{\delta} = [1\ 0\ \cdots\ 0]^T$, and μ is the Lagrange multiplier. It is then easily shown that

$$J_{m,0;\min} = \frac{1}{\boldsymbol{\delta}^T\mathbf{R}_m^{-1}\boldsymbol{\delta}}. \tag{29}$$

In this case, using (18), (29) becomes

$$\begin{aligned} J_{m,0;\min} &= \frac{E\left\{x_0^2[n]\right\}}{\boldsymbol{\delta}^T\tilde{\mathbf{R}}_m^{-1}\boldsymbol{\delta}} \\ &= E\left\{x_0^2[n]\right\}\frac{\det(\tilde{\mathbf{R}}_m)}{\det(\tilde{\mathbf{R}}_{m,1:L})}. \end{aligned} \tag{30}$$

Using (25), it is clear that proposition 2 is verified.

In the general case, for any interpolator, we have

$$\det(\tilde{\mathbf{R}}_m) \leq \frac{J_{m,l;\min}}{E\left\{x_l^2[n]\right\}} \leq 1, \quad l = 0, 1, \ldots, L. \tag{31}$$

As we can see, the determinant of the spatial correlation matrix is related to the minimum mean-squared error and to the correlation of the signals. Let's take the two-channel case. It is obvious that the cross-correlation coefficient between the two signals x_0 and x_1 is linked to the determinant of the corresponding spatial correlation matrix

$$\rho_{m,01}^2 = 1 - \det(\tilde{\mathbf{R}}_{m,0:1}). \tag{32}$$

By analogy to the cross-correlation coefficient definition between two random signals, we define the multichannel correlation coefficient among the signals $x_l, l = 0, 1, \ldots, L$, as

$$\rho_{m,0:L}^2 = 1 - \det(\tilde{\mathbf{R}}_{m,0:L}). \tag{33}$$

From proposition 2, we give a new bound for $\rho_{m,0:L}^2$

$$1 - \frac{J_{m,0;\min}}{E\left\{x_0^2[n]\right\}} \leq \rho_{m,0:L}^2 \leq 1. \tag{34}$$

Basically, the coefficient $\rho_{m,0:L}$ will measure the amount of correlation among all the channels. This coefficient has some interesting properties. For example, if one of the signals, say x_0, is completely decorrelated from the others because the microphone is defective, or it picks up only noise, or the signal is saturated, this signal will not affect $\rho_{m,0:L}$ since $\rho_{m,0l} = 0, \forall\ l$. In this case

$$\rho_{m,0:L}^2 = \rho_{m,1:L}^2. \tag{35}$$

In other words, the measure "drops" the signals which have no correlation with the others. This makes sense from a correlation point of view, since we want to measure the degree of correlation only from the channels who have something in common. In the extreme cases where all the signals are uncorrelated, we have $\rho_{m,0:L}^2 = 0$, and where any two signals (or more) are perfectly correlated, we have $\rho_{m,0:L}^2 = 1$.

Obviously, the multichannel coefficient $\rho_{m,0:L}^2$ can be used for time delay estimation in the following way:

$$\begin{aligned}\hat{\tau} &= \arg\max_{\mathrm{m}} \left(\rho_{\mathrm{m,0:L}}^2\right) \\ &= \arg\min_{\mathrm{m}} [\det(\tilde{\mathbf{R}}_{\mathrm{m,0:L}})].\end{aligned} \tag{36}$$

This method can be seen as a multichannel correlation approach for the estimation of time delay and it is clear that (36) is equivalent to (15).

VI. Simulation Experiments

We have proposed a multichannel correlation approach for the time delay estimation problem. A series of Monte-Carlo simulation experiments were conducted to study the characteristics of the proposed algorithm, and the difference in TDE performance when more microphones are used. Three sets of experimental results are presented here: one involves a set in noisy but nonreverberant environment and the other two pertain to reverberation conditions.

