Time-Delay Estimation via Linear Interpolation and Cross Correlation

Jacob Benesty, *Senior Member, IEEE*, Jingdong Chen, *Member, IEEE*, and Yiteng Huang, *Member, IEEE*

***Abstract*—Time-delay estimation (TDE), which aims at measuring the relative time difference of arrival (TDOA) between different channels is a fundamental approach for identifying, localizing, and tracking radiating sources. Recently, there has been a growing interest in the use of TDE based locator for applications such as automatic camera steering in a room conferencing environment where microphone sensors receive not only the direct-path signal, but also attenuated and delayed replicas of the source signal due to reflections from boundaries and objects in the room. This multipath propagation effect introduces echoes and spectral distortions into the observation signal, termed as reverberation, which severely deteriorates a TDE algorithm in its performance. This paper deals with the TDE problem with emphasis on combating reverberation using multiple microphone sensors. The multichannel cross correlation coefficient (MCCC) is rederived here, in a new way, to connect it to the well-known linear interpolation technique. Some interesting properties and bounds of the MCCC are discussed and a recursive algorithm is introduced so that the MCCC can be estimated and updated efficiently when new data snapshots are available. We then apply the MCCC to the TDE problem. The resulting new algorithm can be treated as a natural generalization of the generalized cross correlation (GCC) TDE method to the multichannel case. It is shown that this new algorithm can take advantage of the redundancy provided by multiple microphone sensors to improve TDE against both reverberation and noise. Experiments confirm that the relative time-delay estimation accuracy increases with the number of sensors.**

***Index Terms*—Cross correlation, cross-correlation coefficient, linear interpolation, multichannel, time-delay estimation (TDE).**

I. Introduction

TIME-DELAY estimation (TDE), which aims at measuring the relative time difference of arrival (TDOA) among spatially separated sensors, has played an important role in radar, sonar, and seismology for localizing radiating sources. Traditional methods for TDE are based on a "measure" of the cross-correlation from measurements made with an array of sensors. Nowadays, we use the same kind of techniques to localize and track acoustic sources in a room environment for applications such as automatic camera tracking for video-conferencing [1]–[3] and microphone array beam steering [4]–[9] for suppressing noise and reverberation in various communication and voice processing systems.

The generalized cross-correlation (GCC) method, proposed by Knapp and Carter in 1976 [10], is the most popular technique for TDE. The delay estimate between two sensors is obtained as the time-lag that maximizes the cross-correlation between filtered versions of the received signals. This method is well studied and it performs fairly well in moderately noisy and nonreverberant environments [11], [12]. However, this method tends to break down when applied to a microphone array system in a room environment where TDE becomes more complicated owing to the sophisticated reverberation effect. Many new ideas have been recently proposed to better deal with noise and reverberation by taking advantage of the nature of a speech signal [13], [14], by utilizing redundant information from multiple sensor pairs [15], or from the blind channel identification point of view [16], [17]. However, reverberation remains a problem and in a highly reverberant room, all known methods fail. One important problem we try to tackle in this paper is how the GCC method can be generalized to the multichannel case (more than two processes). The objective of this generalization is to take advantage of the redundancy available from multiple sensors to make the estimator more robust to noise and reverberation.

The definition of the cross-correlation coefficient is very well-known and widely used in signal processing. In this paper, we redefine the multichannel cross-correlation coefficient (MCCC) in a way that connects it to the well-known linear interpolation technique. Some interesting properties and bounds of the MCCC will be discussed. A recursive algorithm is introduced so that the MCCC can be estimated and updated efficiently when new data snapshots are available. We will show in detail how the MCCC can be used for TDE and many simulations will confirm that the relative delay estimation accuracy increases with the number of sensors.

II. Linear Interpolation

We assume that we have L signals $x_0(n), x_1(n), \ldots, x_{L-1}(n)$, and we seek to determine how any one of these signals can be interpolated from the others. To interpolate $x_i(n)$ from the rest, we need to minimize the criterion [18], [19]

$$\begin{aligned} J_i(n) &= \sum_{p=0}^{n} \lambda^{n-p} \left[-\sum_{l=0}^{L-1} c_{il}(n) x_l(p) \right]^2 \\ &= \sum_{p=0}^{n} \lambda^{n-p} \left[-\mathbf{c}_i^T(n)\mathbf{x}(p) \right]^2 \\ &= \mathbf{c}_i^T(n)\mathbf{R}(n)\mathbf{c}_i(n) \end{aligned} \tag{1}$$

Manuscript received September 22, 2003; revised March 6, 2004. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. M. Mohan Sondhi.

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

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

Digital Object Identifier 10.1109/TSA.2004.833008

1063-6676/04$20.00 © 2004 IEEE

with the constraint

$$\mathbf{c}_i^T(n)\mathbf{u}_i = c_{ii} = -1 \tag{2}$$

where $\lambda (0 < \lambda \leq 1)$ is a forgetting factor

$$\mathbf{c}_i(n) = \left[c_{i0}(n)\; c_{i1}(n)\; \ldots\; c_{i(L-1)}(n)\right]^T$$

is a vector used to compute the interpolation error, (this vector without the component c_{ii} is the ith $(0 \leq i \leq L-1)$ interpolator of the vector signal)

$$\begin{aligned}\mathbf{x}(n) &= [x_0(n) \quad x_1(n) \quad \ldots \quad x_{L-1}(n)]^T\\ \mathbf{u}_i &= [0 \quad \ldots \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0]^T\end{aligned}$$

is a vector of length L where its ith component is equal to one and all others are zero, and

$$\mathbf{R}(n) = \sum_{p=0}^{n} \lambda^{n-p}\mathbf{x}(p)\mathbf{x}^T(p) \tag{3}$$

is an estimate of the signal covariance matrix. Matrix $\mathbf{R}(n)$ is positive semi-definite; but in the rest, we suppose that it is positive definite so it is invertible.

By using a Lagrange multiplier, it is easy to see that the solution to this optimization problem is

$$\mathbf{R}(n)\mathbf{c}_i(n) = -E_i(n)\mathbf{u}_i \tag{4}$$

where

$$E_i(n) = \mathbf{c}_i^T(n)\mathbf{R}(n)\mathbf{c}_i(n) = \frac{1}{\mathbf{u}_i^T\mathbf{R}^{-1}(n)\mathbf{u}_i} \tag{5}$$

is the interpolation error energy.

Since

$$-\frac{\mathbf{c}_i(n)}{E_i(n)} = \mathbf{R}^{-1}(n)\mathbf{u}_i \tag{6}$$

then the ith column of $\mathbf{R}^{-1}(n)$ is $-\mathbf{c}_i(n)/E_i(n)$. We deduce that $\mathbf{R}^{-1}(n)$ can be factorized as follows:

$$\begin{aligned}\mathbf{R}^{-1}(n) =& \begin{bmatrix} 1 & -c_{10}(n) & \ldots & -c_{(L-1)0}(n)\\ -c_{01}(n) & 1 & \ldots & -c_{(L-1)1}(n)\\ \vdots & \vdots & \ddots & \vdots\\ -c_{0(L-1)}(n) & -c_{1(L-1)}(n) & \ldots & 1\end{bmatrix}\\ &\times \begin{bmatrix}\frac{1}{E_0(n)} & 0 & \cdot & 0\\ 0 & \frac{1}{E_1(n)} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{E_{L-1}(n)}\end{bmatrix}\\ &\triangleq \mathbf{C}^T(n)\mathbf{D}_E^{-1}(n).\end{aligned} \tag{7}$$

Since $\mathbf{R}^{-1}(n)$ is a symmetric matrix, (7) becomes

$$\begin{aligned}\mathbf{R}^{-1}(n) =& \begin{bmatrix}\frac{1}{E_0(n)} & 0 & \cdot & 0\\ 0 & \frac{1}{E_1(n)} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{E_{L-1}(n)}\end{bmatrix}\\ &\times \begin{bmatrix} 1 & -c_{01}(n) & \ldots & -c_{0(L-1)}(n)\\ -c_{10}(n) & 1 & \ldots & -c_{(L-1)1}(n)\\ \vdots & \vdots & \ddots & \vdots\\ -c_{(L-1)0}(n) & -c_{(L-1)1}(n) & \ldots & 1\end{bmatrix}\\ &\triangleq \mathbf{D}_E^{-1}(n)\mathbf{C}(n).\end{aligned} \tag{8}$$

The first and last columns of $\mathbf{R}^{-1}(n)$ contain respectively the normalized forward and backward predictors and all the columns between contain the normalized interpolators. $\mathbf{C}(n)$ is simply the matrix of the interpolators and $\mathbf{D}_E(n)$ is a diagonal matrix containing all the respective interpolation error energies.

