Time Difference of Arrival Estimation Exploiting Multichannel Spatio-Temporal Prediction

Hongsen He, Lifu Wu, Jing Lu, Xiaojun Qiu, and Jingdong Chen, *Senior Member, IEEE*

***Abstract*—To localize sound sources in room acoustic environments, time differences of arrival (TDOA) between two or more microphone signals must be determined. This problem is often referred to as time delay estimation (TDE). The multichannel cross-correlation-coefficient (MCCC) algorithm, which is an extension of the traditional cross-correlation method from two- to multiple-channel cases, exploits spatial information among multiple microphones to improve the robustness of TDE. In this paper, we propose a multichannel spatio-temporal prediction (MCSTP) algorithm, which can be viewed as a generalization of the MCCC principle from using only spatial information to using both spatial and temporal information. A recursive version of this new algorithm is then developed, which can achieve similar performance as MCSTP, but is computationally more efficient. Experimental results in reverberant and noisy environments demonstrate the advantages of this new method for TDE.**

***Index Terms*—Microphone arrays, multichannel recursive prediction, multichannel spatio-temporal prediction (MCSTP), pre-whitening, spatial prediction, spatio-temporal prediction, time delay estimation (TDE).**

I. Introduction

TIME delay estimation (TDE), which aims at estimating the relative time difference of arrival (TDOA) using the signals received at an array of sensors, plays an important role in radar, sonar, seismology, and voice communications for localizing and tracking radiating sources. This paper focuses on the problem of TDE in room acoustic environments using microphone arrays, which is a critical problem for teleconferencing applications. Commonly used approaches to this problem include the generalized cross-correlation (GCC) method [1], [2], the blind channel identification based techniques [3]–[6], the information theory based algorithms [7], and the methods exploiting some unique characteristics of speech signals [8]. Due to its simplicity and ease of implementation, GCC [1], [2] is popularly used in the existing systems. However, the GCC algorithm is sensitive to reverberation and tends to deteriorate or even break down when reverberation is strong.

In order to improve the robustness of TDE with respect to noise and reverberation, the so-called multichannel cross-correlation-coefficient (MCCC) method was developed [9], [10]. Such algorithm exploits the redundancy among multiple microphones to deal with background noise and reverberation, thereby enhancing TDE between two sensors (typically the reference sensor and the sensor next to the reference). The robustness of MCCC with respect to noise is greatly improved as compared to the traditional cross-correlation method that uses only two sensors as demonstrated in [9]. However, the MCCC algorithm is still sensitive to reverberation. One way to make MCCC more immune to reverberation is to pre-whiten the microphone signals [11] before computing MCCC. This improved version of MCCC, now using both spatial and temporal information, can be viewed as a generalization of the phase transform (PHAT) method from two- to multiple-channel cases. But, this way of using spatial and temporal information may not be optimal as will become clear later on.

In this paper, we propose a new multichannel spatio-temporal prediction (MCSTP) algorithm for TDE, which naturally exploits spatio-temporal information in an optimal way in the minimum-mean-square-error (MMSE) sense. We also develop a recursive version of MCSTP, which can achieve similar performance as MCSTP, but is computationally more efficient. Experiments demonstrate the advantages of the proposed algorithm for TDE in reverberant and noisy environments.

II. Time Delay Estimation by Exploiting MCSTP

A. Signal Model

Assume that there is a broadband sound source in the far field which radiates a plane wave, and we use an array of M microphones to collect the signals as shown in Fig. 1. If we choose the first microphone as the reference point, the signal captured by the mth microphone at time n is then written as

$$x_m(n) = \alpha_m s[n - t - f_m(D)] + w_m(n), \quad m = 1, 2, \ldots, M, \tag{1}$$

where α_m, $m = 1, 2, \ldots, M$, are the attenuation factors due to propagation effects, $s(n)$ is the unknown zero-mean and reasonably broadband source signal, t is the propagation time from the source to microphone 1, $w_m(n)$, is the additive noise at the mth microphone, which is assumed to be uncorrelated with both the source signal and the noise observed at other microphones, D is the TDOA (i.e., relative delay) between the first and second microphones due to the source, and $f_m(D)$ is the relative delay

Manuscript received September 10, 2012; accepted September 24, 2012. Date of publication October 09, 2012; date of current version December 31, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. James Johnston.

H. He is with the Key Laboratory of Modern Acoustics and Institute of Acoustics, Nanjing University, Nanjing 210093, China, and also with the School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China (e-mail: hshe@nju.edu.cn).

L. Wu, J. Lu, and X. Qiu are with the Key Laboratory of Modern Acoustics and Institute of Acoustics, Nanjing University, Nanjing 210093, China (e-mail: wulifu@ustc.edu; lujing@nju.edu.cn; xjqiu@nju.edu.cn).

J. Chen is with Northwestern Polytechnical University, Xi'an 710072, China (e-mail: jingdongchen@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2223674

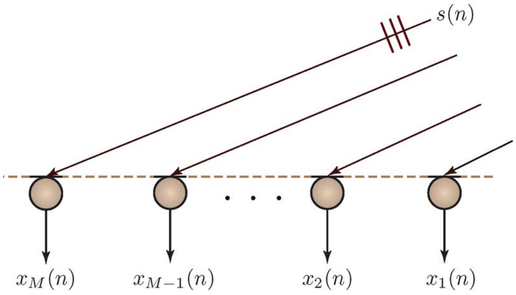


Fig. 1. An equispaced linear array of M microphones.

between microphones 1 and m. The function f_m depends not only on D but also on the microphone array geometry. In this paper, we use an equispaced linear array. Therefore, we have $f_m(D) = (m-1)D$ under the far-field assumption. With the above signal model, the objective of TDE is to estimate the time delay D given the signals received at M microphones.

For a hypothesized time delay p, we use the time shifted signal $x_m[n + f_m(p)]$ (when $p = D$, it can be checked that the desired signal components received at different microphones are aligned). To simplify the notation, let us write $x_m[n + f_m(p)]$ as $x_m(n, p)$ and define

$$\mathbf{x}(n,p) \triangleq [\, x_1(n,p) \quad x_2(n,p) \quad \cdots \quad x_M(n,p)\,]^T, \tag{2}$$

where $(\cdot)^T$ denotes the transpose of a vector or matrix.

B. Time Delay Estimation by Exploiting Spatial and Temporal Forward Prediction

First, let us consider to predict $\mathbf{x}(n,p)$ using the most recent L vectors $\mathbf{x}(n-1,p), \mathbf{x}(n-2,p), \ldots, \mathbf{x}(n-L,p)$, i.e.,

$$\widehat{\mathbf{x}}(n,p) = \mathbf{A}_{L,1}(p)\mathbf{x}(n-1,p) + \mathbf{A}_{L,2}(p)\mathbf{x}(n-2,p) + \cdots + \mathbf{A}_{L,L}(p)\mathbf{x}(n-L,p), \tag{3}$$

where $\mathbf{A}_{L,l}(p)$, $l = 1, 2, \ldots, L$, are the coefficient matrices of the multichannel forward predictor, and L is the prediction order. The prediction error vector can then be written as

$$\begin{aligned}\mathbf{e}_{\mathrm{f},L}(n,p) &\triangleq \mathbf{x}(n,p) - \widehat{\mathbf{x}}(n,p)\\ &= \mathbf{x}(n,p) - \sum_{l=1}^{L} \mathbf{A}_{L,l}(p)\mathbf{x}(n-l,p)\\ &= [\,\mathbf{I} \quad -\mathbf{A}_{L,1}(p) \quad \cdots \quad -\mathbf{A}_{L,L}(p)\,]\begin{bmatrix}\mathbf{x}(n,p)\\ \mathbf{x}(n-1,p)\\ \vdots\\ \mathbf{x}(n-L,p)\end{bmatrix}\\ &= \mathbf{A}^T(p)\mathbf{y}_{L+1}(n,p),\end{aligned} \tag{4}$$

where

$$\mathbf{e}_{\mathrm{f},L}(n,p) \triangleq [\, e_{\mathrm{f},L,1}(n,p) \quad e_{\mathrm{f},L,2}(n,p) \quad \cdots \quad e_{\mathrm{f},L,M}(n,p)\,]^T, \tag{5}$$

$$\mathbf{A}(p) = [\,\mathbf{I} \quad -\mathbf{A}_{L,1}(p) \quad \cdots \quad -\mathbf{A}_{L,L}(p)\,]^T \tag{6}$$

is the coefficient matrix [of size $M(L+1) \times M$] of the multichannel forward prediction-error filter, $\mathbf{I}$ denotes the identity matrix of size $M \times M$, and

$$\mathbf{y}_{L+1}(n,p) = [\,\mathbf{x}^T(n,p) \quad \mathbf{x}^T(n-1,p) \quad \cdots \quad \mathbf{x}^T(n-L,p)\,]^T \tag{7}$$

is the time-shifted signal vector received at M microphones. It is easy to see that the coefficient matrix $\mathbf{A}(p)$ should satisfy the following constraint:

$$\mathbf{A}^T(p)\mathbf{U}_{\mathrm{f}} = \mathbf{I}, \tag{8}$$

where

$$\mathbf{U}_{\mathrm{f}} = [\,\mathbf{I} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}\,]^T, \tag{9}$$

and $\mathbf{0}$ is a null matrix of size $M \times M$.