A. Performance Criteria

Following [9] and [10], we distinguish an estimate as either an *anomaly* or a *nonanomaly* according to its absolute error. If the absolute error $|\hat{\tau}_i - \tau| > T_\mathrm{c}/2$, 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 i-th delay estimate respectively, and T_c is the signal correlation time. In our experiment, T_c is computed as the 3 dB width of the main lobe of the source signal autocorrelation function, which is equal to four (4) samples. The TDE performance is evaluated in terms of the percentage of anomalous estimates over the total estimates $(P_{\hat{\tau}})$, the bias $(B_{\hat{\tau}})$, and the standard deviation $(\sigma_{\hat{\tau}})$ of the nonanomalous estimates. These measures are defined as

$$P_{\hat{\tau}} = \frac{N_\mathrm{a}}{N_\mathrm{T}},$$

$$B_{\hat{\tau}} = \frac{1}{N_\mathrm{na}} \left| \sum_{i \in \mathcal{X}_\mathrm{na}} (\hat{\tau}_i - \tau) \right|,$$

and

$$\sigma_{\hat{\tau}} = \sqrt{\frac{1}{N_\mathrm{na}} \sum_{i \in \mathcal{X}_\mathrm{na}} |\hat{\tau}_i - \tau|^2} \tag{37}$$

where N_T denotes the total number of estimates, N_a is the number of estimates that are identified as anomalies, N_na is the number of nonanomalous estimates, and $\mathcal{X}_\mathrm{na}$ represents the subset of nonanomalous estimates. The smaller are the $P_{\hat{\tau}}, B_{\hat{\tau}}$, and $\sigma_{\hat{\tau}}$, the better the estimator is.

B. Experiment Setup

In an attempt to simulate real reverberant acoustic environments, the image model technology [18] is used. We consider a rectangular room with plane reflective boundaries (walls, ceiling and floor). Each boundary is characterized by a uniform reflection coefficient, which is independent of the frequency and the incidence angle of the source signal. The following parameter values are used.

- Room dimension: 120 × 180 × 150 inch $(x \times y \times z)$.
- Reflection coefficients: $r_i (i = 1, 2, \ldots, 6)$ varying between 0 and 1.
- Source Position: a point omnidirectional source is located at (22.5, 150.0, 112.5).
- Microphone positions: a linear microphone array which consists of ten (10) ideal point receivers (microphones) is placed in parallel with the x-axis. The first microphone (microphone 0) is located at (60.0, 7.5, 30.0), and the tenth at (100.5, 7.5, 30.0). The spacing between two adjacent microphones is 4.5 in. The directivity pattern of each microphone is assumed to be omnidirectional.
- SNR: varying between -10 dB and 0 dB.

An illustration of the setup is shown in Fig. 2. A low-pass sampled version of the impulse response of the acoustic transmission channel between the source and each microphone is generated using the image method. A 4-min speech signal from a female speaker, digitized with 16-bit resolution at 16 kHz, is then convolved with the ten synthetic impulse responses. Finally, mutually independent white Gaussian noise is properly scaled and added to each microphone signal to control the SNR.

C. Results and Interpretation

The algorithm used to obtain the time delay estimates can be summarized as follows.

- The microphone signals are partitioned into nonoverlapping frames with a frame width of 128 ms (2048 samples).

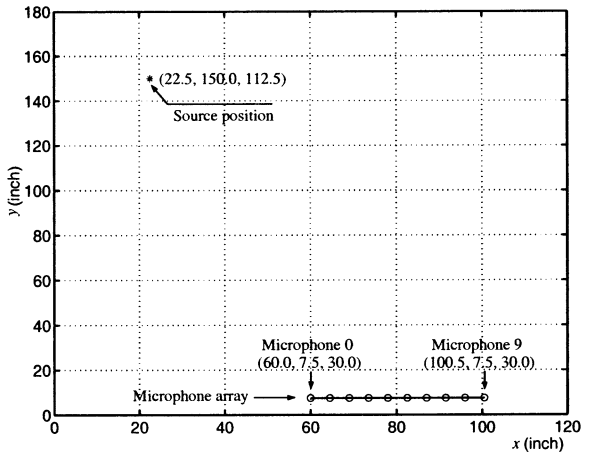


Fig. 2. Layout of the microphone array and source positions in the simulation environment.