We define, respectively, the *a priori* and *a posteriori* interpolation error signals as

$$e_i(n) \triangleq -\mathbf{c}_i^T(n-1)\mathbf{x}(n) \tag{9}$$

$$\varepsilon_i(n) \triangleq -\mathbf{c}_i^T(n)\mathbf{x}(n). \tag{10}$$

Using expression (8), we now define the *a priori* and *a posteriori* Kalman gain vectors

$$\begin{aligned}\mathbf{k}'(n) &\triangleq \mathbf{R}^{-1}(n-1)\mathbf{x}(n)\\ &= \left[\frac{e_0}{E_0(n-1)} \quad \frac{e_1(n)}{E_1(n-1)} \quad \cdots \quad \frac{e_{L-1}(n)}{E_{L-1}(n-1)}\right]^T\end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{k}(n) &\triangleq \mathbf{R}^{-1}(n)\mathbf{x}(n)\\ &= \left[\frac{\varepsilon_0}{E_0(n)} \quad \frac{\varepsilon_1(n)}{E_1(n)} \quad \cdots \quad \frac{\varepsilon_{L-1}(n)}{E_{L-1}(n)}\right]^T.\end{aligned} \tag{12}$$

The ith component of the *a priori* (resp. *a posteriori*) Kalman gain vector is the ith *a priori* (resp. *a posteriori*) interpolation error signal normalized with the ith interpolation error energy at time $n-1$ (resp. n). From (3), we can derive the following recursion:

$$\lambda\mathbf{R}(n-1) = \mathbf{R}(n) - \mathbf{x}(n)\mathbf{x}^T(n). \tag{13}$$

Using this recursion, it can be shown that the *a posteriori* Kalman gain vector is related to $\mathbf{k}'(n)$ by [20]

$$\mathbf{k}(n) = \lambda^{-1}\varphi(n)\mathbf{k}'(n) \tag{14}$$

where

$$\begin{aligned}\varphi(n) &= \frac{\lambda}{\lambda + \mathbf{x}^T(n)\mathbf{R}^{-1}(n-1)\mathbf{x}(n)}\\ &= 1 - \mathbf{x}^T(n)\mathbf{R}^{-1}(n)\mathbf{x}(n).\end{aligned} \tag{15}$$

Writing (4) at time n and $n-1$, we obtain

$$-\frac{\mathbf{R}(n)\mathbf{c}_i(n)}{E_i(n)} = \mathbf{u}_i = -\frac{\lambda\mathbf{R}(n-1)\mathbf{c}_i(n-1)}{\lambda E_i(n-1)}. \tag{16}$$

Replacing in (16), $\lambda\mathbf{R}(n-1)$ by the right-hand side of (13), we get

$$\mathbf{c}_i(n) = \frac{E_i(n)}{\lambda E_i(n-1)}\left[\mathbf{c}_i(n-1) + \mathbf{k}(n)e_i(n)\right]. \tag{17}$$

Now, if we premultiply both sides of (17) by $\mathbf{u}_i^T$ and utilize (2) and (12), we find that

$$E_i(n) = \lambda E_i(n-1) + e_i(n)\varepsilon_i(n). \tag{18}$$

This means that the interpolation error energy can be computed recursively. This relation is well-known for the forward $(i=0)$ and backward $(i=L)$ predictors [20]. It is used to obtain fast versions of the recursive least-squares (RLS) algorithm.

Also, the interpolator vectors can be computed recursively

$$\mathbf{c}_i(n) = \frac{1}{1-k_i(n)e_i(n)}\left[\mathbf{c}_i(n-1) + \mathbf{k}(n)e_i(n)\right]. \tag{19}$$

If we premultiply both sides of (19) by $-\mathbf{x}^T(n)$, we obtain a relation between the *a priori* and *a posteriori* interpolation error signals

$$\frac{\varepsilon_i(n)}{e_i(n)} = \frac{\varphi(n)}{1-k_i(n)e_i(n)}. \tag{20}$$

III. Multichannel Cross-Correlation Coefficient

The definition of multiple coherence function, derived from the concepts of the ordinary coherence function between two signals and the partial (conditioned) coherence function, was presented in [21] to measure the correlation between the output of a MISO (multiple-input/single-output) system and its inputs. In this section, we derive the multichannel cross-correlation coefficient (MCCC) in a new way such that it is related to the multichannel correlation matrix. We show that with our definition, the MCCC can be treated as a generalization of the classical cross-correlation coefficient to the case where we have more than two processes. We will factorize $\mathbf{R}(n)$ and $\mathbf{R}^{-1}(n)$ and deduce some interesting properties related to cross-correlation and linear interpolation. The covariance matrix can be factorized as follows:

$$\mathbf{R}(n) = \mathbf{D}_x^{\frac{1}{2}}(n)\widetilde{\mathbf{R}}(n)\mathbf{D}_x^{\frac{1}{2}}(n) \tag{21}$$

where

$$\mathbf{D}_x^{\frac{1}{2}}(n) = \begin{bmatrix} \sqrt{r_{00}(n)} & 0 & \cdots & 0 \\ 0 & \sqrt{r_{11}(n)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{r_{(L-1)(L-1)}(n)} \end{bmatrix} \tag{22}$$

$$\widetilde{\mathbf{R}}(n) = \begin{bmatrix} 1 & \rho_{10}(n) & \cdots & \rho_{(L-1)0}(n) \\ \rho_{10}(n) & 1 & \cdots & \rho_{(L-1)1}(n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{(L-1)0}(n) & \rho_{(L-1)1}(n) & \cdots & 1 \end{bmatrix} \tag{23}$$

$$r_{ij}(n) = \sum_{p=0}^{n}\lambda^{n-p}x_i(p)x_j(p), \quad i,j=0,1,\ldots,L-1 \tag{24}$$

and

$$\rho_{ij}(n) = \frac{r_{ij}(n)}{\sqrt{r_{ii}(n)r_{ij}(n)}}, \quad i,j=0,1,\ldots,L-1. \tag{25}$$

$\rho_{ij}(n)$ is the cross-correlation coefficient between $x_i(n)$ and $x_j(n)$.

Since matrix $\widetilde{\mathbf{R}}(n)$ is symmetric, positive definite, and its diagonal elements are all equal to one, it can be shown that (see Appendix)

$$0 < \det\left[\widetilde{\mathbf{R}}(n)\right] \le 1 \tag{26}$$

where "det" stands for determinant.

We can now generalize the definition of squared cross-correlation coefficient to the multichannel case. We define the squared MCCC among the L signals $x_0(n), x_1(n), \ldots, x_{L-1}(n)$, as

$$\rho_L^2(n) \triangleq 1 - \det\left[\widetilde{\mathbf{R}}(n)\right] = 1 - \frac{\det\left[\mathbf{R}(n)\right]}{\prod_{l=0}^{L-1} r_{ll}(n)}. \tag{27}$$

This definition is identical to the one given in [22], [23] using the Gram determinant. For two $(L=2)$ processes $x_0(n)$ and $x_1(n)$, we have

$$\rho_2^2(n) = \frac{r_{01}^2(n)}{r_{00}(n)r_{11}(n)} \tag{28}$$

which is the classical definition of the squared cross-correlation coefficient.

We have the following properties [23].

- $0 \le \rho_L^2(n) \le 1$ (the case $\rho_L^2(n) = 1$ happens when matrix $\mathbf{R}(n)$ is nonnegative definite).
- If two or more signals are perfectly correlated, then $\rho_L^2(n) = 1$.
- If all the processes are completely uncorrelated with each other, then $\rho_L^2(n) = 0$.
- If one of the signals is completely uncorrelated with the $L-1$ other signals, then the MCCC will measure the correlation among those $L-1$ remaining signals.