Now we can define the mean-square error (MSE) of the multichannel forward prediction as

$$\begin{aligned}J_{\mathrm{f}}[\mathbf{A}(p)] &\triangleq \mathrm{tr}\left\{E\left[\mathbf{e}_{\mathrm{f},L}(n,p)\mathbf{e}^T_{\mathrm{f},L}(n,p)\right]\right\}\\ &= \mathrm{tr}\left\{E\left[\mathbf{A}^T(p)\mathbf{y}_{L+1}(n,p)\mathbf{y}^T_{L+1}(n,p)\mathbf{A}(p)\right]\right\}\\ &= \mathrm{tr}\left[\mathbf{A}^T(p)\mathbf{R}_{L+1}(p)\mathbf{A}(p)\right],\end{aligned} \tag{10}$$

where $E(\cdot)$ denotes the mathematical expectation, $\mathrm{tr}(\cdot)$ stands for the trace of a matrix,

$$\begin{aligned}\mathbf{R}_{L+1}(p) &= E\left[\mathbf{y}_{L+1}(n,p)\mathbf{y}^T_{L+1}(n,p)\right]\\ &= \begin{bmatrix}\mathbf{R}(0,p) & \mathbf{R}(1,p) & \cdots & \mathbf{R}(L,p)\\ \mathbf{R}^T(1,p) & \mathbf{R}(0,p) & \cdots & \mathbf{R}(L-1,p)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{R}^T(L,p) & \mathbf{R}^T(L-1,p) & \cdots & \mathbf{R}(0,p)\end{bmatrix}\end{aligned} \tag{11}$$

is the spatio-temporal correlation matrix of size $M(L+1) \times M(L+1)$, and

$$\mathbf{R}(l,p) = E\left[\mathbf{x}(n,p)\mathbf{x}^T(n-l,p)\right], \quad l = 0, 1, \ldots, L. \tag{12}$$

In order to estimate $\mathbf{A}(p)$, let us rewrite the constraint given in (8) into the following form:

$$\mathbf{A}^T(p)\mathbf{U}_{\mathrm{f}}\boldsymbol{\eta}_m = \boldsymbol{\eta}_m, \quad m = 1, 2, \ldots, M, \tag{13}$$

where

$$\boldsymbol{\eta}_m = \Big[\,\underbrace{0 \quad \cdots \quad 0}_{m-1} \quad 1 \quad \underbrace{0 \quad \cdots \quad 0}_{M-m}\,\Big]^T \tag{14}$$

is a unit vector. Using a set of Lagrange multipliers to adjoin the constraints (13) to the cost function (10), we get

$$\mathscr{L}[\mathbf{A}(p), \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M] = \mathrm{tr}\left[\mathbf{A}^T(p)\mathbf{R}_{L+1}(p)\mathbf{A}(p)\right] + \sum_{m=1}^{M} \boldsymbol{\mu}^T_m\left[\mathbf{A}^T(p)\mathbf{U}_{\mathrm{f}}\boldsymbol{\eta}_m - \boldsymbol{\eta}_m\right], \tag{15}$$

where vectors $\boldsymbol{\mu}_m$, $m = 1, 2, \ldots, M$, are the Lagrange multipliers. Taking the gradient of $\mathscr{L}$ with respect to $\mathbf{A}(p)$ and equating the result to zero, we obtain the optimal coefficient matrix for the multichannel forward prediction:

$$\mathbf{A}_{\mathrm{o}}(p) = \mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\mathrm{f}} \left[\mathbf{U}_{\mathrm{f}}^T \mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\mathrm{f}}\right]^{-1}, \tag{16}$$

where we have assumed that the matrix $\mathbf{R}_{L+1}(p)$ is of full rank[1].

Substituting the optimal prediction matrix $\mathbf{A}_{\mathrm{o}}(p)$ into (4), we obtain the optimal prediction error signal vector $\mathbf{e}_{\mathrm{o,f},L}(n,p)$. The cross-correlation matrix of the prediction error signals is then

$$\mathbf{E}_{\mathrm{o,f},L}(p) \triangleq E[\mathbf{e}_{\mathrm{o,f},L}(n,p)\mathbf{e}_{\mathrm{o,f},L}^T(n,p)] = [\mathbf{U}_{\mathrm{f}}^T \mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\mathrm{f}}]^{-1}. \tag{17}$$

This matrix can be factorized as [9], [10]:

$$\mathbf{E}_{\mathrm{o,f},L}(p) = \boldsymbol{\Sigma}(p)\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p)\boldsymbol{\Sigma}(p), \tag{18}$$

where

$$\boldsymbol{\Sigma}(p) = \begin{bmatrix} \sqrt{e_{\mathrm{o,f},L,1,1}(p)} & 0 & \cdots & 0 \\ 0 & \sqrt{e_{\mathrm{o,f},L,2,2}(p)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{e_{\mathrm{o,f},L,M,M}(p)} \end{bmatrix} \tag{19}$$

is a diagonal matrix, $e_{\mathrm{o,f},L,m,m}(p), m = 1, 2, \ldots, M$, is the mth diagonal element of the matrix $\mathbf{E}_{\mathrm{o,f},L}(p)$, which corresponds to the variance of the prediction error signal for the mth channel,

$$\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p) = \begin{bmatrix} 1 & \rho_{12}(p) & \cdots & \rho_{1M}(p) \\ \rho_{21}(p) & 1 & \cdots & \rho_{2M}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M1}(p) & \rho_{M2}(p) & \cdots & 1 \end{bmatrix} \tag{20}$$

is a symmetric and generally positive semi-definite matrix ,

$$\rho_{ij}(p) = \frac{e_{\mathrm{o,f},L,i,j}(p)}{\sqrt{e_{\mathrm{o,f},L,i,i}(p)e_{\mathrm{o,f},L,j,j}(p)}}, \quad i,j = 1, 2, \ldots, M, \tag{21}$$

is the correlation coefficient between the aligned prediction error signals at the ith and jth microphones, $e_{\mathrm{o,f},L,i,j}(p)$, $i,j = 1, 2, \ldots, M$, is the (i,j)th element of the matrix $\mathbf{E}_{\mathrm{o,f},L}(p)$.

Since the matrix $\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p)$ is symmetric and positive semi-definite, and its diagonal elements are all equal to one, it can be shown that [9], [10]

$$0 \leq \det\left[\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p)\right] \leq 1, \tag{22}$$

where $\det(\cdot)$ stands for the determinant of a square matrix.

A natural way of using the multichannel cross-correlation matrix in TDOA estimation is through the so-called MCCC [9], [10], which measures the correlation among the prediction error signals. Given the normalized MCSTP error correlation matrix $\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p)$, we can now define the squared MCCC among the M aligned prediction error signals $e_{\mathrm{f},L,m}(n,p)$, $m = 1, 2, \ldots, M$, following the MCCC definition given in [9], i.e.,

$$\rho_{\mathrm{MCSTP}}^2(p) = 1 - \det\left[\widetilde{\mathbf{E}}_{\mathrm{o,f},L}(p)\right] = 1 - \frac{\det[\mathbf{E}_{\mathrm{o,f},L}(p)]}{\prod_{m=1}^M e_{\mathrm{o,f},L,m,m}(p)}. \tag{23}$$

Basically, the coefficient $\rho_{\mathrm{MCSTP}}^2(p)$ measures the amount of correlation among the MCSTP error signals of all the M channels. This coefficient has the following properties: 1) $0 \leq \rho_{\mathrm{MCSTP}}^2(p) \leq 1$; 2) if two or more prediction error signals are perfectly correlated, $\rho_{\mathrm{MCSTP}}^2(p) = 1$; 3) if all the prediction error signals are completely uncorrelated with each other, $\rho_{\mathrm{MCSTP}}^2(p) = 0$; 4) if one of the prediction error signals is completely uncorrelated with all the other prediction error signals, the $\rho_{\mathrm{MCSTP}}^2(p)$ will measure the correlation among those $M-1$ remaining prediction error signals.

Given $\rho_{\mathrm{MCSTP}}^2(p)$, the TDOA estimate can be obtained as

$$\widehat{D} = \arg\max_{p} \ \rho_{\mathrm{MCSTP}}^2(p), \tag{24}$$

where $\widehat{D}$ is an estimate of D, $p \in [-p_{\max}, p_{\max}]$, and $p_{\max}$ is the maximum possible delay. Note that this method can be extended to TDE of multiple sources by searching for multiple peaks. In this paper, however, we focus only on TDE of a single source.

We should point out that the estimator given in (24) is fundamentally different from that given in [9] though both use the concept of MCCC. Specifically, the estimator in (24) uses the MCSTP error signals to construct MCCC, while the estimator in [9] forms MCCC directly using the microphone signals. Since there may exist self correlation among microphone signals while there is no self correlation in the MCSTP error signals, the estimator in (24) is expected to have better performance than that in [9], which will be demonstrated in the following sections. Note that the variance of the prediction error signal for each channel is a function of the parameter p, and therefore the denominator of the second term in (23) is very important; however, it is negligible for the MCCC algorithm in [9] since the variances of the microphone signals do not change with p.

C. Analysis of the MCSTP Algorithm

In order to analyze the performance of MCSTP, we consider to use a noise-free, reverberation signal model in this subsection. The signal received at the mth microphone at time n is modeled as

$$x_m(n) = h_m * s(n), \quad m = 1, 2, \ldots, M, \tag{25}$$

[1] In practical applications, there are always noise and reverberation. So, the signals from different microphones are not fully correlated and the matrix $\mathbf{R}_{L+1}(p)$ is generally of full rank. However, in the ideal case where there is no noise and reverberation, one microphone signal can be completely predicted by the signal from another microphone. In this situation, the matrix $\mathbf{R}_{L+1}(p)$ may be rank deficient if more than two microphones are used. If this unrealistic situation is a concern, one can circumvent the issue by finding the optimal prediction matrix through minimizing $\mathrm{tr}\,[\mathbf{A}^T(p)\mathbf{R}_{L+1}(p)\mathbf{A}(p)] + \lambda \cdot \mathrm{tr}\,[\mathbf{A}(p)\mathbf{A}^T(p)]$, where $\lambda \in (0,1]$ is a weighting factor.

where h_m is the impulse response from the unknown source $s(n)$ to the mth microphone. The signals given in (25) can be written into the following vector/matrix form (considering the most recent L signal samples)[2]

$$\mathbf{x}(n) = \mathbf{H}^T \mathbf{s}(n), \tag{27}$$

where

$$\mathbf{x}(n) = \left[\mathbf{x}_1^T(n) \quad \mathbf{x}_2^T(n) \quad \cdots \quad \mathbf{x}_M^T(n) \right]^T, \tag{28}$$

$$\mathbf{x}_m(n) = \left[x_m(n) \quad x_m(n-1) \quad \cdots \quad x_m(n-L+1) \right]^T, \quad m = 1, 2, \ldots, M, \tag{29}$$

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n), \tag{30}$$

$$\mathbf{s}(n) = \left[s(n) \quad s(n-1) \quad \cdots \quad s(n-N+1) \right]^T, \tag{31}$$

$$\mathbf{h}_m = \left[h_m(0) \quad h_m(1) \quad \cdots \quad h_m(L_h-1) \quad 0 \quad \cdots \quad 0 \right]^T, \tag{32}$$

$$\mathbf{H} = \left[\mathbf{H}_1 \quad \mathbf{H}_2 \quad \cdots \quad \mathbf{H}_M \right], \tag{33}$$

$$\mathbf{H}_m = \begin{bmatrix} h_m(0) & 0 & \cdots & 0 \\ h_m(1) & h_m(0) & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ h_m(L_h-1) & \cdots & \cdots & 0 \\ 0 & h_m(L_h-1) & \cdots & h_m(0) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & h_m(L_h-1) \end{bmatrix} \tag{34}$$

is a Sylvester matrix of size $N \times L$, $N = L + L_h - 1$, and L_h is the length of the longest acoustic impulse responses among the M channels h_m, $m = 1, 2, \ldots, M$. Note that $L_h > L$ in general, and the matrix $\mathbf{H}_m$ shows how the microphone signals are generated with the multichannel reverberation model [12].