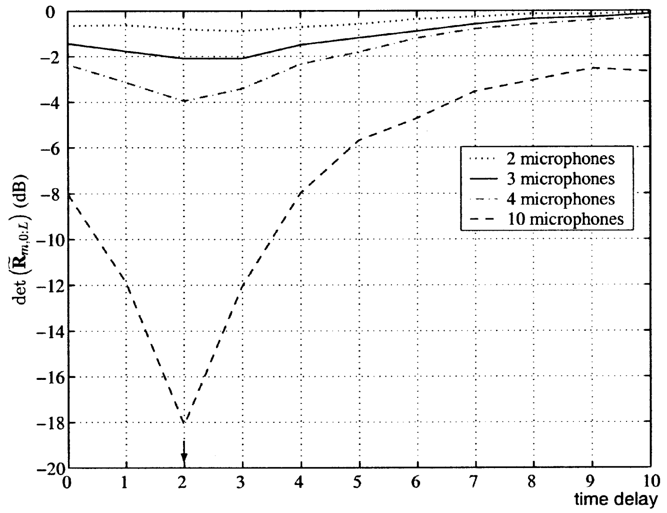


Fig. 3. Comparison of $\det(\tilde{\mathbf{R}}_{m,0:L})$ as a function of lag-time m for different microphones in a noisy environment where $\mathrm{SNR} = -5$ dB. Arrow shows the position of the true delay.

- A voice activity detector (VAD) based on the short-time energy and zero crossing rate is then applied to the signal at microphone 0 to identify regions of speech and nonspeech. These automatically labeled regions are then manually checked for accuracy and consistency.
- For each speech frame, the multichannel correlation approach given by (36) is applied to obtain a time delay estimate. The noise-only frame is disregarded.

In the absence of reverberation, the performance of the TDE algorithm is mainly influenced by the level (SNR) and the characteristics of noise. Fig. 3 plots the $\det(\tilde{\mathbf{R}}_{m,0:L})$ as a function of lag-time m for the case where $\mathrm{SNR} = -5$ dB. It shows the results using two, three, four, and ten microphones respectively. The true time delay is $\tau = 2$ (samples). When two and three microphones are employed, the estimated delay is 3 (samples). As the number of microphones is increased to four, the estimated delay is equal to the true delay. It is remarkable that as the number of microphones is increased, the valley of the cost function tends to be sharper, which will enable an easier search of the minimum. This demonstrates the effectiveness of the multichannel correlation approach in taking advantage of the redundant information provided by multiple microphones to mitigate the effect of noise.