The inverse covariance matrix can be factorized as follows:

$$\mathbf{R}^{-1}(n) = \mathbf{D}_E^{-\frac{1}{2}}(n)\widetilde{\mathbf{C}}(n)\mathbf{D}_E^{-\frac{1}{2}}(n) \tag{29}$$

[See (30) at bottom of next page]. Clearly, $\widetilde{\mathbf{C}}(n)$ is symmetric and

$$\sqrt{\frac{E_j}{E_i}}c_{ij}(n) = \sqrt{\frac{E_i}{E_j}}c_{ji}(n), \quad i,j=0,1,\ldots,L-1. \tag{31}$$

Since matrix $\widetilde{\mathbf{C}}(n)$ is symmetric, positive definite, and its diagonal elements are all equal to one, we also have

$$0 < \det\left[\widetilde{\mathbf{C}}(n)\right] \leq 1. \tag{32}$$

From (21) and (29), we find

$$\det\left[\widetilde{\mathbf{R}}(n)\right]\det\left[\widetilde{\mathbf{C}}(n)\right] = \prod_{l=0}^{L-1}\frac{E_l(n)}{r_{ll}(n)}. \tag{33}$$

We then get interesting bounds

$$\prod_{l=0}^{L-1}\frac{E_l(n)}{r_{ll}(n)} \leq \det\left[\widetilde{\mathbf{R}}(n)\right] \leq 1 \tag{34}$$

$$\prod_{l=0}^{L-1}\frac{E_l(n)}{r_{ll}(n)} \leq \det\left[\widetilde{\mathbf{C}}(n)\right] \leq 1. \tag{35}$$

Hence

$$0 \leq \rho_L^2(n) \leq 1 - \prod_{l=0}^{L-1}\frac{E_l(n)}{r_{ll}(n)} \leq 1. \tag{36}$$

Moreover, (5) can be rewritten the following way:

$$\begin{aligned} E_i(n) &= \frac{1}{\mathbf{u}_i^T\mathbf{R}^{-1}(n)\mathbf{u}_i} \\ &= \frac{r_{ii}(n)}{\mathbf{u}_i^T\widetilde{\mathbf{R}}^{-1}(n)\mathbf{u}_i} \\ &= r_{ii}(n)\frac{\det\left[\widetilde{\mathbf{R}}(n)\right]}{\det\left[\widetilde{\mathbf{R}}_{-i}(n)\right]} \end{aligned} \tag{37}$$

where $\widetilde{\mathbf{R}}_{-i}(n)$ is a matrix of size $(L-1)\times(L-1)$ obtained from $\widetilde{\mathbf{R}}(n)$ by removing its ith row and ith column. As a result, we have

$$\begin{aligned} \rho_L^2(n) &= 1 - \frac{E_i(n)}{r_{ii}(n)}\det\left[\widetilde{\mathbf{R}}_{-i}(n)\right] \\ &\geq 1 - \frac{E_i(n)}{r_{ii}(n)} \end{aligned} \tag{38}$$

and the bound in (36) becomes

$$0 \leq 1 - \frac{E_i(n)}{r_{ii}(n)} \leq \rho_L^2(n) \leq 1 - \prod_{l=0}^{L-1}\frac{E_l(n)}{r_{ll}(n)} \leq 1, \quad \forall i \in [0,\ L-1]. \tag{39}$$

Now if we define $\rho_{i:j}^2$ as

$$\rho_{i:j}^2 = 1 - \det\left[\widetilde{\mathbf{R}}_{i:j}(n)\right] \tag{40}$$

where $j \geq i$, and $\widetilde{\mathbf{R}}_{i:j}(n)$ is a $(j-i)\times(j-i)$ matrix whose elements are taken as a block from the (i,i)-position to the (j,j)-position from the matrix $\widetilde{\mathbf{R}}(n)$. We then have

$$\begin{aligned} \rho_{0:L-1}^2(n) &= \rho_L^2(n) \\ &= 1 - \frac{E_0(n)}{r_{00}(n)}\det\left[\widetilde{\mathbf{R}}_{-0}(n)\right] \\ &= 1 - \frac{E_0(n)}{r_{00}(n)}\left[1 - \rho_{1:L-1}^2(n)\right] \end{aligned} \tag{41}$$

where

$$\rho_{1:L-1}^2(n) = 1 - \det\left[\widetilde{\mathbf{R}}_{-0}(n)\right] \tag{42}$$

is the squared MCCC among the $L-1$ signals $x_1(n), x_2(n), \ldots, x_{L-1}(n)$. The interesting thing about (41) is that it gives a relation on the order of the MCCC, with $\rho_{L-1:L-1}^2(n) = 0$. Expression (42) can also be rewritten as

$$\rho_{1:L-1}^2(n) = 1 - \frac{E_{1,1:L-1}(n)}{r_{11}(n)}\det\left[\widetilde{\mathbf{R}}_{-0:1}(n)\right] \tag{43}$$

where $E_{1,1:L-1}(n)$ is the forward prediction error energy (of order $L-2$) using the signals $x_1(n), x_2(n), \ldots, x_{L-1}(n)$ and $\widetilde{\mathbf{R}}_{-0:1}(n)$ is a matrix of size $(L-2)\times(L-2)$ obtained from $\widetilde{\mathbf{R}}(n)$ by removing its first two rows and columns. Replacing (43) in (41), we obtain

$$\rho_{0:L-1}^2(n) = 1 - \frac{E_0(n)E_{1,1:L-1}(n)}{r_{00}(n)r_{11}(n)}\det\left[\widetilde{\mathbf{R}}_{-0:1}(n)\right] \tag{44}$$

and continuing the same process, we finally have

$$\rho_{0:L-1}^2(n) = 1 - \prod_{l=0}^{L-1}\frac{E_{l,l:L-1}(n)}{r_{ll}(n)} \tag{45}$$

where $E_{l,l:L-1}(n)$ is the forward prediction error energy (of order $L-1-l$) using the signals $x_l(n), x_{l+1}(n), \ldots, x_{L-1}(n)$ and $E_{0,0:L-1}(n) = E_0(n)$. This shows how the MCCC is related to the different orders of the forward linear prediction energies. The same principle can be shown with the different orders of the backward linear prediction energies

$$\rho_{L-1:0}^2(n) = \rho_L^2(n) = 1 - \prod_{l=0}^{L-1}\frac{E_{L-1-l,L-1-l:0}(n)}{r_{ll}(n)} \tag{46}$$

$$\widetilde{\mathbf{C}}(n) = \begin{bmatrix} 1 & -\sqrt{\frac{E_1}{E_0}}c_{01}(n) & \cdots & -\sqrt{\frac{E_{L-1}}{E_0}}c_{0(L-1)}(n) \\ -\sqrt{\frac{E_0}{E_1}}c_{10}(n) & 1 & \cdots & -\sqrt{\frac{E_{L-1}}{E_1}}c_{1(L-1)}(n) \\ \vdots & \vdots & \ddots & \vdots \\ -\sqrt{\frac{E_0}{E_{L-1}}}c_{(L-1)0}(n) & -\sqrt{\frac{E_1}{E_{L-1}}}c_{(L-1)1}(n) & \cdots & 1 \end{bmatrix} \tag{30}$$

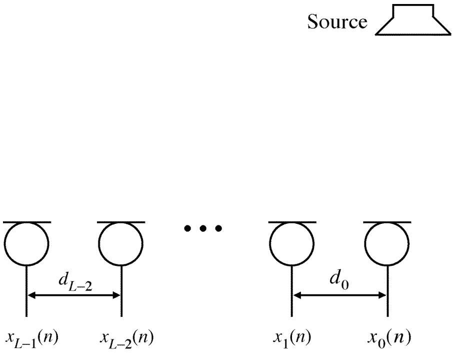


Fig. 1. Linear microphone array.

where $E_{L-1-l,L-1-l:0}(n)$ is the backward prediction error energy (of order $L-1-l$) and $E_{L-1,L-1:0}(n) = E_{L-1}(n)$. Obviously, we can generalize this approach to any linear interpolator.