If we use the most recent L samples captured by each microphone to predict $x_m(n)$ in a forward manner, the prediction error is

$$e_m(n) \triangleq x_m(n) - \mathbf{a}_m^T \mathbf{x}(n-1) = \mathbf{h}_m^T \mathbf{s}(n) - \mathbf{a}_m^T \mathbf{x}(n-1), \tag{35}$$

where

$$\mathbf{a}_m = \left[a_m(0) \quad a_m(1) \quad \cdots \quad a_m(ML-1) \right]^T. \tag{36}$$

The prediction errors of M channels can be combined into a vector, i.e.,

$$\mathbf{e}(n) = \begin{bmatrix} \mathbf{h}_1^T \mathbf{s}(n) - \mathbf{a}_1^T \mathbf{x}(n-1) \\ \mathbf{h}_2^T \mathbf{s}(n) - \mathbf{a}_2^T \mathbf{x}(n-1) \\ \vdots \\ \mathbf{h}_M^T \mathbf{s}(n) - \mathbf{a}_M^T \mathbf{x}(n-1) \end{bmatrix} = \mathbf{P}^T \mathbf{s}(n) - \mathbf{A}^T \mathbf{x}(n-1) = \mathbf{P}^T \mathbf{s}(n) - \mathbf{A}^T \mathbf{H}^T \mathbf{s}(n-1), \tag{37}$$

where

$$\mathbf{A} = \left[\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_M \right], \tag{38}$$

and

$$\mathbf{P} = \left[\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_M \right]. \tag{39}$$

Then, the MSE of the multichannel forward prediction is given by

$$\begin{aligned} J_{\mathrm{f}}(\mathbf{A}) &\triangleq E\left[\mathbf{e}^T(n) \mathbf{e}(n) \right] \\ &= E\big[\mathbf{s}^T(n) \mathbf{P} \mathbf{P}^T \mathbf{s}(n) - 2\mathbf{s}^T(n-1) \mathbf{H} \mathbf{A} \mathbf{P}^T \mathbf{s}(n) \\ &\quad + \mathbf{s}^T(n-1) \mathbf{H} \mathbf{A} \mathbf{A}^T \mathbf{H}^T \mathbf{s}(n-1) \big]. \end{aligned} \tag{40}$$

Taking the gradient of $J_{\mathrm{f}}(\mathbf{A})$ with respect to the coefficient matrix $\mathbf{A}$ and equating the result to zero, we obtain the optimal coefficient matrix

$$\mathbf{A}_{\mathrm{o}} = \{ \mathbf{H}^T E[\mathbf{s}(n-1) \mathbf{s}^T(n-1)] \mathbf{H} \}^{\dagger} \, \mathbf{H}^T E[\mathbf{s}(n-1) \mathbf{s}^T(n)] \mathbf{P}, \tag{41}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo inverse. Let us introduce a new matrix $\mathbf{Q}$, which is basically the right-hand side of (41), but replacing $\mathbf{P}$ with $\mathbf{H}$, i.e.,

$$\mathbf{Q} \triangleq \left\{ \mathbf{H}^T E\left[\mathbf{s}(n-1) \mathbf{s}^T(n-1) \right] \mathbf{H} \right\}^{\dagger} \mathbf{H}^T E[\mathbf{s}(n-1) \mathbf{s}^T(n)] \mathbf{H}. \tag{42}$$

Notice that the mth ($m = 1, 2, \ldots, M$) column of the matrix $\mathbf{P}$ corresponds to column $(m-1)L+1$ of the matrix $\mathbf{H}$, and therefore the mth column of the matrix $\mathbf{A}_{\mathrm{o}}$ corresponds to column $(m-1)L+1$ of the matrix $\mathbf{Q}$.

We assume that a speech signal can be modeled as an autoregressive (AR) process excited by white noise [13], and the Z-transform of the AR process is $1/a(z)$, where (assuming that the order of the AR process is N for simplicity)

$$a(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_N z^{-N}. \tag{43}$$

Then, the source signal vector can be expressed as

$$\mathbf{s}(n) = \mathbf{C}^T \mathbf{s}(n-1) + \mathbf{v}(n), \tag{44}$$

where

$$\mathbf{C} = \begin{bmatrix} a_1 & 1 & 0 & \cdots & 0 \\ a_2 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & 1 \\ a_N & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{45}$$

is a companion matrix of size $N \times N$, and

$$\mathbf{v}(n) = \left[v(n) \quad 0 \quad \cdots \quad 0 \right]^T. \tag{46}$$

Since the source signal and noise are assumed uncorrelated, we can deduce from (44) that:

$$E\left[\mathbf{s}(n-1) \mathbf{s}^T(n) \right] = E\left[\mathbf{s}(n-1) \mathbf{s}^T(n-1) \right] \mathbf{C}. \tag{47}$$

Let us further assume that: 1) the matrix $E\left[\mathbf{s}(n-1) \mathbf{s}^T(n-1) \right]$ is positive definite, and denote $E\left[\mathbf{s}(n-1) \mathbf{s}^T(n-1) \right] = \mathbf{S}^T \mathbf{S}$,

[2] Note that the zeros shared by all the impulse responses at the beginning are removed. In order to better understand the MCCC based on the MCSTP, we consider the case that $h_1(0)$ corresponds to the direct path component of $\mathbf{h}_1$ according to Fig. 1, and

$$\mathbf{h}_m = \Big[\underbrace{\delta \quad \delta \quad \cdots \quad \delta}_{f_m(D)} \quad \underbrace{h_m(0) \quad \cdots \quad h_m(L_h-1) \quad 0 \quad \cdots \quad 0}_{N - f_m(D)} \Big]^T, \quad m = 2, \ldots, M, \tag{26}$$

where δ is a very small positive number.

where $\mathbf{S}$ is an invertible matrix; 2) room transfer functions from the source to multiple microphones do not share common zeros, which makes the matrix $\mathbf{H}$ be full row rank. Then, the matrix $\mathbf{Q}$ in (42) can be simplified as

$$\mathbf{Q} = \left(\mathbf{H}^T\mathbf{S}^T\mathbf{S}\mathbf{H}\right)^{\dagger}\mathbf{H}^T\mathbf{S}^T\mathbf{S}\mathbf{C}\mathbf{H} = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{C}\mathbf{H}. \tag{48}$$

Given $\mathbf{Q}$, we can now get the mth column of the optimal coefficient matrix $\mathbf{A}_{\rm o}$ (which is the $[(m-1)L+1]$th column of the matrix $\mathbf{Q}$),

$$\mathbf{a}_{{\rm o},m} = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{C}\mathbf{h}_m. \tag{49}$$

Finally, the prediction error of the mth channel is obtained as

$$\begin{aligned} e_{{\rm o},m}(n) &= \mathbf{h}_m^T\mathbf{s}(n) - \mathbf{a}_{{\rm o},m}^T\mathbf{H}^T\mathbf{s}(n-1) \\ &= \mathbf{h}_m^T\left[\mathbf{s}(n) - \mathbf{C}^T\mathbf{s}(n-1)\right] \\ &= \mathbf{h}_m^T\mathbf{v}(n) = h_m(0)v(n). \end{aligned} \tag{50}$$

It can be seen from (50) that the prediction error is a whitened version of the reverberant signal captured by a microphone; so the condition number of the correlation matrix $\mathbf{E}_{{\rm o,f},L}(p)$ in (17) is much smaller than that of the spatial correlation matrix corresponding to the MCCC algorithm. It can also be found from (50) that the reverberant components of the microphone signal are eliminated, which indicates that this pre-whitening is optimal in terms of robustness to reverberation. Therefore, the robustness of the proposed MCSTP algorithm to reverberation can be improved as compared to the MCCC algorithm with or without pre-whitening. Notice that for the MCSTP-based TDOA estimator in Section II-B, when $p = D$, the signals of all the channels are aligned, indicating that all the $h_m(0), m = 1, 2, \ldots, M$, are now direct path components for the corresponding channels.

D. Time Delay Estimation by Exploiting Spatial and Temporal Backward Prediction

In backward prediction, the $\mathbf{x}(n-L,p)$ vector is predicted using $\mathbf{x}(n,p), \mathbf{x}(n-1,p), \ldots, \mathbf{x}(n-L+1,p)$, i.e.,

$$\begin{aligned} \widehat{\mathbf{x}}(n-L,p) = \mathbf{B}_{L,1}(p)\mathbf{x}(n,p) + \mathbf{B}_{L,2}(p)\mathbf{x}(n-1,p) + \\ \cdots + \mathbf{B}_{L,L}(p)\mathbf{x}(n-L+1,p), \end{aligned} \tag{51}$$

where $\mathbf{B}_{L,l}(p)$, $l = 1, 2, \ldots, L$, are the coefficient matrices of the multichannel backward predictor. The error signal vector of the multichannel backward prediction is written as:

$$\begin{aligned} &\mathbf{e}_{{\rm b},L}(n,p) \\ &\triangleq \mathbf{x}(n-L,p) - \widehat{\mathbf{x}}(n-L,p) \\ &= \mathbf{x}(n-L,p) - \sum_{l=1}^{L}\mathbf{B}_{L,l}(p)\mathbf{x}(n-l+1,p) \\ &= \left[-\mathbf{B}_{L,1}(p) \quad \cdots \quad -\mathbf{B}_{L,L}(p) \quad \mathbf{I}\right]\begin{bmatrix} \mathbf{x}(n,p) \\ \vdots \\ \mathbf{x}(n-L+1,p) \\ \mathbf{x}(n-L,p) \end{bmatrix} \\ &= \mathbf{B}^T(p)\mathbf{y}_{L+1}(n,p), \end{aligned} \tag{52}$$

where

$$\mathbf{B}(p) = \left[-\mathbf{B}_{L,1}(p) \quad \cdots \quad -\mathbf{B}_{L,L}(p) \quad \mathbf{I}\right]^T \tag{53}$$

is the coefficient matrix [of size $M(L+1) \times M$] of the multichannel backward prediction-error filter. It is obvious that the $\mathbf{B}(p)$ matrix should satisfy

$$\mathbf{B}^T(p)\mathbf{U}_{\rm b} = \mathbf{I}, \tag{54}$$

where

$$\mathbf{U}_{\rm b} = \left[\mathbf{0} \quad \cdots \quad \mathbf{0} \quad \mathbf{I}\right]^T. \tag{55}$$

Following the same line of principles in Section II-B, we can deduce the optimal coefficient matrix of the multichannel backward prediction as

$$\mathbf{B}_{\rm o}(p) = \mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\rm b}\left[\mathbf{U}_{\rm b}^T\mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\rm b}\right]^{-1}. \tag{56}$$

Substituting the optimal prediction matrix $\mathbf{B}_{\rm o}(p)$ into (52), we obtain the optimal prediction error signal vector $\mathbf{e}_{{\rm o,b},L}(n,p)$. The cross-correlation matrix of the corresponding prediction error signals is then

$$\mathbf{E}_{{\rm o,b},L}(p) \triangleq E[\mathbf{e}_{{\rm o,b},L}(n,p)\mathbf{e}_{{\rm o,b},L}^T(n,p)] = [\mathbf{U}_{\rm b}^T\mathbf{R}_{L+1}^{-1}(p)\mathbf{U}_{\rm b}]^{-1}. \tag{57}$$

Similar to the forward prediction case, we can define the squared MCCC of the MCSTP error signals based on the cross-correlation matrix $\mathbf{E}_{{\rm o,b},L}(p)$ and then estimate TDOA by searching the lag time corresponding to the maximum of the MCCC.