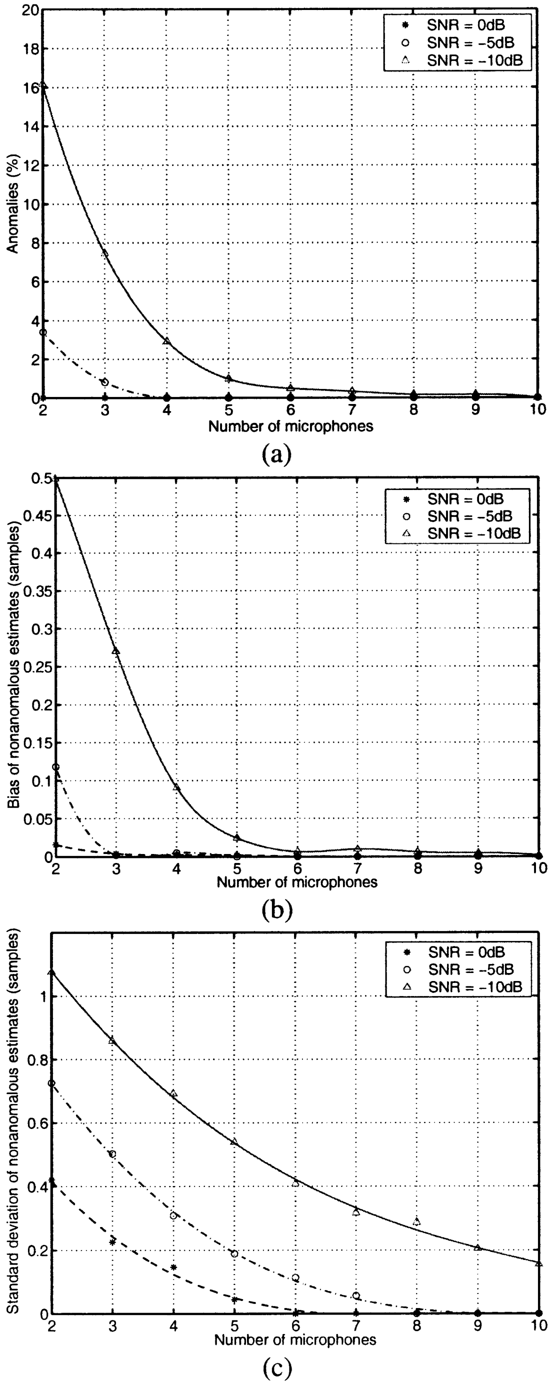


Fig. 4. Percentage of (a) anomalous time delay estimates, (b) bias, and (c) standard deviation of nonanomalous time delay estimates versus different number of microphones in nonreverberant environments.

Fig. 4 presents the TDE results obtained in the white Gaussian noise condition in the absence of reverberation, where the percentage of anomalies, the bias and standard deviation of nonanomalous estimates are plotted, respectively, all as a function of the number of microphones.

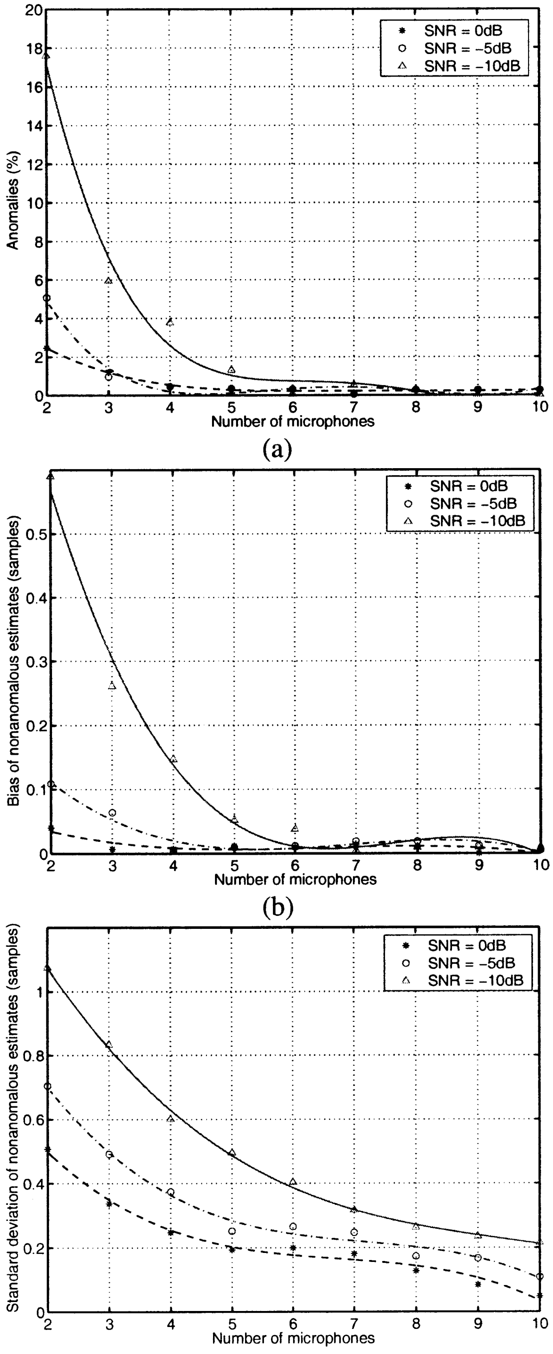
As clearly shown in Fig. 4, when the number of microphones is fixed, the TDE performance deteriorates as the level of noise increases. For example, for two microphones (in this case the multichannel correlation approach is equivalent to the GCC method), when $\mathrm{SNR} = 0$ dB, no anomaly is observed. As SNR decreases to -5 dB, the probability of anomalies grows up to nearly 4%. As SNR further drops down to -10 dB, the anomalies reach 16%, which is more than 4 times that of $\mathrm{SNR} = -5$ dB. Similarly, both the absolute value of bias and standard deviation of the nonanomalous estimates grow as SNR decreases.