For two $(L = 2)$ processes $x_0(n)$ and $x_1(n)$, we have

$$\rho_2^2(n) = \frac{r_{01}^2(n)}{r_{00}(n)r_{11}(n)} = 1 - \frac{E_0(n)}{r_{00}(n)} = 1 - \frac{E_1(n)}{r_{11}(n)}. \tag{47}$$

IV. Application to Time-Delay Estimation

A. Signal Model

Suppose that we have a linear array as shown in Fig. 1, which consists of L microphones whose outputs are denoted as $x_l(n)$, $l = 0, 1, \ldots, L-1$. Without loss of generality, we select microphone 0 as the reference point and consider the following propagation model:

$$\begin{bmatrix} x_0(n) \\ x_1(n) \\ x_2(n) \\ \vdots \\ x_{L-1}(n) \end{bmatrix} = \begin{bmatrix} \alpha_0 & 0 & 0 & \ldots & 0 \\ 0 & \alpha_1 & 0 & \ldots & 0 \\ 0 & 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & 0 & \alpha_{L-1} \end{bmatrix} \times \begin{bmatrix} s(n-t) \\ s(n-t-\tau) \\ s\left[n-t-f_2(\tau)\right] \\ \vdots \\ s\left[n-t-f_{L-1}(\tau)\right] \end{bmatrix} + \begin{bmatrix} w_0(n) \\ w_1(n) \\ w_2(n) \\ \vdots \\ w_{L-1}(n) \end{bmatrix} \tag{48}$$

where α_l, $l = 0, 1, 2, \ldots, L-1$, 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 on τ but also on the microphone array geometry. It can be specified for arbitrary arrays in one, two, or three dimensions. In this paper we are considering only linear arrays. In the far-field case (i.e., plane wave propagation), if the array is equispaced, we have

$$f_l(\tau) = l\tau \tag{49}$$

and if it is not equispaced, we have

$$f_l(\tau) = \frac{\sum_{i=0}^{l-1} d_i}{d_0}\tau \tag{50}$$

where d_i is the distance between microphones i and $i+1$, $i = 0, 1, 2, \ldots, L-2$. In the near-field case, f_l depends also on the position of the source. Again in this paper, we focus only on the far-field case. In such a situation, 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 $w_l(n)$ is a zero-mean Gaussian random process that is uncorrelated with $s(n)$ and the noise signals at other microphones. It is also assumed that $s(n)$ is reasonably broad-band.

B. Two-Channel Case

Consider the two signals $x_0(n)$ and $x_1(n+m)$ where m is an integer. The value of m that gives the maximum

$$\rho_2^2(n,m) = \frac{r_{01}^2(n,m)}{r_{00}(n)r_{11}(n,m)} \tag{51}$$

where

$$r_{01}(n,m) = \sum_{p=0}^{n} \lambda^{n-p} x_0(p) x_1(p+m) \tag{52}$$

$$r_{11}(n,m) = \sum_{p=0}^{n} \lambda^{n-p} x_1^2(p+m) \tag{53}$$

corresponds to the time-delay between microphones 0 and 1. Mathematically, the solution to our problem is then given by

$$\hat{\tau} = \arg\max\nolimits_{\mathrm{m}} \rho_2^2(n,m) \tag{54}$$

where $\hat{\tau}$ is an estimate of τ, $m \in [-\tau_{\max}, \tau_{\max}]$, and $\tau_{\max}$ is the maximum possible delay. 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 in-phase, i.e. $m \approx \tau$. This approach is similar to the generalized cross-correlation method proposed by Knapp and Carter [10].

C. Multichannel Case

We are interested in estimating only one time-delay (τ) from multiple sensors. Obviously, two sensors are enough to estimate τ. However, the redundant information that is available when more than two sensors are used, will help to improve the estimator, especially in the presence of high level of noise and reverberation.

Consider the following vector:

$$\mathbf{x}(n,m) = \left[x_0(n) \;\; x_1[n+f_1(m)] \;\; \ldots \;\; x_{L-1}[n+f_{L-1}(m)]\right]^T.$$

We can check that for $m = \tau$, all the signals $x_l[n + f_l(\tau)]$, $l = 0, 1, \ldots, L-1$, are aligned. This observation is essential because it already gives an idea on how to estimate τ. An estimate of the covariance matrix corresponding to the signal $\mathbf{x}(n,m)$ is

$$\begin{aligned}\mathbf{R}(n,m) &= \sum_{p=0}^{n} \lambda^{n-p}\mathbf{x}(p,m)\mathbf{x}^T(p,m)\\ &= \lambda\mathbf{R}(n-1,m) + \mathbf{x}(n,m)\mathbf{x}^T(n,m).\end{aligned} \quad (55)$$

Therefore, the squared MCCC is

$$\rho_L^2(n,m) = 1 - \frac{\det\left[\mathbf{R}(n,m)\right]}{\prod_{l=0}^{L-1} r_{ll}(n,m)} \quad (56)$$

where

$$r_{ll}(n,m) = \sum_{p=0}^{n} \lambda^{n-p} x_l^2\left[p + f_l(m)\right], \quad l = 0,1,\ldots,L-1. \quad (57)$$

The value of m that gives the maximum of $\rho_L^2(n,m)$, for different m, corresponds to the time-delay between microphones 0 and 1. Hence, the solution to our problem is

$$\hat{\tau} = \arg\max_{m} \rho_L^2(n,m) \quad (58)$$

where again $m \in [-\tau_{\max}, \tau_{\max}]$, and $\tau_{\max}$ is the maximum possible delay. This 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 with more than two sensors (instead of estimating multiple time-delays independently) in an optimal way in a least-squares sense.

D. Recursive Estimation of the Squared MCCC

From Sections II and III, we can see that there are many different ways to estimate the squared MCCC. Here we propose to estimate the elements of $\rho_L^2(n,m)$ recursively. The recursive estimation of $r_{ll}(n,m)$ is straightforward. Indeed, we have

$$r_{ll}(n,m) = \lambda r_{ll}(n-1,m) + x_l^2\left[n + f_l(m)\right], \; l=0,1,\ldots,L-1 \quad (59)$$

it is then easy to compute $\prod_{l=0}^{L-1} r_{ll}(n,m)$.

From (55), we have

$$\begin{aligned}&\frac{1}{\lambda}\mathbf{R}(n,m)\mathbf{R}^{-1}(n-1,m)\\ &\qquad = \mathbf{I} + \frac{1}{\lambda}\mathbf{x}(n,m)\mathbf{x}^T(n,m)\mathbf{R}^{-1}(n-1,m).\end{aligned} \quad (60)$$

One can notice that the right-hand side of (60) is of the form $\mathbf{I} + \mathbf{y}\mathbf{z}^T$. So it has one eigenvalue equal to $1 + \mathbf{y}^T\mathbf{z}$, and the rest all equal to unity. The determinant, which is the product of all the eigenvalues is therefore equal to $1 + \mathbf{y}^T\mathbf{z}$. We then have

$$\begin{aligned}&\frac{1}{\lambda^L}\det\left[\mathbf{R}(n,m)\mathbf{R}^{-1}(n-1,m)\right]\\ &\quad = \det\left[\mathbf{I} + \frac{1}{\lambda}\mathbf{x}(n,m)\mathbf{x}^T(n,m)\mathbf{R}^{-1}(n-1,m)\right]\\ &\quad = 1 + \frac{1}{\lambda}\mathbf{x}^T(n,m)\mathbf{R}^{-1}(n-1,m)\mathbf{x}(n,m)\\ &\quad = \frac{1}{\varphi(n,m)}\end{aligned} \quad (61)$$