E. Time Delay Estimation Based on Recursive Spatio-Temporal Prediction

It is observed from (11) that the spatio-temporal correlation matrix $\mathbf{R}_{L+1}(p)$ has a high dimension; thus, finding its inverse is computationally very expensive. In order to reduce the computational complexity, we develop a recursive version of the previous MCSTP algorithm by borrowing the basic idea in [14]. The algorithm is summarized in Table I. The detailed derivations are shown in the Appendix.

Besides the complexity advantage of the recursive version, another benefit of using the recursive method is that it provides the predictor of all different orders. This can provide a way to determine the optimal order for the prediction, i.e., the optimal value is reached if the prediction error is under a threshold. This can be very useful in practice when the choice of the prediction order is not easy to determine in advance.

F. Comparison of Computational Complexity

This subsection briefly compares the computational complexity of the MCCC, MCCC with pre-whitening, MCSTP, and recursive MCSTP algorithms. The computational complexity is evaluated in terms of the number of real-valued multiplications/divisions required for implementation of each algorithm. The number of additions/subtractions is neglected because they are much quicker to compute in most generic hardware platforms. Assume that the frame length is $L_{\rm f}$, the number

TABLE I
TDE ALGORITHM BASED ON THE RECURSIVE MCSTP.

Algorithm steps:	(Real-valued) multiplications:
for $-p_{\max} \le p \le p_{\max}$	
Initialization: $\mathbf{E}_{\mathrm{o,f},0}(p) = \mathbf{E}_{\mathrm{o,b},0}(p) = \mathbf{R}(0,p)$	$M^2(L_\mathrm{f}+1)$
for $1 \le l \le L$	
if $l == 1$	
$\mathbf{K}_{\mathrm{b},1}(p) = \mathbf{R}(1,p)$	$M^2 L_\mathrm{f}$
$\mathbf{A}_{\mathrm{o},1} = \mathbf{E}^{-1}_{\mathrm{o,b},0}(p)\mathbf{K}_{\mathrm{b},1}(p)$	$2M^3 - M$
$\mathbf{B}_{\mathrm{o},1} = \mathbf{E}^{-1}_{\mathrm{o,f},0}(p)\mathbf{K}_{\mathrm{b},1}(p)$	$2M^3 - M$
$\mathbf{E}_{\mathrm{o,f},1}(p) = \mathbf{E}_{\mathrm{o,f},0}(p) - \mathbf{K}_{\mathrm{b},1}(p)\mathbf{A}_{\mathrm{o},1}(p)$	M^3
$\mathbf{E}_{\mathrm{o,b},1}(p) = \mathbf{E}_{\mathrm{o,b},0}(p) - \mathbf{K}^T_{\mathrm{b},1}(p)\mathbf{B}_{\mathrm{o},1}(p)$	M^3
else	
$\mathbf{K}_{\mathrm{b},l}(p) = \mathbf{R}(l,p) - \mathbf{R}_\mathrm{f}{}^T(1/(l-1),p)\mathbf{B}_{\mathrm{o},l-1}(p)$	$M^2 L_\mathrm{f} + M^3(l-1)$
$\mathbf{P}(p) = \mathbf{E}^{-1}_{\mathrm{o,b},l-1}(p)\mathbf{K}^T_{\mathrm{b},l}(p)$	$2M^3 - M$
$\mathbf{Q}(p) = \mathbf{E}^{-1}_{\mathrm{o,f},l-1}(p)\mathbf{K}_{\mathrm{b},l}(p)$	$2M^3 - M$
$\mathbf{A}_{\mathrm{o},l}(p) = \begin{bmatrix} \mathbf{A}_{\mathrm{o},l-1}(p) - \mathbf{B}_{\mathrm{o},l-1}(p)\mathbf{P}(p) \\ \mathbf{P}(p) \end{bmatrix}$	$M^3(l-1)$
$\mathbf{B}_{\mathrm{o},l}(p) = \begin{bmatrix} \mathbf{Q}(p) \\ \mathbf{B}_{\mathrm{o},l-1}(p) - \mathbf{A}_{\mathrm{o},l-1}(p)\mathbf{Q}(p) \end{bmatrix}$	$M^3(l-1)$
$\mathbf{E}_{\mathrm{o,f},l}(p) = \mathbf{E}_{\mathrm{o,f},l-1}(p) - \mathbf{K}_{\mathrm{b},l}(p)\mathbf{P}(p)$	M^3
$\mathbf{E}_{\mathrm{o,b},l}(p) = \mathbf{E}_{\mathrm{o,b},l-1}(p) - \mathbf{K}^T_{\mathrm{b},l}(p)\mathbf{Q}(p)$	M^3
end	
end	
$\rho^2_{\mathrm{MCSTP}}(p) = 1 - \frac{\det[\mathbf{E}_{\mathrm{o,f},L}(p)]}{\prod_{m=1}^{M} e_{\mathrm{o,f},L,m,m}(p)}$	$\frac{1}{3}M^3 + \frac{8}{3}M + 1$
end	
$\widehat{D} = \arg\max_p \rho^2_{\mathrm{MCSTP}}(p)$	Total: $(2p_{\max}+1)\{\frac{3}{2}M^3L^2 + \frac{9}{2}M^3L$ $+M^2[(L+1)L_\mathrm{f}+1] - 2ML + \frac{19}{3}M^3 + \frac{2}{3}M + 1\}$

of multiplications for computing the inverse of a matrix of size $M \times M$ (using the LU decomposition) is assumed to be $M^3 - M$ [15], and the determinant of a matrix of size $M \times M$ is computed through LU decomposition, which requires $M^3/3 + 5M/3$ multiplications [15]. Then, the number of multiplications required by the MCCC algorithm for each frame is $(2p_{\max}+1)[M^2(L_\mathrm{f}+1) + M^3/3 + 8M/3 + 1]$, and that required by the MCCC algorithm with pre-whitening for each frame is

$$M\left[\frac{5L_\mathrm{f}\log_2(L_\mathrm{f})}{2} - \frac{31L_\mathrm{f}}{4} + 13\right] + (2p_{\max}+1)\left[M^2(L_\mathrm{f}+1) + \frac{M^3}{3} + \frac{8M}{3} + 1\right].$$

One can check that the number of multiplications needed by MCSTP with direct inverse is

$$(2p_{\max}+1)\Bigg[M^3L^3 + 3M^3L^2 + 3M^3L + M^2(L+1)^2(L_\mathrm{f}+1) - ML + \frac{7M^3}{3} + \frac{2M}{3} + 1\Bigg],$$

while that required by the recursive MCSTP algorithm for each frame, as shown in Table I, is

$$(2p_{\max}+1)\Bigg\{\frac{3M^3L^2}{2} + \frac{9M^3L}{2} + M^2[(L+1)L_\mathrm{f}+1] - 2ML + \frac{19M^3}{3} + \frac{2M}{3} + 1\Bigg\}.$$

Fig. 2 plots the computational complexity of the four algorithms as a function of the prediction order when a frame is processed, where four microphones are considered (the frame length and $p_{\max}$ are shown in Section III). Clearly, the computational complexity of the recursive algorithm is significantly lower than that of MCSTP with direct inverse. It is seen that both the MCSTP and recursive MCSTP algorithms have a higher complexity than the MCCC-type methods, but their performance is much better as will be seen in Section III.

III. SIMULATION EXPERIMENTS

A. Experimental Environment

Experiments are carried out in a simulated room of size 7 m × 6 m × 3 m. An equispaced linear array consisting of six

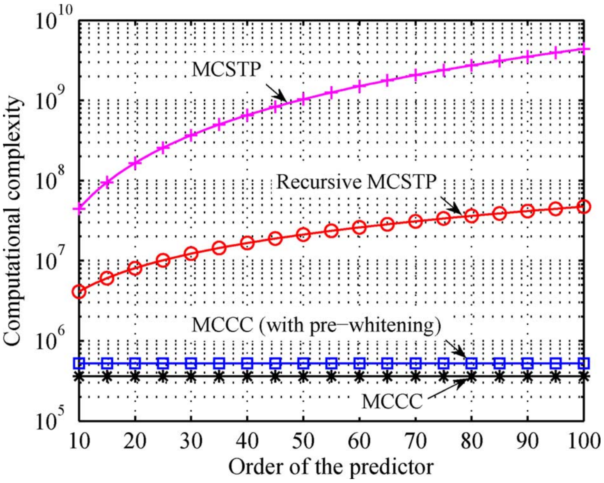


Fig. 2. Computational complexity of the MCCC, MCCC with pre-whitening, proposed MCSTP, and recursive MCSTP algorithms when a frame is processed where the frame length is 2048 samples and four microphones are considered.

omnidirectional microphones is used with the inter-element spacing being 0.1 m. For ease of exposition, positions in the room are designated by (x, y, z) coordinates with reference to the southwest corner of the room floor. The first and sixth microphones of the array are at (3.25, 3.00, 1.40) and (3.75, 3.00, 1.40), respectively. The sound source is located at (2.49, 1.27, 1.40).

The impulse responses from the source to the six microphones are generated using the image model [16]. The microphones' outputs are obtained by convolving the source signal with the corresponding generated impulse responses and then adding zero-mean white Gaussian noise to the results to control the signal-to-noise ratio (SNR).

B. Performance Criteria

In the simulations, the microphone signals are partitioned into nonoverlapping frames with a frame length of 128 ms. Each frame is windowed with a Hamming window, and a time delay estimate is then obtained. Two performance metrics [17], [18], namely the probability of anomalous estimates and the root mean square error (RMSE) of nonanomalous estimates, are used to evaluate the performance of the proposed algorithm. The following criterion is used to distinguish between an anomalous and a nonanomalous estimates. For the ith delay estimate $\widehat{D}_i$, if the absolute error $\left|\widehat{D}_i - D\right| > T_c/2$, where D is the true delay, and T_c is the signal self correlation time, the estimate is identified as an anomalous estimate. Otherwise, the estimate would be deemed as a nonanomalous one [9], [10]. For the particular source signals used in this study, such as speech and non-speech signals, which is sampled at 16 kHz, T_c is equal to 4.0 samples.