In the same SNR condition, the number of anomalies, the bias and standard deviation of nonanomalous estimates, all reduce as more microphones are employed. For instance, in the condition where $\mathrm{SNR} = -10$ dB, the percentage of anomalies is over 16% for two microphones, but it diminishes to approximately 0 when more than 8 microphones are used. Similarly, the bias of the nonanomalous estimates is nearly 0.5 when only two microphones are available. Its value vanishes as the number of microphones is increased up to 6. It is remarkable that the performance obtained using 6 microphones in $\mathrm{SNR} = -10$ dB condition is almost as good as that achieved by two microphones in 0 dB. This demonstrates the powerfulness of the multichannel correlation approach in taking advantage of the redundancy among multiple microphones to deal with noise.

In reverberation condition, each microphone receives delayed and attenuated replicas of source signal due to reflections of the source wave from room boundaries in addition to the directional path signal. In such a case, the transmission of an acoustical signal between a source and microphones is not accurately characterized by the signal model given in (1). The TDE performance will be affected by not only background noise, but reverberation as well. Fig. 5 presents the results obtained in a light reverberation condition where all boundary reflection coefficients are $r_1 = r_2 = \cdots = r_6 = 0.5$. The reverberation time T_{60}, which is defined as the time for the sound to die away to a level 60 dB below its original level and is measured by the Schroeder's method [19] using the reverse-time integrated impulse response, is approximately 0.12 s. Each set of measured data points is fitted by a third order polynomial curve, displaying a clear trend of dependence of the TDE performance on the number of microphones.

Again when the number of microphones is fixed, the TDE performance degrades as the SNR drops. In the same SNR condition, a better performance is obtained when more microphones are available. It is noted that the probability of anomalies and the bias of nonanomalous estimates obtained in this reverberant environment are similar to those achieved in the nonreverberant condition. This is due to the fact that even though noise and reverberation coexist, in the studied SNR conditions, the background noise is the dominant distortion source that degrades the TDE performance.

As seen from Fig. 5, the probability of anomalies decreased monotonously as the number of microphones is increased. The trend of bias is in general downwards as more microphones are employed. We notice however, in some occasional situations such as eight microphones, its bias is slightly higher than that for seven microphones. This is mainly because that when eight as opposed to seven microphones are used, some anomalous estimates will become nonanomalies. This part of nonanomalies due to one more microphones may have a higher bias than other nonanomalies.