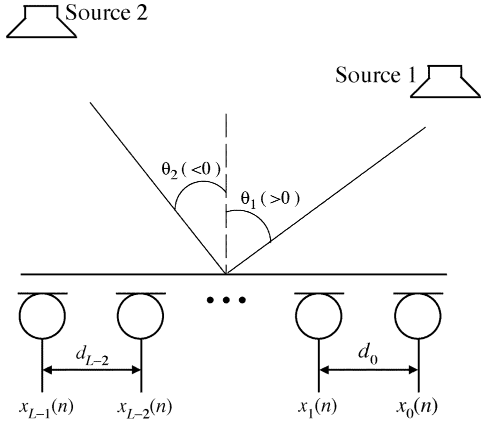


Fig. 2. Linear microphone array in a multiple-source situation.

and finally we get

$$\det\left[\mathbf{R}(n,m)\right] = \frac{\lambda^L}{\varphi(n,m)}\det\left[\mathbf{R}(n-1,m)\right]. \quad (62)$$

The inverse of matrix $\mathbf{R}(n,m)$ that appears in $\varphi(n,m)$ can also be calculated recursively

$$\begin{aligned}\mathbf{R}^{-1}(n,m) = \lambda^{-1}&\mathbf{R}^{-1}(n-1,m)\\ &-\lambda^{-2}\varphi(n,m)\mathbf{k}'(n,m)\mathbf{k}'^T(n,m)\end{aligned} \quad (63)$$

where

$$\mathbf{k}'(n,m) = \mathbf{R}^{-1}(n-1,m)\mathbf{x}(n,m). \quad (64)$$

V. Time-Delay Estimation of Multiple Sources

time-delay estimation of multiple sources using microphone arrays is a difficult problem. Consider a generic scenario where there are M sources $s_1(n), s_2(n), \ldots, s_M(n)$, and $M \leq L$ (This means that we have at least as many microphones as sources). We assume that the M sources are mutually uncorrelated. This assumption holds in general. In a real system, time-delay estimation is achieved on a frame-by-frame basis. There may exist some correlation among signals if the frame size is not large enough. However, such effect is neglected here.

One way to estimate the delays of the M sources is to compute $\rho_L^2(n,m)$, and then search for M largest peaks between maximum negative and maximum positive possible delays, each one corresponding to the time delay of one of the M sources. However, in the search of multiple peaks, errors can be made to discern a fake peak as a real one. If one has some *a priori* knowledge of the minimum angular separation among the M sources, the searching task can be easier. As a special example, we consider a case for two sources, i.e., $M = 2$. Generalization to more than two uncorrelated sources is rather straightforward. For the angular separation constraint, we assume that we know that $s_1(n)$ impinges on the array with a positive bearing and $s_2(n)$ impinges on the array with a negative bearing as shown in Fig. 2. For the source $s_1[n]$, we select microphone 0 as the reference point, and for $s_2[n]$, we choose microphone $L-1$ as the reference. Using the same propagation model as previously, the microphone array vector signal is

$$\mathbf{x}(n) = \mathbf{D}_\alpha \mathbf{s}_1(n) + \mathbf{D}_\beta \mathbf{s}_2(n) + \mathbf{w}(n) \tag{65}$$

where

$$\begin{aligned}
\mathbf{x}(n) &= [x_0(n) \quad x_1(n) \quad \dots \quad x_{L-1}(n)]^T \\
\mathbf{D}_\alpha &= \mathrm{diag}(\alpha_0, \alpha_1, \dots, \alpha_{L-1}) \\
\mathbf{D}_\beta &= \mathrm{diag}(\beta_0, \beta_1, \dots, \beta_{L-1}) \\
\mathbf{s}_1(n) &= [s_1(n-t_1) \quad s_1(n-t_1-\tau_1) \quad s_1\left[n-t_1-f_2(\tau_1)\right] \\
&\qquad \dots \quad s_1\left[n-t_1-f_{L-1}(\tau_1)\right]]^T \\
\mathbf{s}_2(n) &= [s_2\left[n-t_2-g_{L-1}(\tau_2)\right] \quad \dots \quad s_2\left[n-t_2-g_2(\tau_2)\right] \\
&\qquad s_2(n-t_2-\tau_2) \quad s_2(n-t_2)]^T, \\
\mathbf{w}(n) &= [w_0(n) \quad w_1(n) \quad \dots \quad w_{L-1}(n)]^T
\end{aligned}$$

α_l [resp. β_l] are the attenuation factors of source $s_1(n)$ [resp. $s_2(n)$] due to propagation effects, t_1 [resp. t_2] is the propagation time from the unknown source $s_1(n)$ [resp. $s_2(n)$] to microphone 0 [resp. $L-1$], τ_1 [resp. τ_2] is the relative delay between microphones 0 and 1 [resp. $L-2$ and $L-1$], $f_l(\tau_1)$ [resp. $g_{L-1-l}(\tau_2)$] is the relative delay between microphones 0 and l [resp. $L-1$ and $L-1-l$], and $w_l(n)$ is an additive noise signal at the lth microphone.

Consider the following vectors:

$$\begin{aligned}
\mathbf{x}_1(n, m_1) &= [x_0(n) \quad x_1\left[n + f_1(m_1)\right] \\
&\qquad \dots \quad x_{L-1}\left[n + f_{L-1}(m_1)\right]]^T \\
\mathbf{x}_2(n, m_2) &= [x_0\left[n + g_{L-1}(m_2)\right] \quad x_1\left[n + g_{L-2}(m_2)\right] \\
&\qquad \dots \quad x_{L-1}(n)]^T
\end{aligned}$$

where m_1 and m_2 are two positive integers. We can check that for $m_1 = \tau_1$ [resp. $m_2 = \tau_2$] all the signals $x_l[n + f_l(\tau_1)]$, $l = 0, 1, \dots, L-1$ [resp. $x_l[n + g_{L-1-l}(\tau_2)]$, $l = 0, 1, \dots, L-1$], are aligned with respect to $s_1(n)$ [resp. $s_2(n)$]. Estimates of the covariance matrices corresponding to the signals $\mathbf{x}_1(n, m_1)$ and $\mathbf{x}_2(n, m_2)$ are

$$\mathbf{R}_1(n, m_1) = \lambda \mathbf{R}_1(n-1, m_1) + \mathbf{x}_1(n, m_1)\mathbf{x}_1^T(n, m_1) \tag{66}$$

$$\mathbf{R}_2(n, m_2) = \lambda \mathbf{R}_2(n-1, m_2) + \mathbf{x}_2(n, m_2)\mathbf{x}_2^T(n, m_2). \tag{67}$$

We see now that the solution to our problem is

$$\hat{\tau}_1 = \arg\min_{m_1} \det\left[\mathbf{R}_1(n, m_1)\right] \tag{68}$$

$$\hat{\tau}_2 = \arg\min_{m_2} \det\left[\mathbf{R}_2(n, m_2)\right] \tag{69}$$

where $\hat{\tau}_1$ and $\hat{\tau}_2$ are, respectively, estimates of τ_1 and τ_2. With our assumptions and since $\mathbf{R}_1(n, m_1)$ and $\mathbf{R}_2(n, m_2)$ are positive definitive, each one of the two functions $\det[\mathbf{R}_1(n, m_1)]$ and $\det[\mathbf{R}_2(n, m_2)]$ has a unique minimum corresponding to the solutions τ_1 and τ_2. In practice, the more microphone we have, the more obvious the solutions are. Therefore, in principle, we can easily estimate relative delays from two independent sources at the same time by applying a constraint on the angular separation.

VI. Experiments

A. Experimental Setup

The measurements used in this paper were made in the Varechoic chamber at Bell Laboratories [24]. A diagram of the floor plan layout is shown in Fig. 3. For convenience, positions in the floor plan will be designated by (x, y) coordinates with reference to the southwest corner and corresponding to meters along the (South, West) walls. The chamber of size $6.7 \text{ m} \times 6.1 \text{ m} \times 2.9 \text{ m}$ $(x \times y \times z)$ is a room with 368 electronically controlled panels that vary the acoustic absorption of the walls, floor, and ceiling [25]. Each panel consists of two perforated sheets whose holes, if aligned, expose sound absorbing material behind, but if shifted to misalign, form a highly reflective surface. The panels are individually controlled so that the holes on one particular panel are either fully open (absorbing state) or fully closed (reflective state). Therefore, by varying the binary state of each panel in any combination, 2^{368} different room characteristics can be simulated.