The RMSE of the nonanomalous estimates is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{na}}} \sum_{i \in S_{\mathrm{na}}} \left|\widehat{D}_i - D\right|^2}, \tag{58}$$

where N_{na} is the number of the nonanomalous estimates for TDOA, and S_{na} denotes the subset of the nonanomalous estimates.

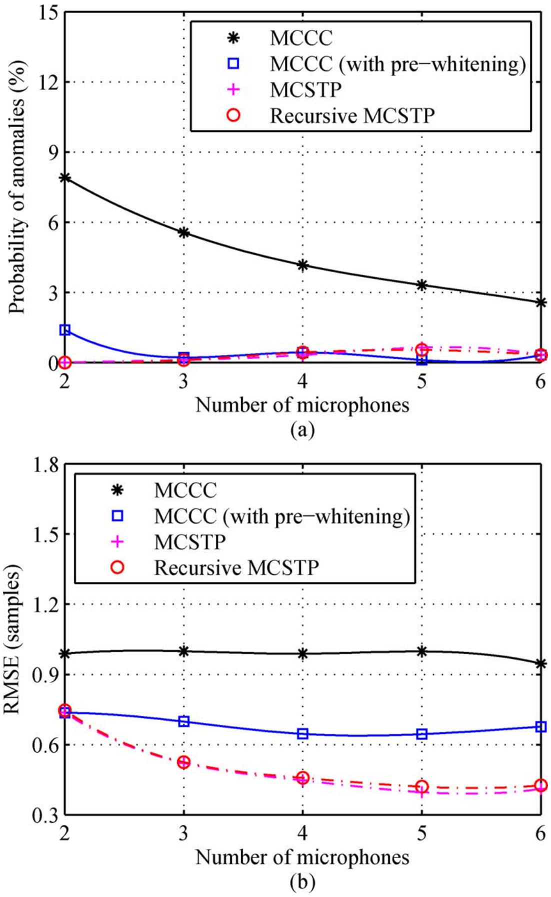


Fig. 3. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates versus the number of microphones in a moderately reverberant environment ($T_{60} = 300$ ms). The prediction order is 80.

C. Results and Discussions

First of all, we assume that the source signal is a speech signal from a female talker and the length of the signal is 2 minutes. The total number of frames is 936 (the frame length is 2048 samples). The true time delay from the sound source to the first two microphones is 2.0 samples.

The first set of experiments is to investigate the effectiveness of the proposed MCSTP algorithm in reverberant but noise-free environments. Fig. 3 shows the TDE results in a moderately reverberant environment ($T_{60} = 300$ ms), where the prediction order is set to 80. The probability of anomalous estimates and the RMSE of nonanomalous estimates are plotted as a function of the number of microphones, respectively. It is seen from Fig. 3 that the performance of all the algorithms generally increases with the number of microphones, which indicates that more spatial redundancy can help improve the robustness of TDE. The MCCC algorithm without pre-whitening is found most sensitive to reverberation among the studied algorithms. However, its performance is greatly improved when microphone signals are pre-whitened (note that the MCCC algorithm is basically a multichannel generalization of the PHAT algorithm when a pre-whitening process is used). It is also observed from Fig. 3 that the probability of anomalous estimates of the MCSTP algorithm is less than one percent for all the different conditions. For the case of two microphones, the MCSTP algorithm has a smaller probability of anomalous estimates than the

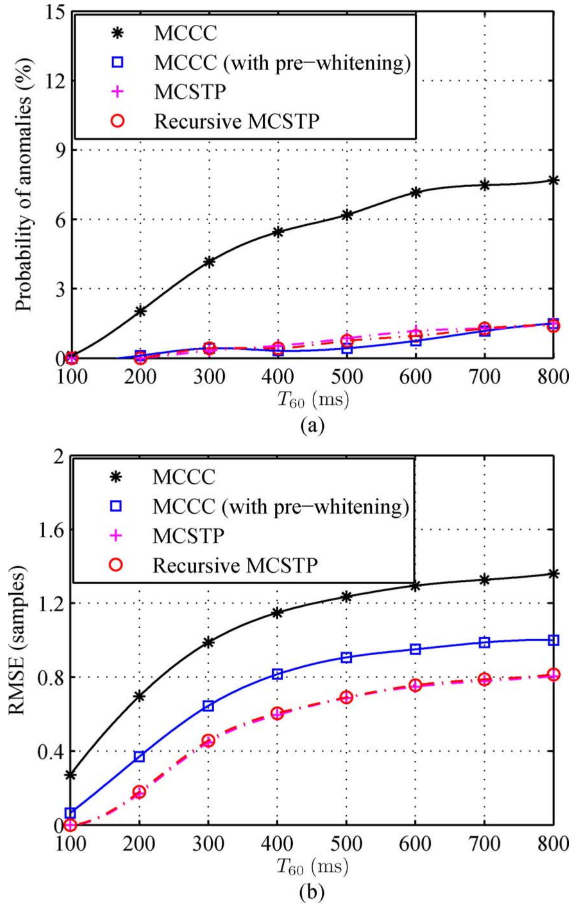


Fig. 4. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates versus T_{60}. Four microphones are used, and the prediction order is 80.

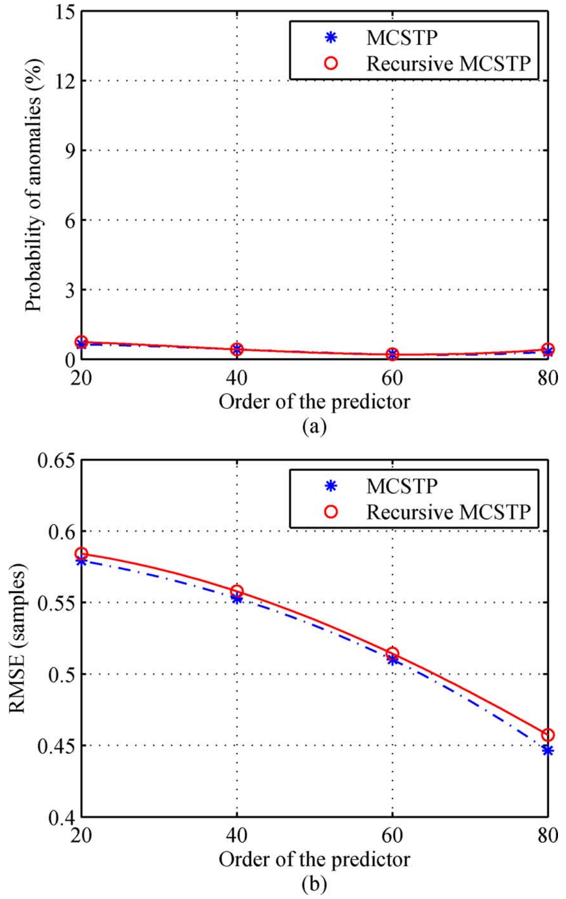


Fig. 5. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates as a function of the prediction order in a moderately reverberant environment ($T_{60} = 300$ ms). Four microphones are used.

MCCC method with pre-whitening though both have a similar value of RMSE of the nonanomalous estimates. When multiple microphones are used, the probability of anomalous estimates of MCSTP and MCCC with pre-whitening is similar. However, the RMSE of nonanomalous estimates of the MCSTP algorithm is smaller than that of the MCCC algorithm with pre-whitening. This demonstrates the robustness of the proposed MCSTP algorithm to reverberation. It is also seen from Fig. 3 that the recursive MCSTP algorithm obtains similar performance as MCSTP regardless of the number of microphones used.

Fig. 4 presents the TDE results as a function of the reverberation time T_{60} for the case where four microphones are used, and the prediction order is again set to 80. It is seen from Fig. 4 that the MCCC algorithm with pre-whitening exhibits better robustness to reverberation as compared to its counterpart without pre-whitening. It is also seen from Fig. 4 that the probability of anomalous estimates of the MCSTP algorithm is comparable to that of MCCC with pre-whitening; however, the RMSE of nonanomalous estimates of the MCSTP algorithm is evidently smaller than that of MCCC with pre-whitening. This further demonstrates the robustness of the proposed MCSTP algorithm to reverberation. It is also observed from Fig. 4 that the recursive MCSTP algorithm obtains similar performance as MCSTP regardless of the reverberation condition.

Fig. 5 depicts the TDE results versus the prediction order in a moderately reverberant environment ($T_{60} = 300$ ms) where four microphones are used. It is seen from Fig. 5 that the probability of anomalous estimates of the MCSTP method and its recursive version is small (less than one percent) and does not change much, while the RMSE of nonanomalous estimates of them decreases as the prediction order is increased, indicating that properly increasing the prediction order can improve TDE performance of the MCSTP method. It is also seen that the MCSTP algorithm and its recursive version always achieve similar performance.

The second set of experiments is to examine the performance of the four studied TDE algorithms in the situations where there are both noise and reverberation. Figs. 6 and 7 depict, respectively, the TDE results versus SNR in a moderately ($T_{60} = 300$ ms) and a lightly ($T_{60} = 120$ ms) reverberant environments, where four microphones are used, and the prediction order is again set to 80. When reverberation is dominant (e.g., $\mathrm{SNR} > 15$ dB for the moderate reverberation condition and $\mathrm{SNR} > 20$ dB for the light reverberation condition), one can see that both the MCSTP algorithm and the MCCC method with pre-whitening obtain better performance than the MCCC algorithm, again showing that using both the spatial and temporal information can help improve robustness of TDE against reverberation. However, if noise is more dominant (e.g., $\mathrm{SNR} < 5$ dB for the moderate reverberation case and $\mathrm{SNR} < 10$ dB for the light reverberation condition), the MCCC algorithm obtains

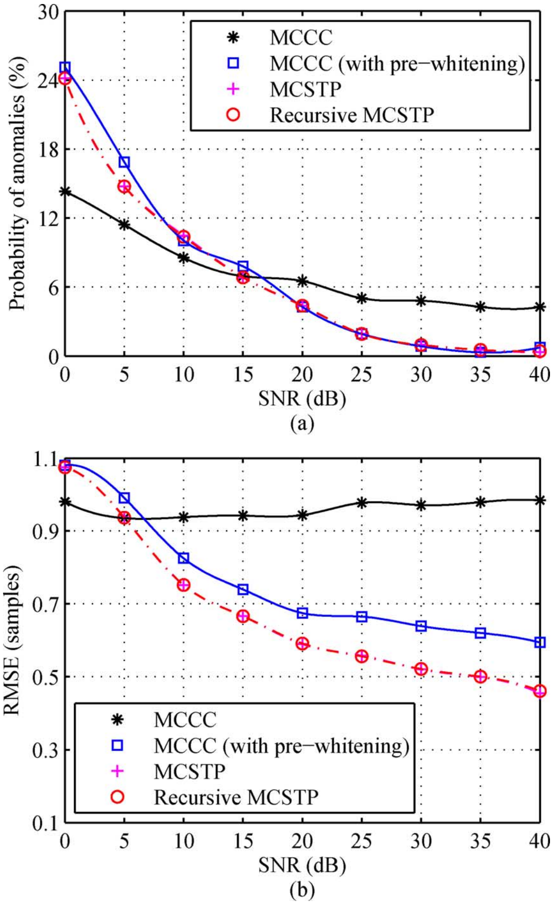


Fig. 6. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates versus SNR in a moderately reverberant environment ($T_{60} = 300$ ms). Four microphones are used and the prediction order is 80.