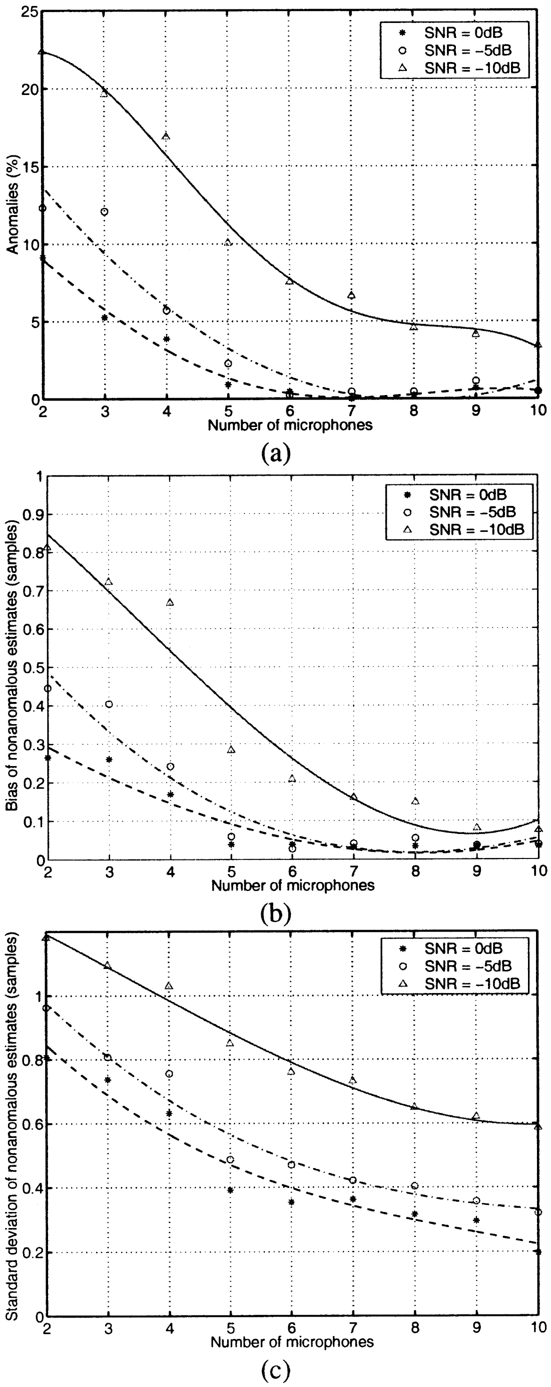
(a)

(b)

(c)

Fig. 5. Percentage of (a) anomalous time delay estimates, (b) bias, and (c) standard deviation of nonanomalous time delay estimates versus different number of microphones in reverberation conditions where $r_i = 0.5, i = 1, 2, \ldots, 6$. The fitting curve is a third order polynomial.

Like the probability of anamolies and the bias, when the SNR is fixed, the standard deviation of the nonanomalous estimates in this light reverberation circumstance also reduces when the number of microphones is increased. It is interesting to notice that, by comparing Fig. 5 with Fig. 4, the standard deviation obtained in the reverberant environment is higher than that in nonreverberant condition and the former reduces slower than the latter as the number of microphones increases. Even worse, the deviation in the reverberation condition does not vary much when more than six microphones are employed. This is understandable. In the reverberant environment, reflected signals with



(a)

(b)

(c)

Fig. 6. Percentage of (a) anomalous time delay estimates, (b) bias, and (c) standard deviation of nonanomalous time delay estimates versus different number of microphones in highly reverberant environments where boundary $r_i = 0.8, i = 1, 2, \ldots, 6$. The fitting curve is a third-order polynomial.

different delay reach the microphone sensors, which will shift the peaks (or valleys) of the cost function.

In many practical situations, the boundary reflection coefficients are very likely to be greater than 0.5. The reverberation time is therefore much longer than 0.1 s. The third experiment tests the TDE performance in a moderate condition where $r_1 = r_2 = \cdots = r_6 = 0.8$ respectively. The corresponding reverberation time T_{60} is approximately 0.23 s. The results obtained are plotted in Fig. 6. Again, a third-order polynomial curve is fitted to the data to display the trend of dependence of TDE performance on the number of microphones.

It can be seen that, in the same SNR condition, this moderate reverberation condition exhibits much higher percentage of anomalies as compared to the nonreverberant and lightly reverberant environments. This is due to the fact that as the boundary reflection coefficients grow, more reflected signals will reach the microphones with a stronger level and a different delay. As a result, the erroneous peaks of the cost function increase, which will eventually lead to mistakes in extremum searching. Likewise, the bias and deviation in this moderate reverberation condition are larger than that obtained in previous experiments.

It is worthwhile pointing out that even in this stronger reverberation situation, a better performance is achieved as more microphones are used. This again confirms the effectiveness of the multichannel correlation approach in fully utilizing the redundant information provided by multiple sensors to eliminate the effect of noise and reverberation.