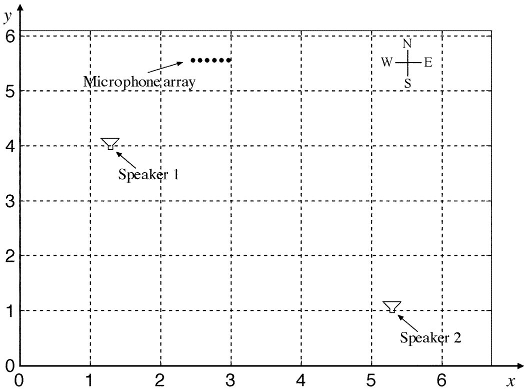


Fig. 3. Layout of the microphone array and source positions in the Varechoic chamber (coordinate values measured in meters); six microphones are placed at (2.437, 5.600, 1.400), (2.537, 5.600, 1.400), (2.637, 5.600, 1.400), (2.737, 5.600, 1.400), (2.837, 5.600, 1.400), (2.937, 5.600, 1.400), respectively; two loudspeaker sources are located at (1.337, 4.162, 1.600), and (5.337, 1.162, 1.600), respectively.

A linear microphone array which consists of six omni-directional microphones was employed in the measurement and the spacing between adjacent microphones is 10 cm. The array was mounted 1.4 m above the floor and parallel to the north wall at a distance of 50 cm. The six microphone positions are denoted as M1 (2.437, 5.600, 1.400), M2 (2.537, 5.600, 1.400), M3 (2.637, 5.600, 1.400), M4 (2.737, 5.600, 1.400), M5 (2.837, 5.600, 1.400), and M6 (2.937, 5.600, 1.400), respectively. The sources were simulated by placing two loudspeakers: one at (1.337, 4.162, 1.600), and the other at (5.337, 1.162, 1.600). The transfer functions of the acoustic channels between two loudspeakers and six microphones were measured at a 48 kHz sampling rate. Then the obtained channel impulse responses were downsampled to a 16 kHz sampling rate and truncated to 4096 samples. These measured impulse responses will be treated as the actual impulse responses in the TDE experiments.

B. Performance Measure

To better evaluate the performance of a time-delay estimator, it would be helpful to classify an estimate into two comprehensive categories: the class of success and the class of failure [11], [12]. An estimate $\hat{\tau}_i$ for which the absolute error $|\hat{\tau}_i - \tau_i|$ exceeds $\mathrm{T_c}/2$, where $\mathrm{T_c}$ is the signal correlation time, and τ_i the true delay, is identified as a failure or an anomaly which follows the terminology used in [12]. Otherwise, an estimate would be deemed as a success or a nonanomalous one. In this paper, $\mathrm{T_c}$ is

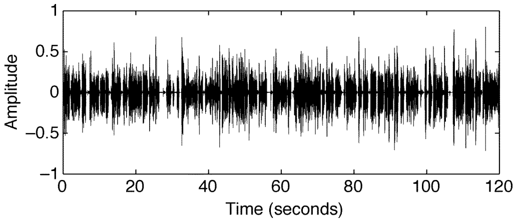


Fig. 4. Segment of the speech signal used as the source and sampled at 16 kHz.

defined as the width of the main lobe of the source signal autocorrelation function (taken between the -3-dB points). For the particular speech signal used here, which is sampled at 16 kHz, $\mathbf{T}_c$ is equal to 4.0 samples (0.25 ms).

After time-delay estimates are classified into the two classes, the TDE performance is evaluated in terms of the percentage of anomalies over the total estimates, and the mean square error (MSE) of the nonanomalous estimates.

C. Experimental Results

Several experiments were conducted to study the characteristics of the proposed multichannel TDE algorithm, as well as how the TDE performance is affected by the number of microphone sensors. For brevity, we report two sets of experimental results here.

We first consider a situation where there is only one source located in the far field (loudspeaker 1 in Fig. 3). The source signal is a speech (from a female speaker) sampled at 16 kHz and a duration of 4 min. A segment of the signal is shown in Fig. 4. The six-channel observation signals are obtained by convolving the speech source with the corresponding measured channel impulse responses and adding a zero-mean, white, Gaussian noise to each one of these outputs for a given signal-to-noise ratio (SNR).

Two experimental conditions are considered. One consists of light reverberation whose 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 measured by the Schroeder's method [26], is approximately 240 ms. The other pertains to a heavily reverberant environment where $\mathrm{T}_{60} = 580$ ms. In both cases, $\mathrm{SNR} = -5$ dB. The multichannel signals are partitioned into nonoverlapping frames with a frame size of 128 milliseconds (equivalently 2048 samples). For each frame, a delay estimate is measured according to the estimator given by (58) with a forgetting factor $\lambda = 0.95$. Therefore, with a four-minute speech sequence, a total of 1875 time-delay estimates are yielded, based on which the statistics of the performance is computed.

Fig. 5 presents the estimation results using two and six microphones respectively in a condition where $\mathbf{T}_{60} = 580$ ms, and $\mathrm{SNR} = -5$ dB. The x-axis shows the location in time at which a delay estimate is made, using the signals over a short window. The true delay in this case is $\tau = -3$ (samples). Apparently, the estimation accuracy with six microphones is much higher than that using two microphones. Using the performance measure described previously, we found that the percentages of anomalies in both conditions are rather small (approximately zero). The MSE of the nonanomalous estimates, as a function of the number of microphone sensors, is graphically portrayed in Fig. 6. As seen from Fig. 6(a), the estimator yields reasonably good performance in the light reverberation condition. The MSE is approximately -12 dB when only two sensors are used (in this case, the estimator is equivalent to the classical cross-correlation method, one member of the GCC family). It is reduced to -25 dB when one more microphone is added, and diminishes when more than four sensors are available. This demonstrates

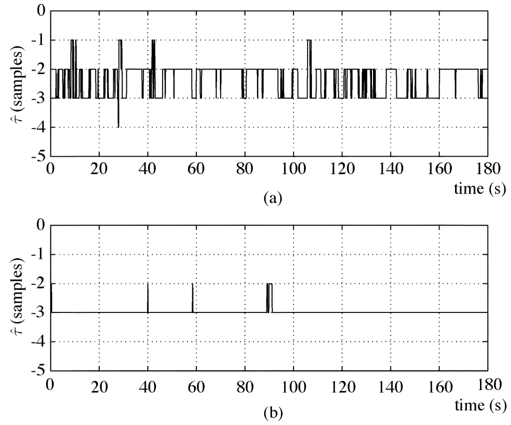


Fig. 5. TDE results in a condition where $\mathrm{T}_{60} = 580$ ms, $\mathrm{SNR} = -5$ dB, and the true delay $\tau = -3$ (samples): (a) with two microphones (b) with six microphones. The x-axis shows the location in time at which a delay estimate is made, using the signals over a short window.

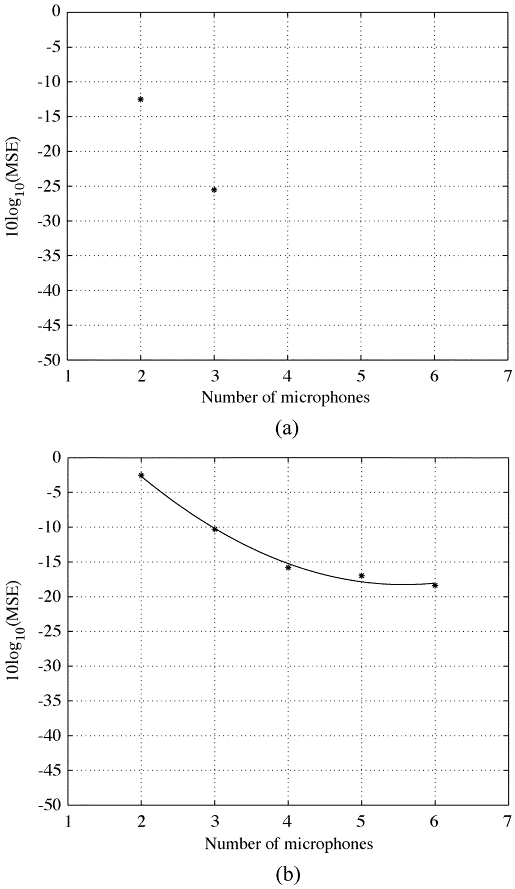


Fig. 6. MSE of the nonanomalous delay estimates in reverberant and noisy environments. (a) $\mathrm{T}_{60} = 240$ ms, and $\mathrm{SNR} = -5$ dB; (b) $\mathrm{T}_{60} = 580$ ms, and $\mathrm{SNR} = -5$ dB. The fitting curve in (b) is a second-order polynomial. (Note: in (a), when four or more microphones are used, all the time delays are correctly identified, and MSE of the nonanomalous estimates becomes zero. Thus $10\log_{10}(\mathrm{MSE})$ becomes minus infinity, which is not displayed in the figure.)