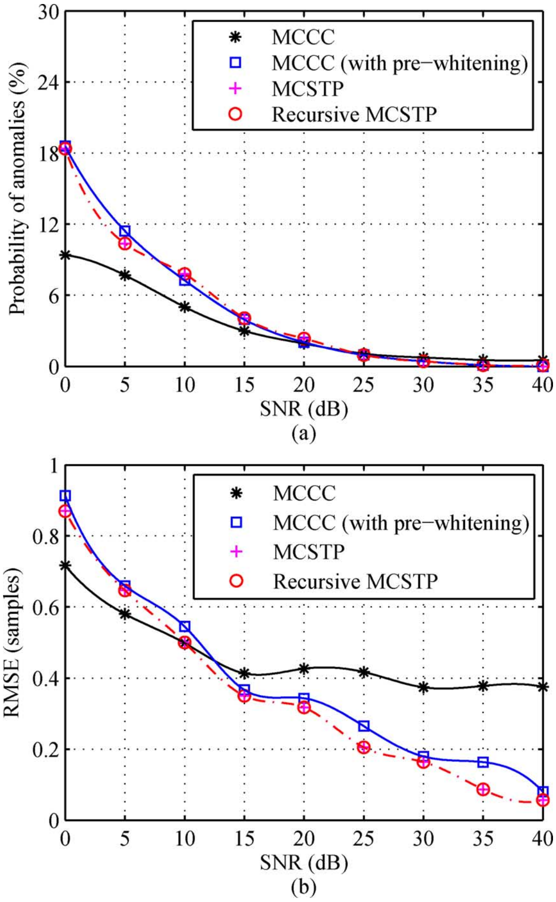


Fig. 7. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates versus SNR in a lightly reverberant environment ($T_{60} = 120$ ms). Four microphones are used and the prediction order is 80.

better performance. This is understandable. The motivation of using MCSTP or pre-whitening is to remove the impact of signal self correlation (either caused by reverberation or due to the fact that the source signal is self correlated) on TDE. When spatially and temporally white noise is very strong, it becomes difficult to reliably estimate the predictor or the pre-whitening filter.

In the previous experiments, the source signals are assumed to be speech. In the third set of experiments, we investigate the case of non-speech source signals. To this end, we first setup a recording system by a noisy urban road and record a traffic noise signal. This traffic noise is then used as the source signal to generate the microphone array outputs. Fig. 8 presents the TDE results as a function of the reverberation time T_{60} for the case where four microphones are used with the prediction order of 80. It is clearly seen from Fig. 8 that the MCSTP algorithm produces better performance than the MCCC algorithm with or without pre-whitening, which shows that the MCSTP algorithm works not only for speech signals but for non-speech signals as well.

We also carried out some experiments to study the impact of different source positions on the TDE performance. When the source position changes, the reverberation structure may change significantly though the reverberation time T_{60} stays approximately the same. This will lead to some fluctuation in the probability of anomalous estimates as well as the RMSE of nonanomalous estimates for all the TDE algorithms [19]. However, the impact of source position on TDE performance is negligible as compared to that of noise and reverberation. Therefore, the results are not plotted here to make the presentation more concise.

IV. Conclusions

In this paper, a new TDOA estimator based on MCSTP is developed. This new estimator can exploit both the spatial and temporal information embedded in the multichannel microphone signals to improve TDOA estimation performance. A theoretical analysis is presented to illustrate the underlying reason why the MCSTP algorithm is robust to reverberation. A recursive version of the MCSTP algorithm is also developed, which can achieve similar performance as MCSTP, but is more efficient in terms of computational complexity. Experiments show that MCSTP is better than MCCC (using only spatial information) in performance in the presence of reverberation, indicating that using both spatial and temporal information can help deal with reverberation. The MCSTP method is also superior to MCCC combined with pre-whitening (using both spatial and temporal information) in reverberant and noisy environments, justifying that MCSTP can jointly use spatial and temporal information in an optimal way.

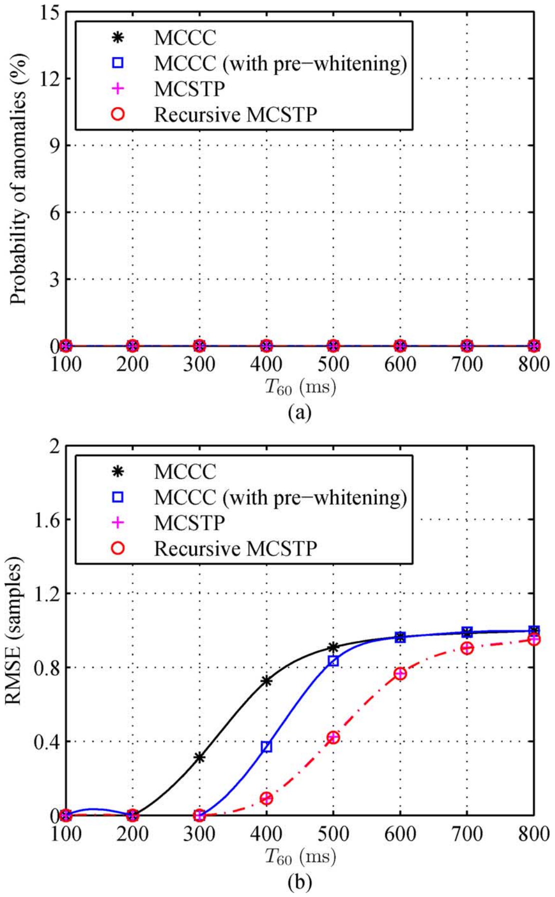


Fig. 8. Probability of (a) anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates versus T_{60}. Four microphones are used and the prediction order is 80. The source signal is a traffic noise signal pre-recorded by a busy urban main road.

Appendix
Derivations of TDE Algorithm Based on the Recursive MCSTP

The error signal vector of the multichannel forward prediction is expressed as

$$\begin{aligned}\mathbf{e}_{\mathrm{f},L}(n,p) &\triangleq \mathbf{x}(n,p) - \sum_{i=1}^{L}\mathbf{A}_{L,i}(p)\mathbf{x}(n-i,p)\\ &= \mathbf{x}(n,p) - \mathbf{A}_L^T(p)\mathbf{y}_L(n-1,p), \end{aligned} \tag{59}$$

where

$$\mathbf{A}_L(p) = [\mathbf{A}_{L,1}(p) \quad \mathbf{A}_{L,2}(p) \quad \cdots \quad \mathbf{A}_{L,L}(p)]^T \tag{60}$$

is the coefficient matrix (of size $ML \times M$) of the multichannel forward predictor, and

$$\begin{aligned}&\mathbf{y}_L(n-1,p)\\ &= [\mathbf{x}^T(n-1,p) \quad \mathbf{x}^T(n-2,p) \quad \cdots \quad \mathbf{x}^T(n-L,p)]^T \end{aligned} \tag{61}$$

is the time-shifted signal vector received at the M microphones. Then, the MSE of the multichannel forward predictor is given by

$$\begin{aligned}J_{\mathrm{f}}[\mathbf{A}_L(p)] &\triangleq E\left[\mathbf{e}_{\mathrm{f},L}^T(n,p)\mathbf{e}_{\mathrm{f},L}(n,p)\right]\\ &= E[\mathbf{x}^T(n,p)\mathbf{x}(n,p) - 2\mathbf{y}_L^T(n-1,p)\mathbf{A}_L(p)\mathbf{x}(n,p)]\\ &\quad + E\left[\mathbf{y}_L^T(n-1,p)\mathbf{A}_L(p)\mathbf{A}_L^T(p)\mathbf{y}_L(n-1,p)\right]. \end{aligned} \tag{62}$$

The derivative of $J_{\mathrm{f}}[\mathbf{A}_L(p)]$ with respect to the coefficient matrix $\mathbf{A}_L(p)$ is

$$\begin{aligned}\frac{\partial J_{\mathrm{f}}[\mathbf{A}_L(p)]}{\partial \mathbf{A}_L(p)} &= -2E\left[\mathbf{y}_L(n-1,p)\mathbf{x}^T(n,p)\right]\\ &\quad +2E\left[\mathbf{y}_L(n-1,p)\mathbf{y}_L^T(n-1,p)\right]\mathbf{A}_L(p). \end{aligned} \tag{63}$$

Thus, the Wiener-Hopf equations for the multichannel forward prediction can be obtained as follows:

$$\mathbf{R}_L(p)\mathbf{A}_{\mathrm{o},L}(p) = \mathbf{R}_{\mathrm{f}}(1/L,p), \tag{64}$$

where $\mathbf{A}_{\mathrm{o},L}(p)$ is the optimal coefficient matrix of the multichannel forward prediction,

$$\begin{aligned}\mathbf{R}_L(p) &= E\left[\mathbf{y}_L(n-1,p)\mathbf{y}_L^T(n-1,p)\right]\\ &= E\left[\mathbf{y}_L(n,p)\mathbf{y}_L^T(n,p)\right]\\ &= \begin{bmatrix} \mathbf{R}(0,p) & \mathbf{R}(1,p) & \cdots & \mathbf{R}(L-1,p)\\ \mathbf{R}^T(1,p) & \mathbf{R}(0,p) & \cdots & \mathbf{R}(L-2,p)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{R}^T(L-1,p) & \mathbf{R}^T(L-2,p) & \cdots & \mathbf{R}(0,p) \end{bmatrix}, \end{aligned} \tag{65}$$

and

$$\begin{aligned}\mathbf{R}_{\mathrm{f}}(1/L,p) &= E\left[\mathbf{y}(n-1,p)\mathbf{x}^T(n,p)\right]\\ &= [\mathbf{R}(1,p) \quad \mathbf{R}(2,p) \quad \cdots \quad \mathbf{R}(L,p)]^T. \end{aligned} \tag{66}$$