VII. Conclusion

Although many research efforts have been devoted to it, time delay estimation remains to be a difficult problem in practical noisy and reverberant environments.

In this paper, the linear spatial prediction and linear spatial interpolation techniques were readdressed from the point view of time delay estimation. The spatial correlation matrix is then introduced and its properties were discussed.

The spatial correlation matrix can be written in different ways. We proposed a way which had included some a priori information of the microphone array geometry and the relation among the different time delays. Given the relative delay, τ, between microphones 0 and 1, we supposed that the relative delay between microphone 0 and l is a function of τ. Thus, if τ is known, any microphone signal can be predicted from the others. This can be useful for multichannel coding. If τ is not known, it can be estimated by minimizing the spatial prediction error or, equivalently, by using the determinant of the spatial correlation matrix. It was shown that this multichannel correlation TDE algorithm is a generalized version of the popularly used GCC method. The advantage of the new approach, as compared to the GCC method, is that it can take into account the redundant information provided by multiple microphones. Experimental results demonstrated that this redundancy can make the estimation of τ more robust to noise and reverberation.

Acknowledgment

The authors would like to thank Dr. G. W. Elko for very helpful discussions during the development of this work. They are also grateful to D. R. Morgan for carefully reading a draft and providing many constructive comments and suggestions that have improved the clarity of this paper.

References

[1] D. R. Fischell and C. H. Coker, “A speech direction finder,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 1984, pp. 19.8.1–19.8.4.

[2] 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 Trans. Signal Processing*, vol. 50, pp. 1843–1854, Aug. 2002.

[3] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. II, 1994, pp. 273–276.

[4] ——, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 1996, pp. 921–924.

[5] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.

[6] Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Norwell, MA: Kluwer, 2000, ch. 11, pp. 239–259.

[7] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 943–956, Nov. 2001.

[8] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 320–327, Aug. 1976.

[9] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 998–1003, Dec. 1982.

[10] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in presence of room reverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 148–152, Mar. 1996.

[11] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.

[12] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for time delay estimation under reverberant conditions," *Signal Process.*, vol. 59, pp. 253–266, 1997.

[13] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[14] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, pp. 384–391, Jan. 2000.

[15] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, pp. 11–24, Jan. 2003.

[16] S. Haykin, "Radar array processing for angle of arrival estimation," in *Array Signal Process.*, S. Haykin, Ed: Prentice-Hall, 1985, ch. 4, pp. 194–292.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: John Hopkins Univ. Press, 1996.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[19] M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

Jingdong Chen (M'99) 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. He studied and proposed several techniques covering speech enhancement and HMM adaptation by signal transformation.

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis and 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 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.

Dr. Chen 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 (M'92) was born in Marrakech, Morocco, in 1963. He received the Masters degree in microwaves from Pierre and 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. program (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 on multichannel adaptive filters and acoustic echo cancellation. He joined Bell Labs, Lucent Technologies (formerly AT&T) in October 1995, first as a Consultant and then as a Member of Technical Staff. Since then, he has been working on stereophonic acoustic echo cancellation, adaptive filters, source localization, robust network echo cancellation, and blind deconvolution. He co-authored the book *Advances in Network and Acoustic Echo Cancellation* (Berlin, Germany: Springer-Verlag, 2001). He is also a co-editor/co-author of the books *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003) and *Acoustic Signal Processing for Telecommunication* (Norwell, MA: Kluwer, 2000).

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control.

Yiteng (Arden) Huang (S'97–M'01) received the B.S. degree from 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.

During his Ph.D. studies from 1998 to 2001, he was a Research Assistant with the Center of Signal and Image Processing, Georgia Tech, and was a Teaching Assistant with the School of Electrical and Computer Engineering, Georgia Tech. In the summers from 1998 to 2000, he worked with Bell Laboratories, Murray Hill, NJ, and engaged in research on passive acoustic source localization with microphone arrays. Upon graduation, he joined Bell Laboratories as a Member of 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. He is a co-editor/co-author of the book *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003).

Dr. Huang is currently an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000–2001 Outstanding Graduate Teching Assistant Award from the School 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.