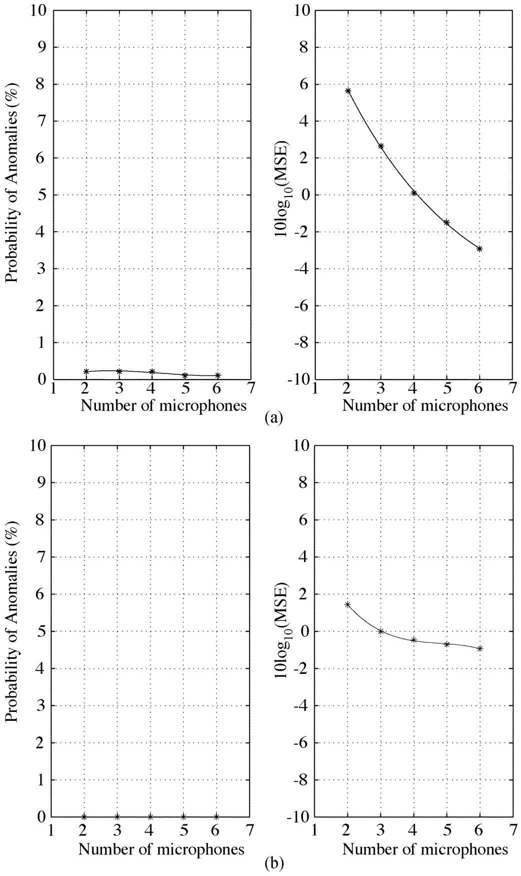


Fig. 7. Performance of time-delay estimation of two sources in a noisy but nonreverberant environment where $\mathrm{SNR} = 5\ \mathrm{dB}$. (a) Probability of anomalies and MSE of nonanomalous estimates for the first source. (b) Probability of anomalies and MSE of nonanomalous estimates for the second source.

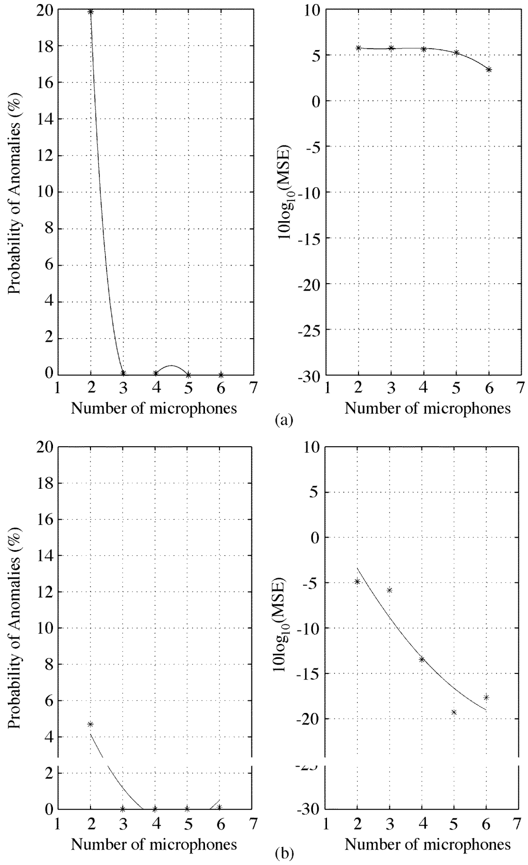


Fig. 8. Performance of time-delay estimation of two sources in a reverberant and noisy environment where $\mathrm{T}_{60} = 580\ \mathrm{ms}$ and $\mathrm{SNR} = 5\ \mathrm{dB}$. (a) Probability of anomalies and MSE of nonanomalous estimates for the first source. (b) Probability of anomalies and MSE of nonanomalous estimates for the second source.

the effectiveness of the algorithm in taking advantage of the redundant information provided by multiple microphones to mitigate the effect of noise and reverberation.

Comparing Fig. 6(a) with 6(b), one can see that the MSE of the estimator deteriorates significantly when reverberation time becomes longer. This is understandable. As reverberation becomes stronger, more reflections (some have a stronger energy level, and some have a longer delay) will reach the microphone sensors. As a result, the peak of the cost function shifts away from the true delay, which will eventually lead to performance degradation. It is remarkable that, even in the heavily reverberant environment, the TDE accuracy increases with the number of microphones, corroborating the powerfulness of the multichannel TDE approach in exploiting redundancy to combat distortion.

The second set of experiments concerns the time-delay estimation of multiple sources. As opposed to the above experiment, this time we assume that there are two sources in the far field: one is located at the position of loudspeaker 1, and the other at loudspeaker 2 in Fig. 3. The first source is a speech signal as used in the previous experiment. The second source is a music signal also sampled at 16 kHz. We further assume that the two sources have equal energy levels. Again, independent Gaussian noise is added to the multichannel signals to control the SNR.

We investigated two situations. One is free of reverberation. The other, similar to the above single-source TDE case, consists of strong reverberation where $\mathbf{T}_{60} = 580\ \mathrm{ms}$. The summary characteristics of the TDE performance are presented in Figs. 7 and 8. As clearly seen, when reverberation is absent, the relative time delays due to both sources are correctly identified, with their anomalies being less than one percent. It is also remarkable to notice that the delay estimation accuracy grows as more microphones are employed. However, we observe that the TDE performance in this case is much inferior to what obtained in the single-source scenario. This is understandable. When two sources are active at the same time, one source will act as uncorrelated noise to the other, resulting in SNR degradation and performance deterioration.

Comparing Fig. 7 with Fig. 8, one can readily see that when reverberation is present, there are more failures. For example, for the second source, the percentage of anomalies is approximately zero in the absence of reverberation. This number increases almost to 20% when $\mathrm{T}_{60} = 580\ \mathrm{ms}$. Although the general trend of performance is clear (improvement as the number of sensors increases), we observe that for the first source, the performance for the six-sensor case is slightly inferior to that of five-sensor situation. The reason may be that the sixth microphone receives some strong reflection paths. Further work is in progress to investigate what makes this phenomenon happen, and how to deal with when it occurs.

VII. Conclusions

Time-delay estimation in reverberant environments remains a difficult challenge and further research efforts are indispensable. This paper has dealt with TDE, with emphasis on combating reverberation. Starting with the theory of linear interpolation, it has introduced the concepts of multichannel correlation matrix and multichannel cross correlation coefficient. Some interesting properties and bounds of the MCCC were discussed. An efficient recursive algorithm was proposed to estimate and update the MCCC when new data snapshots are available. This new definition of the MCCC was then applied to the problem of time-delay estimation, resulting to a multichannel TDE algorithm. It was shown that this new approach is equivalent to the classical cross correlation method, one member of the GCC family, in the two-sensor case. It can be treated as a natural generalization of the cross correlation method to the multichannel case when more than two sensors are available. An appealing property of this new algorithm is that it can fully utilizes the redundant information provided by multiple sensors to enhance the TDE performance against distortion. Experiments confirmed that the delay estimation accuracy increases with the number of sensors. We also addressed time-delay estimation of multiple sources using the multichannel approach.