By employing the augmented correlation matrix of size $M(L+1) \times M(L+1)$:

$$\mathbf{R}_{L+1}(p) = \begin{bmatrix} \mathbf{R}(0,p) & \mathbf{R}_{\mathrm{f}}^T(1/L,p)\\ \mathbf{R}_{\mathrm{f}}(1/L,p) & \mathbf{R}_L(p) \end{bmatrix} \tag{67}$$

and the Wiener-Hopf equations for the multichannel forward prediction, the augmented multichannel Wiener-Hopf equations for the multichannel forward prediction are derived as follows:

$$\mathbf{R}_{L+1}(p)\begin{bmatrix} \mathbf{I}_{M\times M}\\ -\mathbf{A}_{\mathrm{o},L}(p) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{\mathrm{o,f},L}(p)\\ \mathbf{0}_{ML\times M} \end{bmatrix}, \tag{68}$$

where

$$\begin{aligned}\mathbf{E}_{\mathrm{o,f},L}(p) &= E\left[\mathbf{e}_{\mathrm{o,f},L}(n,p)\mathbf{e}_{\mathrm{o,f},L}^T(n,p)\right]\\ &= \mathbf{R}(0,p) - \mathbf{R}_{\mathrm{f}}^T(1/L,p)\mathbf{A}_{\mathrm{o},L}(p) \end{aligned} \tag{69}$$

is the correlation matrix (of size $M \times M$) of the forward prediction error vector, with

$$\mathbf{e}_{\mathrm{o,f},L}(n,p) \triangleq \mathbf{x}(n,p) - \mathbf{A}_{\mathrm{o},L}^T(p)\mathbf{y}_L(n-1,p). \tag{70}$$

Similar to the multichannel forward prediction, the error signal vector of the multichannel backward prediction is written as

$$\begin{aligned}\mathbf{e}_{\mathrm{b},L}(n,p) &\triangleq \mathbf{x}(n-L,p) - \sum_{i=1}^{L}\mathbf{B}_{L,i}(p)\mathbf{x}(n-i+1,p)\\ &= \mathbf{x}(n-L,p) - \mathbf{B}_L^T(p)\mathbf{y}_L(n,p), \end{aligned} \tag{71}$$

where

$$\mathbf{B}_L(p) = \left[\mathbf{B}_{L,1}(p) \quad \mathbf{B}_{L,2}(p) \quad \cdots \quad \mathbf{B}_{L,L}(p)\right]^T \tag{72}$$

is the coefficient matrix (of size $ML \times M$) of the multichannel backward predictor, and

$$\mathbf{y}_L(n,p) = \left[\mathbf{x}^T(n,p) \quad \mathbf{x}^T(n-1,p) \quad \cdots \quad \mathbf{x}^T(n-L+1,p)\right]^T \tag{73}$$

is the time-shifted signal vector received at M microphones. Then, the Wiener-Hopf equations for the multichannel backward prediction are achieved by minimizing the MSE of the multichannel backward predictor:

$$\mathbf{R}_L(p)\mathbf{B}_{\mathrm{o},L}(p) = \mathbf{R}_{\mathrm{b}}(1/L,p), \tag{74}$$

where $\mathbf{B}_{\mathrm{o},L}(p)$ is the optimal coefficient matrix of the multichannel backward prediction, and

$$\begin{aligned}\mathbf{R}_{\mathrm{b}}(1/L,p) &= E\left[\mathbf{y}(n,p)\mathbf{x}^T(n-1,p)\right]\\ &= \left[\mathbf{R}^T(L,p) \quad \mathbf{R}^T(L-1,p) \quad \cdots \quad \mathbf{R}^T(1,p)\right]^T.\end{aligned} \tag{75}$$

By employing the augmented correlation matrix of size $M(L+1) \times M(L+1)$:

$$\mathbf{R}_{L+1}(p) = \begin{bmatrix} \mathbf{R}_L(p) & \mathbf{R}_{\mathrm{b}}(1/L,p) \\ \mathbf{R}_{\mathrm{b}}^T(1/L,p) & \mathbf{R}(0,p) \end{bmatrix} \tag{76}$$

and the Wiener-Hopf equations for the multichannel backward prediction, the augmented multichannel Wiener-Hopf equations for the multichannel backward prediction can be found:

$$\mathbf{R}_{L+1}(p)\begin{bmatrix} -\mathbf{B}_{\mathrm{o},L}(p) \\ \mathbf{I}_{M\times M} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{ML\times M} \\ \mathbf{E}_{\mathrm{o,b},L}(p) \end{bmatrix}, \tag{77}$$

where

$$\begin{aligned}\mathbf{E}_{\mathrm{o,b},L}(p) &= E\left[\mathbf{e}_{\mathrm{o,b},L}(n,p)\mathbf{e}_{\mathrm{o,b},L}^T(n,p)\right]\\ &= \mathbf{R}(0,p) - \mathbf{R}_{\mathrm{b}}^T(1/L,p)\mathbf{B}_{\mathrm{o},L}(p)\end{aligned} \tag{78}$$

is the correlation matrix (of size $M \times M$) of the backward prediction error vector, with

$$\mathbf{e}_{\mathrm{o,b},L}(n,p) \triangleq \mathbf{x}(n-L,p) - \mathbf{B}_{\mathrm{o},L}^T(p)\mathbf{y}_L(n,p). \tag{79}$$

In order to find the recursive solution of the multichannel Wiener-Hopf equations, let us construct two systems. One is from (68), (75), and (76):

$$\begin{bmatrix} \mathbf{R}_L(p) & \mathbf{R}_{\mathrm{b}}(1/L,p) \\ \mathbf{R}_{\mathrm{b}}^T(1/L,p) & \mathbf{R}(0,p) \end{bmatrix}\begin{bmatrix} \mathbf{I}_{M\times M} \\ -\mathbf{A}_{\mathrm{o},L-1}(p) \\ \mathbf{0}_{M\times M} \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{\mathrm{o,f},L-1}(p) \\ \mathbf{0}_{(ML-M)\times M} \\ \mathbf{K}_{\mathrm{f},L}(p) \end{bmatrix}, \tag{80}$$

where

$$\mathbf{K}_{\mathrm{f},L}(p) = \mathbf{R}^T(L,p) - \mathbf{R}_{\mathrm{b}}^T(1/(L-1),p)\mathbf{A}_{\mathrm{o},L-1}(p). \tag{81}$$

The other system is as follows by using (66), (67), and (77):

$$\begin{bmatrix} \mathbf{R}(0,p) & \mathbf{R}_{\mathrm{f}}^T(1/L,p) \\ \mathbf{R}_{\mathrm{f}}(1/L,p) & \mathbf{R}_L(p) \end{bmatrix}\begin{bmatrix} \mathbf{0}_{M\times M} \\ -\mathbf{B}_{\mathrm{o},L-1}(p) \\ \mathbf{I}_{M\times M} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{\mathrm{b},L}(p) \\ \mathbf{0}_{(ML-M)\times M} \\ \mathbf{E}_{\mathrm{o,b},L-1}(p) \end{bmatrix}, \tag{82}$$

where

$$\mathbf{K}_{\mathrm{b},L}(p) = \mathbf{R}(L,p) - \mathbf{R}_{\mathrm{f}}^T(1/(L-1),p)\mathbf{B}_{\mathrm{o},L-1}(p). \tag{83}$$

If we post-multiply both sides of (82) by $\mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p)$, we get

$$\begin{bmatrix} \mathbf{R}(0,p) & \mathbf{R}_{\mathrm{f}}^T(1/L,p) \\ \mathbf{R}_{\mathrm{f}}(1/L,p) & \mathbf{R}_L(p) \end{bmatrix}\begin{bmatrix} \mathbf{0}_{M\times M} \\ -\mathbf{B}_{\mathrm{o},L-1}(p) \\ \mathbf{I}_{M\times M} \end{bmatrix}\mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p) = \begin{bmatrix} \mathbf{K}_{\mathrm{b},L}(p)\mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p) \\ \mathbf{0}_{(ML-M)\times M} \\ \mathbf{K}_{\mathrm{f},L}(p) \end{bmatrix}. \tag{84}$$

Subtracting (84) from (80) results

$$\mathbf{R}_{L+1}(p)\begin{bmatrix} \mathbf{I}_{M\times M} \\ -\mathbf{A}_{\mathrm{o},L-1}(p) + \mathbf{B}_{\mathrm{o},L-1}\mathbf{E}_{\mathrm{o,b},L-1}^{-1}\mathbf{K}_{\mathrm{f},L}(p) \\ -\mathbf{E}_{\mathrm{o,b},L-1}^{-1}\mathbf{K}_{\mathrm{f},L}(p) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{\mathrm{o,f},L-1}(p) - \mathbf{K}_{\mathrm{b},L}\mathbf{E}_{\mathrm{o,b},L-1}^{-1}\mathbf{K}_{\mathrm{f},L}(p) \\ \mathbf{0}_{ML\times M} \end{bmatrix}. \tag{85}$$

Comparing (68) with (85), we can obtain the following two recursions:

$$\mathbf{A}_{\mathrm{o},L}(p) = \begin{bmatrix} \mathbf{A}_{\mathrm{o},L-1}(p) - \mathbf{B}_{\mathrm{o},L-1}\mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p) \\ \mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p) \end{bmatrix} \tag{86}$$

and

$$\mathbf{E}_{\mathrm{o,f},L}(p) = \mathbf{E}_{\mathrm{o,f},L-1}(p) - \mathbf{K}_{\mathrm{b},L}(p)\mathbf{E}_{\mathrm{o,b},L-1}^{-1}(p)\mathbf{K}_{\mathrm{f},L}(p). \tag{87}$$

Similarly, if both sides of (80) are post-multiplied by $\mathbf{E}_{\mathrm{o,f},L-1}^{-1}(p)\mathbf{K}_{\mathrm{b},L}(p)$, we obtain:

$$\begin{bmatrix} \mathbf{R}(0,p) & \mathbf{R}_{\mathrm{f}}^T(1/L,p) \\ \mathbf{R}_{\mathrm{f}}(1/L,p) & \mathbf{R}_L(p) \end{bmatrix}\begin{bmatrix} \mathbf{I}_{M\times M} \\ -\mathbf{A}_{\mathrm{o},L-1}(p) \\ \mathbf{0}_{M\times M} \end{bmatrix}\mathbf{E}_{\mathrm{o,f},L-1}^{-1}(p)\mathbf{K}_{\mathrm{b},L}(p) = \begin{bmatrix} \mathbf{K}_{\mathrm{b},L}(p) \\ \mathbf{0}_{(ML-M)\times M} \\ \mathbf{K}_{\mathrm{f},L}(p)\mathbf{E}_{\mathrm{o,f},L-1}^{-1}(p)\mathbf{K}_{\mathrm{b},L}(p) \end{bmatrix}. \tag{88}$$

Subtracting (88) from (82) yields

$$\mathbf{R}_{L+1}(p)\begin{bmatrix} -\mathbf{E}_{\mathrm{o,f},L-1}^{-1}\mathbf{K}_{\mathrm{b},L}(p) \\ -\mathbf{B}_{\mathrm{o},L-1}(p) + \mathbf{A}_{\mathrm{o},L-1}\mathbf{E}_{\mathrm{o,f},L-1}^{-1}\mathbf{K}_{\mathrm{b},L}(p) \\ \mathbf{I}_{M\times M} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{ML\times M} \\ \mathbf{E}_{\mathrm{o,b},L-1}(p) - \mathbf{K}_{\mathrm{f},L}\mathbf{E}_{\mathrm{o,f},L-1}^{-1}\mathbf{K}_{\mathrm{b},L}(p) \end{bmatrix}. \tag{89}$$

Comparing (77) with (89), we can again obtain the following two recursions:

$$\mathbf{B}_{\mathrm{o},L}(p) = \begin{bmatrix} \mathbf{E}^{-1}_{\mathrm{o,f},L-1}(p)\mathbf{K}_{\mathrm{b},L}(p) \\ \mathbf{B}_{\mathrm{o},L-1}(p) - \mathbf{A}_{\mathrm{o},L-1}\mathbf{E}^{-1}_{\mathrm{o,f},L-1}(p)\mathbf{K}_{\mathrm{b},L}(p) \end{bmatrix} \tag{90}$$

and

$$\mathbf{E}_{\mathrm{o,b},L}(p) = \mathbf{E}_{\mathrm{o,b},L-1}(p) - \mathbf{K}_{\mathrm{f},L}(p)\mathbf{E}^{-1}_{\mathrm{o,f},L-1}(p)\mathbf{K}_{\mathrm{b},L}(p). \tag{91}$$

From the prediction error vectors $\mathbf{e}_{\mathrm{o,f},L-1}(n,p)$ and $\mathbf{e}_{\mathrm{o,b},L-1}(n-1,p)$, we get:

$$\begin{aligned} &E\left[\mathbf{e}_{\mathrm{o,f},L-1}(n,p)\mathbf{e}^T_{\mathrm{o,b},L-1}(n-1,p)\right] \\ &= E\{\left[\mathbf{x}(n,p) - \mathbf{A}^T_{\mathrm{o},L-1}(p)\mathbf{y}_{L-1}(n-1,p)\right] \\ &\quad \times \left[\mathbf{x}^T(n-L,p) - \mathbf{y}^T_{L-1}(n-1,p)\mathbf{B}_{\mathrm{o},L-1}(p)\right]\} \\ &= E\left\{\left[\mathbf{x}(n,p)\mathbf{x}^T(n-L,p)\right]\right\} - E\left\{\left[\mathbf{x}(n,p)\mathbf{y}^T_{L-1}(n-1,p)\right]\right\} \\ &\quad \times \mathbf{B}_{\mathrm{o},L-1}(p) \\ &\quad - \mathbf{A}^T_{\mathrm{o},L-1}(p)E\left\{\left[\mathbf{y}_{L-1}(n-1,p)\mathbf{x}^T(n-L,p)\right]\right\} \\ &\quad + \mathbf{A}^T_{\mathrm{o},L-1}(p)E\left\{\left[\mathbf{y}_{L-1}(n-1,p)\mathbf{y}^T_{L-1}(n-1,p)\right]\right\} \\ &\quad \times \mathbf{B}_{\mathrm{o},L-1}(p). \end{aligned} \tag{92}$$

It follows that

$$\begin{aligned} &E\left[\mathbf{e}_{\mathrm{o,f},L-1}(n,p)\mathbf{e}^T_{\mathrm{o,b},L-1}(n-1,p)\right] \\ &= \mathbf{R}(L,p) - \mathbf{R}^T_{\mathrm{f}}(1/(L-1),p)\mathbf{B}_{\mathrm{o},L-1}(p) \\ &\quad - \mathbf{A}^T_{\mathrm{o},L-1}(p)\mathbf{R}_{\mathrm{b}}(1/(L-1),p) + \mathbf{A}^T_{\mathrm{o},L-1}(p)\mathbf{R}_{L-1}(p)\mathbf{B}_{\mathrm{o},L-1}(p) \\ &= \mathbf{K}_{\mathrm{b},L}(p). \end{aligned} \tag{93}$$

Similarly, we have

$$E\left[\mathbf{e}_{\mathrm{o,b},L-1}(n-1,p)\mathbf{e}^T_{\mathrm{o,f},L-1}(n,p)\right] = \mathbf{K}_{\mathrm{f},L}(p). \tag{94}$$

It is seen from (93) and (94) that the following relation holds:

$$\mathbf{K}_{\mathrm{b},L}(p) = \mathbf{K}^T_{\mathrm{f},L}(p). \tag{95}$$

It should be straightforward then how to deduce the recursive algorithm given in Table I.

REFERENCES

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

[2] G. C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, pp. 463–470, Jun. 1981.

[3] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 1999, pp. 937–940.

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

[5] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1110–1124, Nov. 2003.

[6] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Elsevier Signal Process.*, vol. 85, pp. 177–204, Jan. 2005.

[7] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Process. Lett.*, vol. 12, pp. 561–564, Aug. 2005.

[8] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 1997.

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

[10] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.

[11] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, pp. 1–19, 2006.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[13] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 430–440, Feb. 2007.

[14] J. Benesty, J. Chen, and Y. Huang, "Linear prediction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008.

[15] L. Fox, *An Introduction to Numerical Linear Algebra*. Oxford, U.K.: Clarendon, 1964.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

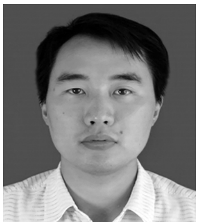
[17] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, pp. 998–1003, Dec. 1982.

[18] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 148–152, Mar. 1996.

[19] J. Chen, J. Benesty, and Y. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP J. Appl. Signal Process.*, pp. 25–36, 2005.

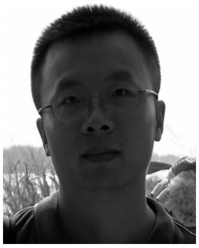
Hongsen He was born in Sichuan, China. He received the B.E. degree in automation from Southwest University of Science and Technology (SWUST), Mianyang, China, in 2000. He joined the School of Information Engineering, SWUST, as a Member of Teaching Staff in July 2000. He is currently pursuing the Ph.D. degree at the Institute of Acoustics, Nanjing University, Nanjing, China.

His main research interests include adaptive filtering, multichannel signal processing, microphone array signal processing, acoustic source localization, and adaptive noise cancellation.



Lifu Wu was born in Anhui, China, in 1981. He received the M.E. degree in electronic engineering from University of Science and Technology of China in 2005 and served as a senior engineer at Fortemedia Inc. from 2005 to 2009. He is currently pursuing the Ph.D. degree at the Key Laboratory of Modern Acoustics and Institute of Acoustics, Nanjing University, Nanjing, China.

His research interests include noise and vibration control, audio and speech signal processing.



Jing Lu received the B.S. degree in Electronic Science and Technology Department in 1999, and the Ph.D. degree in the Institute of Acoustics in 2004, both from Nanjing University. He joined the Institute of Acoustics, Nanjing University, as a lecturer in 2004. From September 2004 to March 2005, he paid a half-year academic visit to the University of Western Australia. In 2007, he was promoted to Associate Professor, and he is currently the Head of Communication Acoustics Group. He has been teaching advanced signal processing for postgraduates since 2005.

His main research interests include active noise control, echo cancellation, speech enhancement, loudspeaker and microphone arrays, and DSP implementations of acoustical signal processing algorithms. He won the Award of Backbone Young Teacher of Nanjing University in 2006, and the May 4th Youth Medal of Nanjing University in 2007. He is currently a senior member of Chinese Institute of Electronics and a member of Chinese Institute of Acoustics.

Xiaojun Qiu graduated in electronics from Peking University, China, in 1989 and received his Ph.D. degree from Nanjing University, China, in 1995 for a dissertation on active noise control. He worked in the University of Adelaide, Australia, as a Research Fellow from 1997 to 2002. He has been working in the Institute of Acoustics, Nanjing University, as a professor on Acoustics and Signal processing since 2002 and now is the Head of the Institute. He visited Germany as a Humboldt Research Fellow in 2008.

His main research areas include noise and vibration control, room acoustics, electro-acoustics and audio signal processing. He is a member of Audio Engineering Society and International Institute of Acoustics and Vibration. He has authored and co-authored more than 250 technique papers and held more than 70 patents on audio acoustics and audio signal processing.

Jingdong Chen (SM'09) received the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998.

Dr. Jingdong Chen is currently a professor at the Northwestern Polytechnical University (NWPU) in Xi'an, China. Before joining NWPU in Jan. 2011, he served as the Chief Scientist of WeVoice Inc. in New Jersey for one year. Prior to this position, he was with Bell Labs in New Jersey for nine years. Before joining Bell Labs, he held positions at the Griffith University in Brisbane, Australia and the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. His research interests include acoustic signal processing, adaptive signal processing, speech enhancement, adaptive noise/echo control, microphone array signal processing, signal separation, and speech communication. Dr. Chen is currently an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and an associate member of the IEEE Signal Processing Society (SPS) Technical Committee (TC) on Audio and Acoustic Signal Processing (AASP). He served as a member of the AASP TC from 2006 to 2009. He was the technical Co-Chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) and helped organize many other conferences. He co-authored the books *Study and Design of Differential Microphone Arrays* (Springer-Verlag, 2012), *Speech Enhancement in the STFT Domain* (Springer-Verlag, 2011), *Optimal Time-Domain Noise Reduction Filters: A Theoretical Study* (Springer-Verlag, 2011), *Speech Enhancement in the Karhunen-Loève Expansion Domain* (Morgan & Claypool, 2011), *Noise Reduction in Speech Processing* (Springer-Verlag, 2009), *Microphone Array Signal Processing* (Springer-Verlag, 2008), and *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006). He is also a co-editor/co-author of the book *Speech Enhancement* (Springer-Verlag, 2005) and a section co-editor of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, Berlin, 2007).

Dr. Chen received the 2008 Best Paper Award from the IEEE Signal Processing Society, the best paper award from the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2011, the Bell Labs Role Model Teamwork Award twice, respectively, in 2009 and 2007, the NASA Tech Brief Award twice, respectively, in 2010 and 2009, the 1998–1999 Japan Trust International Research Grant from the Japan Key Technology Center, the Young Author Best Paper Award from the 5th National Conference on Man-Machine Speech Communications in 1998, and the CAS (Chinese Academy of Sciences) President's Award in 1998.