Appendix
Bounds of the Determinant of Matrix $\widetilde{\mathbf{R}}(n)$

Theorem: The determinant of matrix $\widetilde{\mathbf{R}}(n)$ given in (23) satisfies

$$0 < \det\left[\widetilde{\mathbf{R}}(n)\right] \leq 1. \tag{70}$$

Since $\mathbf{R}(n)$ is symmetric and is supposed to be positive definite, it is clear that $\det[\mathbf{R}(n)] > 0$, which implies that $\det[\widetilde{\mathbf{R}}(n)] > 0$. The only thing we need to prove is that $\det[\widetilde{\mathbf{R}}(n)] \leq 1$. To do so, we first give two lemmas.

Lemma 1: Suppose a matrix $\mathbf{M}$ is partitioned as

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{C} & \mathbf{B} \end{pmatrix}$$

where $\mathbf{A}$ and $\mathbf{B}$ are square matrices and $\mathbf{A}$ is nonsingular. Then

$$\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{B} - \mathbf{C}\mathbf{A}^{-1}\mathbf{D}).$$

Proof of Lemma 1: Let the square matrices $\mathbf{A}$ and $\mathbf{B}$ in the partitioned matrix $\mathbf{M}$ have dimensions $p \times p$ and $q \times q$ respectively. Then it can be verified that $\mathbf{M}$ can be factored as follows:

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{C} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{A}^{-1}\mathbf{D} \\ \mathbf{0} & \mathbf{B} - \mathbf{C}\mathbf{A}^{-1}\mathbf{D} \end{pmatrix} \tag{71}$$

where $\mathbf{I}_p$ and $\mathbf{I}_q$ are two identity matrices of size $p \times p$ and $q \times q$, respectively. Because of the special structures, the determinants of the matrices on the right hand side of (71) can be written down by inspection. The determinant of the first matrix is $\det(\mathbf{A})$, the determinant of the second matrix is 1, and the determinant of the third matrix is $\det(\mathbf{B} - \mathbf{C}\mathbf{A}^{-1}\mathbf{D})$. Since $\det(\mathbf{M})$ is the product of these three determinants, the lemma 1 follows.

Lemma 2: Let $\det(\mathbf{H})$ be the determinant of a positive definite symmetric matrix $\mathbf{H}$ of dimension $L \times L$. Then

$$\det(\mathbf{H}) \leq \prod_{i=0}^{L-1} h_{ii} \tag{72}$$

where h_{ii} are the diagonal elements of $\mathbf{H}$.

Proof of Lemma 2: The proof follows by iterative use of the **Lemma 1**. Partition the $L \times L$ matrix $\mathbf{H}$ as follows:

$$\mathbf{H} = \begin{pmatrix} \tilde{\mathbf{H}} & \mathbf{h} \\ \mathbf{h}^T & h_{LL} \end{pmatrix}$$

where $\tilde{\mathbf{H}}$ is an $(L-1) \times (L-1)$ positive definite systems matrix, $\mathbf{h}$ is an $(L-1) \times 1$ vector. Then the **Lemma 1** tells us that

$$\det(\mathbf{H}) = \det(\tilde{\mathbf{H}})(h_{LL} - \mathbf{h}^T\tilde{\mathbf{H}}^{-1}\mathbf{h}). \tag{73}$$

Since $\mathbf{H}$ and $\tilde{\mathbf{H}}$ are positive definite, their determinants are positive. Therefore the expression in the second pair of parentheses on the right hand side of (73) must be positive. Also, since the inverse of a positive matrix is positive it follows that $\mathbf{h}^T\tilde{\mathbf{H}}^{-1}\mathbf{h}$ is nonnegative. Hence $(h_{LL} - \mathbf{h}^T\tilde{\mathbf{H}}^{-1}\mathbf{h}) \leq h_{LL}$. Therefore

$$\det(\mathbf{H}) \leq h_{LL}\det(\tilde{\mathbf{H}}). \tag{74}$$

Repeated use of this bordering argument proves the lemma.

Proof of Theorem: From **Lemma 2**, we know that the determinant of the matrix $\widetilde{\mathbf{R}}(n)$ is less than or equal to the product of all its diagonal elements. From (23), we know that all the diagonal elements of $\widetilde{\mathbf{R}}(n)$ are equal to 1. We immediately have $\det[\widetilde{\mathbf{R}}(n)] \leq 1$. That completes the proof.

Acknowledgment

The authors would like to thank M. M. Sondhi for carefully reading a preliminary version of this paper, for providing many constructive comments and suggestions, and for providing the proof shown in theAppendix, which is easier to follow than the one presented in the preliminary versions of the paper.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

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

[8] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 45–50, Jan. 1997.

[9] ——, “A localization-error-based method for microphone-array design,” in *Proc. IEEE Int. Conf. Acoustic Speech, Signal Processing*, vol. 2, 1996, pp. 901–904.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] S. M. Griebel and M. S. Brandstein, “microphone array source localization using realizable delay vectors,” in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics*, 2001, pp. 71–74.

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

[17] 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.

[18] S. Kay, “Some results in linear interpolation theory,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–31, pp. 746–749, June 1983.

[19] B. Picinbono and J.-M. Kerilis, “Some properties of prediction and interpolation errors,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 525–531, Apr. 1988.

[20] M. G. Bellanger, *Adaptive Digital Filters and Signal Analysis*. New York: Marcel Dekker, 1987.

[21] J. S. Bendat and A. G. Piersol, *Random Data Analysis and Measurement Procedures*. New York: Wiley, 1986.

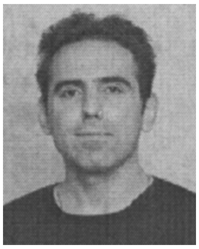
[22] H. Gish and D. Cochran, “Generalized coherence,” in *Proc. IEEE Int. Conf. Acoustic Speech, Signal Processing*, vol. 5, 1988, pp. 2745–2748.

[23] D. Cochran, H. Gish, and D. Sinno, “A geometric approach to multichannel signal detection,” *IEEE Trans. Signal Processing*, vol. 43, pp. 2049–2057, Sept. 1995.

[24] A. Härmä, “Acoustic measurement data from the varechoic chamber,” in *Proc. Technical Memorandum, Agere Systems*, Nov. 2001.

[25] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougld, “The new Varechoic chamber at AT&T Bell Labs,” in *Proc. Wallance Clement Sabine Centennial Symp.*, 1994, pp. 343–346.

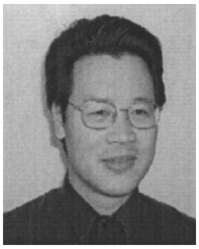
[26] M. R. Schroeder, “New method for measuring reverberation,” *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.



Jacob Benesty (M’92–SM’04) was born in 1963. He received the M.S. degree in microwaves from Pierre & Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Orsay, France, in 1991.

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 first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, QC, Canada, as an Associate Professor. His research interests are in acoustic signal processing and multimedia communications. He is a Member of the editorial board of the EURASIP *Journal on Applied Signal Processing*. He coauthored 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 *Audio Signal Processing for Next Generation Multimedia communication Systems* (Boston, MA: Kluwer Academic, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Boston, MA: Kluwer Academic, 2000).

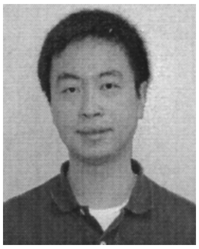
Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control.



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, Xiaan, China, in 1993 and 1995 respectively, and the Ph.D. degree in pattern recognition and intelligence control with a focus on speech recognition in noisy environments from the Chinese Academy of Sciences, Beijing, in 1998.

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, 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, Murray Hill, NJ, as 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.



Yiteng (Arden) Huang (S’97–M’01) received the B.S. degree from the Tsinghua University, Beijing, China, in 1994, the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1998 and 2001, respectively, all in electrical and computer engineering.

During his doctoral studies, from 1998 to 2001, he was a Research Assistant with the Center of Signal and Image Processing, and 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 Lucent, Murray Hill, NJ, and engaged in research on passive acoustic source localization with microphone arrays. Upon graduation, he joined Bell Laboratories as Member of Technical Staff in March 2001. He is a co-editor/coauthor of the book *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003). 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 Associate Editor for IEEE SIGNAL PROCESSING LETTERS. He received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000 to 2001 Outstanding Graduate Teaching 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 to 1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech.