ON TIME DELAY ESTIMATION BASED ON MULTICHANNEL SPATIOTEMPORAL SPARSE LINEAR PREDICTION

Hongsen He[1], Jingdong Chen[2], Jacob Benesty[3], and Tao Yang[1]

[1]School of Information Engineering and
Robot Technology Used for Special Environment
Key Laboratory of Sichuan Province
Southwest University of Science and Technology
Mianyang 621010, China
hongsenhe@gmail.com, yangtao98@tsinghua.org.cn

[2]Center of Immersive and
Intelligent Acoustics
Northwestern Polytechnical University
127 Youyi West Road
Xi'an 710072, China
jingdongchen@ieee.org

[3]INRS-EMT
University of Quebec
800 de la Gauchetiere Ouest
Suite 6900, Montreal
QC H5A 1K6, Canada
benesty@emt.inrs.ca

ABSTRACT

Noise and reverberation can significantly affect the performance of time delay estimation (TDE) in room acoustic environments. The multichannel cross-correlation coefficient (MCCC) algorithm, which extends the traditional cross-correlation method from two to multiple channels, can exploit the spatial information among multiple microphones to improve the robustness of TDE with respect to environmental noise; but this algorithm is not robust to reverberation. The multichannel spatiotemporal prediction (MCSTP) algorithm uses both the spatial and temporal information provided by the array. This algorithm improves significantly the robustness of TDE with respect to reverberation; however, it is found sensitive to noise. In this paper, we develop a multichannel spatiotemporal sparse prediction (MCSTSP) algorithm for TDE. This algorithm obtains a good compromise between robustness of TDE to noise and that to reverberation through making a tradeoff between pre-whitening and non-prewhitening. This is achieved via adjusting a regularization parameter, which is solved by an augmented Lagrangian alternating direction method of multipliers (ADMM). The property of this developed algorithm is justified with numerical experiments in both noisy and reverberant environments.

Index Terms— Time delay estimation (TDE), acoustic source localization, alternating direction method of multipliers (ADMM), microphone arrays, multichannel spatiotemporal sparse prediction (MCSTSP).

1. INTRODUCTION

Time delay estimation (TDE), which aims at measuring the relative time difference of arrival (TDOA) based on the signals captured by an array of sensors, play a crucial role in radar, sonar, and hands-free speech communications for localizing and tracking radiating sources [1]–[3]. TDE has been an active research topic since the landmark work, generalized cross-correlation (GCC) method, was proposed by Knapp and Carter [4], [5]. Besides the GCC method, commonly used TDE approaches also include the blind channel identification based techniques [6]–[9], the information theory based algorithms [10]–[12], and the methods exploiting some characteristics of speech signals [13], [14]. Due to its simplicity and ease of implementation, GCC [4], [5] is popularly used in the existing systems. In room acoustic environments, however, TDE using microphone arrays remains an open problem primarily due to the adverse effect of noise and reverberation.

In this paper, we develop a multichannel spatiotemporal sparse prediction (MCSTSP) algorithm to estimate TDOA. This algorithm uses the sparsity of prediction coefficient matrix of speech signals to construct an F/ℓ_1-norm optimization cost function, which is solved by the augmented Lagrangian alternating direction method of multipliers (ADMM) [15]. Through adjusting a regularization parameter, this developed algorithm can make a proper tradeoff between pre-whitening and non-prewhitening and, as a result, can achieve a good compromise between robustness to noise and robustness to reverberation. The performance of this approach is demonstrated via experiments in both noisy and reverberant environments.

2. TDE VIA MULTICHANNEL SPATIOTEMPORAL SPARSE LINEAR PREDICTION

2.1. Signal Model

Let us start from an ideal signal model for TDE, where there is a broadband sound source in the farfield, which radiates a plane wave, and we use an array of M microphones to collect the signals. If we choose the first microphone as the reference point, the signal captured by the mth microphone at time n is then modeled as

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

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, τ is the TDOA between the first and second microphones due to the source, and $f_m(\tau)$ is the relative delay between microphones 1 and m. In this paper, we consider an equispaced linear array. Therefore, we have $f_m(\tau) = (m-1)\tau$ under the far-field assumption. With the above signal model, the goal of TDE is to estimate the time delay τ given the signals received at M microphones. For a hypothesized time delay p, we use the time shifted signal $x_m[n + f_m(p)]$ to align the microphone signals. To simplify the notation, let us write $x_m[n + f_m(p)]$ as $x_m(n, p)$.

2.2. Algorithm Derivation

Let us stack the samples captured by M microphones into a vector

$$\mathbf{x}(n, p) = \begin{bmatrix} x_1(n,p) & x_2(n,p) & \cdots & x_M(n,p) \end{bmatrix}^T, \tag{2}$$

This work was supported in part by the National Science Foundation of China (NSFC) (Grant No. 61571376), the NSFC "Distinguished Young Scientists Fund" (Grant No. 61425005), the Incubation Program for the Distinguished Youth Foundation of Sichuan Province of China (Grant No. 2014JQ0042), the Open Foundation of the Key Laboratory of Modern Acoustics of Nanjing University (Grant No. 1302), the Doctoral Foundation of Southwest University of Science and Technology (Grant No. 13zx7149), and the Open Foundation of Robot Technology Used for Special Environment Key Laboratory of Sichuan Province (Grant No. 13zxtk06).

where $(\cdot)^T$ denotes the transpose of a vector or matrix. We also define another vector of the mth channel at time $n-1$ as follows:

$$\boldsymbol{x}_m(n-1,p) = \left[\begin{array}{cccc} x_m(n-1,p) & x_m(n-2,p) & \cdots & x_m(n-K,p) \end{array}\right]^T. \quad (3)$$

Now, we consider predicting $\mathbf{x}(n,p)$ in (2) from the past samples of the M channels $\boldsymbol{x}_1(n-1,p), \boldsymbol{x}_2(n-1,p), \ldots, \boldsymbol{x}_M(n-1,p)$, i.e.,

$$\begin{aligned}\widehat{\boldsymbol{x}}(n,p) = \boldsymbol{\mathcal{A}}_1(p)\boldsymbol{x}_1(n-1,p) + \boldsymbol{\mathcal{A}}_2(p)\boldsymbol{x}_2(n-1,p) \\ + \cdots + \boldsymbol{\mathcal{A}}_M(p)\boldsymbol{x}_M(n-1,p),\end{aligned} \quad (4)$$

where $\boldsymbol{\mathcal{A}}_m(p) \in \mathbb{R}^{M\times K}$, $m = 1, 2, \ldots, M$, are the coefficient matrices of the multichannel forward predictor. The prediction error vector can then be written as

$$\begin{aligned}\boldsymbol{\epsilon}(n,p) &= \mathbf{x}(n,p) - \widehat{\boldsymbol{x}}(n,p) \\ &= \mathbf{x}(n,p) - \boldsymbol{\mathcal{A}}^T(p)\boldsymbol{y}(n-1,p),\end{aligned} \quad (5)$$

where

$$\boldsymbol{\epsilon}(n,p) = \left[\begin{array}{cccc} \epsilon_1(n,p) & \epsilon_2(n,p) & \cdots & \epsilon_M(n,p) \end{array}\right]^T, \quad (6)$$

$$\boldsymbol{\mathcal{A}}(p) = \left[\begin{array}{cccc} \boldsymbol{\mathcal{A}}_1(p) & \boldsymbol{\mathcal{A}}_2(p) & \cdots & \boldsymbol{\mathcal{A}}_M(p) \end{array}\right]^T \quad (7)$$

is the $KM \times M$ coefficient matrix of the multichannel forward prediction-error filter, and

$$\boldsymbol{y}(n-1,p) = \left[\begin{array}{cccc} \boldsymbol{x}_1^T(n-1,p) & \boldsymbol{x}_2^T(n-1,p) & \cdots & \boldsymbol{x}_M^T(n-1,p) \end{array}\right]^T \quad (8)$$

is the time-shifted signal vector received at M microphones. In matrix form, the error vector in (5) can be written as

$$\boldsymbol{\mathcal{E}}(n,p) = \mathbf{X}(n,p) - \boldsymbol{\mathcal{Y}}(n,p)\boldsymbol{\mathcal{A}}(p), \quad (9)$$

where

$$\boldsymbol{\mathcal{E}}(n,p) = \left[\begin{array}{cccc} \boldsymbol{\epsilon}(n,p) & \boldsymbol{\epsilon}(n+1,p) & \cdots & \boldsymbol{\epsilon}(n+K+L-1,p) \end{array}\right]^T, \quad (10)$$

$$\mathbf{X}(n,p) = \left[\begin{array}{cccc} \mathbf{x}(n,p) & \mathbf{x}(n+1,p) & \cdots & \mathbf{x}(n+K+L-1,p) \end{array}\right]^T, \quad (11)$$

$$\boldsymbol{\mathcal{Y}}(n,p) = \left[\begin{array}{cccc} \boldsymbol{y}(n-1,p) & \boldsymbol{y}(n,p) & \cdots & \boldsymbol{y}(n+K+L-2,p) \end{array}\right]^T. \quad (12)$$

The configuration of the traditional linear predictor uses a cascade of a long-term predictor and a short-term predictor [16], [17]. The consequent prediction coefficient vector is highly sparse [18]. In the case of multichannel linear prediction, the prediction coefficient matrix is also of sparsity, as illustrated in Fig. 1(a). This sparsity, however, greatly deteriorates when noise is present, which can be seen from Fig. 1(b). Since the prediction coefficient matrix is sparse for clean speech signals, we can use this property to improve the robustness of the estimation of the linear predictor in noise. To this end, we impose a sparse regularization term to the least squares criterion. Then, we propose the following F/ℓ_1-norm optimization criterion to preprocess the microphone signals:

$$\min_{\boldsymbol{\mathcal{A}}(p)} \left\{ \frac{1}{2} \left\| \mathbf{X}(n,p) - \boldsymbol{\mathcal{Y}}(n,p)\boldsymbol{\mathcal{A}}(p) \right\|_F^2 + \lambda \left\| \boldsymbol{\mathcal{A}}(p) \right\|_{\ell_1} \right\}, \quad (13)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\|\cdot\|_{\ell_1}$ stands for the ℓ_1-norm of a matrix (i.e., the sum of the absolute values of all the entries of the matrix), and the parameter $\lambda > 0$ is a scalar regularization parameter.

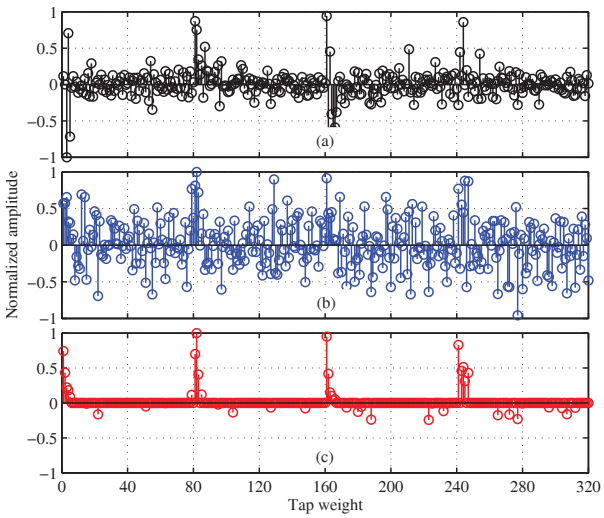

Fig. 1. Illustration of the MCSTSP coefficient matrix, where the predictor length is 80 and four microphones are used. (a) A column vector of the MCSTSP coefficient matrix of a clean speech signal; (b) the column vector of the MCSTSP coefficient matrix estimated using the least squares method (F-norm criterion) at the SNR of 5 dB; (c) the column vector of the MCSTSP coefficient matrix estimated with the F/ℓ_1-norm criterion at the SNR of 5 dB (δ=0.1).

It is obvious that (13) is a convex optimization problem, which can be solved by many existing methods, such as the linear programming [19], the interior point method [20], the primal-dual interior point method [21], etc. In this work, we adopt the ADMM, which can efficiently use the separability of multiple variables [15] to solve this problem.

By means of an auxiliary matrix $\boldsymbol{\mathcal{Z}}(p)$, (13) can be equivalently written as

$$\min_{\boldsymbol{\mathcal{A}}(p),\boldsymbol{\mathcal{Z}}(p)} \left\{ \frac{1}{2} \left\| \mathbf{X}(n,p) - \boldsymbol{\mathcal{Y}}(n,p)\boldsymbol{\mathcal{A}}(p) \right\|_F^2 + \lambda \left\| \boldsymbol{\mathcal{Z}}(p) \right\|_{\ell_1} : \boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p) = \mathbf{0} \right\}. \quad (14)$$

This is an augmented Lagrangian subproblem, which can be formulated as

$$\begin{aligned}\min_{\boldsymbol{\mathcal{A}}(p),\boldsymbol{\mathcal{Z}}(p)} \Bigg\{ \frac{1}{2} \left\| \mathbf{X}(n,p) - \boldsymbol{\mathcal{Y}}(n,p)\boldsymbol{\mathcal{A}}(p) \right\|_F^2 + \lambda \left\| \boldsymbol{\mathcal{Z}}(p) \right\|_{\ell_1} \\ + \langle \boldsymbol{\Theta}(p), \boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p) \rangle + \frac{\beta}{2} \left\| \boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p) \right\|_F^2 \Bigg\},\end{aligned} \quad (15)$$

where $\boldsymbol{\Theta}(p) \in \mathbb{R}^{KM\times M}$ is the multiplier of the linear constraint,

$$\begin{aligned}\langle \boldsymbol{\Theta}(p), \boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p) \rangle &= \frac{1}{2}\mathrm{tr}\Big\{ \boldsymbol{\Theta}^T(p)[\boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p)] \\ &\quad + [\boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p)]^T \boldsymbol{\Theta}(p) \Big\} \\ &= \mathrm{tr}\Big\{ \boldsymbol{\Theta}^T(p)[\boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}(p)] \Big\}\end{aligned} \quad (16)$$

denotes the matrix inner product with $\mathrm{tr}(\cdot)$ being the trace of a matrix, $\beta > 0$ is a penalty parameter for the violation of the linear constraint. The augmented term, i.e., the fourth term within the braces of (15), is introduced to ensure that the objective function is strictly convex. Given $[\boldsymbol{\mathcal{Z}}_k(p), \boldsymbol{\Theta}_k(p)]$, we can obtain $[\boldsymbol{\mathcal{A}}_{k+1}(p), \boldsymbol{\mathcal{Z}}_{k+1}(p), \boldsymbol{\Theta}_{k+1}(p)]$ by alternating minimization of (15) with respect to one variable while keeping the other variables fixed. First, when $\boldsymbol{\mathcal{Z}}(p) = \boldsymbol{\mathcal{Z}}_k(p)$ and $\boldsymbol{\Theta}(p) = \boldsymbol{\Theta}_k(p)$ are fixed, the minimization of (15) with respect to $\boldsymbol{\mathcal{A}}(p)$ is equivalent to

$$\min_{\boldsymbol{\mathcal{A}}(p)} \left\{ \frac{1}{2} \|\mathbf{X}(n,p) - \boldsymbol{\mathcal{Y}}(n,p)\boldsymbol{\mathcal{A}}(p)\|_F^2 + \frac{\beta}{2} \|\boldsymbol{\mathcal{A}}(p) - \boldsymbol{\mathcal{Z}}_k(p) + \boldsymbol{\Theta}_k(p)/\beta\|_F^2 \right\}, \quad (17)$$

whose solution is

$$\boldsymbol{\mathcal{A}}_{k+1}(p) = \left[\boldsymbol{\mathcal{Y}}^T(n,p)\boldsymbol{\mathcal{Y}}(n,p) + \beta\mathbf{I}\right]^{-1} \times \left[\boldsymbol{\mathcal{Y}}^T(n,p)\mathbf{X}(n,p) + \beta\boldsymbol{\mathcal{Z}}_k(p) - \boldsymbol{\Theta}_k(p)\right], \quad (18)$$

where $\mathbf{I}$ denotes the identity matrix of size $KM \times KM$.

Then, when $\boldsymbol{\mathcal{A}}(p) = \boldsymbol{\mathcal{A}}_{k+1}(p)$ and $\boldsymbol{\Theta}(p) = \boldsymbol{\Theta}_k(p)$ are fixed, the minimization of (15) with respect to $\boldsymbol{\mathcal{Z}}(p)$ is equivalent to

$$\min_{\boldsymbol{\mathcal{Z}}(p)} \left\{ \lambda\|\boldsymbol{\mathcal{Z}}(p)\|_{\ell_1} + \frac{\beta}{2} \|\boldsymbol{\mathcal{A}}_{k+1}(p) - \boldsymbol{\mathcal{Z}}(p) + \boldsymbol{\Theta}_k(p)/\beta\|_F^2 \right\}. \quad (19)$$

Let $\boldsymbol{\Phi}(p) = \boldsymbol{\mathcal{A}}_{k+1}(p) - \boldsymbol{\mathcal{Z}}(p) + \boldsymbol{\Theta}_k(p)/\beta$, then (19) can be rewritten as

$$\min_{\boldsymbol{\mathcal{Z}}(p)} \left\{ \lambda\|\boldsymbol{\mathcal{Z}}(p)\|_{\ell_1} + \frac{\beta}{2} \|\boldsymbol{\Phi}(p)\|_F^2 \right\} = \min_{\boldsymbol{\mathcal{Z}}(p)} \left\{ \sum_{i=1}^{KM} \sum_{j=1}^{M} \left[\lambda|(\boldsymbol{\mathcal{Z}}(p))^{i,j}| + \frac{\beta}{2} \left|(\boldsymbol{\Phi}(p))^{i,j}\right|^2 \right] \right\}, \quad (20)$$

where $(\cdot)^{i,j}$ denotes the (i,j)th element of a matrix. It can be seen from (20) that the variables $(\boldsymbol{\mathcal{Z}}(p))^{i,j}$, $i = 1, 2, \ldots, KM, j = 1, 2, \ldots, M$, are decoupled (separable). Hence we obtain a simple problem of minimizing a scalar function given by

$$\min_{(\boldsymbol{\mathcal{Z}}(p))^{i,j}} \left\{ \lambda|(\boldsymbol{\mathcal{Z}}(p))^{i,j}| + \frac{\beta}{2} \left|(\boldsymbol{\mathcal{A}}_{k+1}(p) - \boldsymbol{\mathcal{Z}}(p) + \boldsymbol{\Theta}_k(p)/\beta)^{i,j}\right|^2 \right\}, \quad (21)$$

the solution of which is readily achieved as

$$(\boldsymbol{\mathcal{Z}}_{k+1}(p))^{i,j} = \begin{cases} \frac{(\boldsymbol{\Psi}(p))^{i,j}}{|(\boldsymbol{\Psi}(p))^{i,j}|} \max\left(|(\boldsymbol{\Psi}(p))^{i,j}| - \lambda/\beta, 0\right), & |(\boldsymbol{\mathcal{Z}}(p))^{i,j}| \neq 0, \\ 0, & |(\boldsymbol{\mathcal{Z}}(p))^{i,j}| = 0, \end{cases} \quad (22)$$

where $\boldsymbol{\Psi}(p) = \boldsymbol{\mathcal{A}}_{k+1}(p) + \boldsymbol{\Theta}_k(p)/\beta$. These solutions can also be compactly formulated by a soft-thresholding operator, i.e.,

$$\boldsymbol{\mathcal{Z}}_{k+1}(p) = \mathrm{soft}\left(\boldsymbol{\mathcal{A}}_{k+1}(p) + \boldsymbol{\Theta}_k(p)/\beta, \lambda/\beta\right), \quad (23)$$

where the soft function is defined as

$$\mathrm{soft}\left(\boldsymbol{\Omega}, \mu\right) = \mathrm{sgn}\left(\boldsymbol{\Omega}\right) \odot \max\left\{|\boldsymbol{\Omega}| - \mu, 0\right\}, \quad \forall\, \boldsymbol{\Omega} \in \mathbb{R}^{KM \times M}, \mu > 0, \quad (24)$$

$\mathrm{sgn}\,(\cdot)$ is the signum function, $\odot$ denotes the dot product of two matrices, all the other operations are performed in a component-wise way. Finally, the Lagrangian multiplier matrix $\boldsymbol{\Theta}(p)$ is updated by

$$\boldsymbol{\Theta}_{k+1}(p) = \boldsymbol{\Theta}_k(p) + \beta\left(\boldsymbol{\mathcal{A}}_{k+1}(p) - \boldsymbol{\mathcal{Z}}_{k+1}(p)\right). \quad (25)$$

Therefore, we obtain the solution of (14) by iteratively calculating (18), (23), (25), and so get the suboptimal prediction coefficient matrix $\boldsymbol{\mathcal{A}}^{\mathrm{s}}(p)$. It is found from Fig. 1(c) that we can obtain a sparse prediction coefficient matrix via this F/ℓ_1-norm optimization algorithm in a noisy environment.

Substituting $\boldsymbol{\mathcal{A}}^{\mathrm{s}}(p)$ into (9), we then achieve the suboptimal prediction error matrix $\boldsymbol{\mathcal{E}}^{\mathrm{s}}(n,p)$, from which we can obtain the cross-correlation matrix of the prediction error signals $\boldsymbol{\mathcal{R}}(p)$. Therefore, we define an MCCC function according to $\boldsymbol{\mathcal{R}}(p)$ [22]–[25] and then obtain a new TDOA estimator based on MCSTSP.

2.3. Regularization Parameter

It is found from (14) that the parameter λ plays an important role in controlling the sparse level of the prediction coefficient matrix. This parameter is mostly affected by the microphone signals, i.e., $\mathbf{X}(n,p)$ and $\boldsymbol{\mathcal{Y}}(n,p)$. Herein, we consider to determine λ by the following choice:

$$\lambda = \delta\|\mathbf{X}^T(n,p)\boldsymbol{\mathcal{Y}}(n,p)\|_{\ell_\infty}, \quad (26)$$

where

$$\|\mathbf{Z}\|_{\ell_\infty} = \max_j \sum_{i=1}^{M} |z_{i,j}| \quad (27)$$

denotes the ℓ_∞-norm for any matrix $\mathbf{Z}$ with the (i,j)th entry of $z_{i,j}$ and δ is a positive number.

3. EXPERIMENTS

3.1. Experimental Environments

All the experiments are carried out in a simulated room of size $7\ \mathrm{m} \times 6\ \mathrm{m} \times 3\ \mathrm{m}$. An equispaced linear array, which consists of six omnidirectional microphones with the inter-element spacing being 0.1 m, is used to collect the microphones' output signals. 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 source signal is a prerecorded clean speech signal, which is sampled at 16 kHz, and the length of the signal is approximately 1 min. The impulse responses from the source to the six microphones are generated using the image model [26]. The length of the impulse responses is 2048 samples. 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).

In the simulations, the microphone signals are partitioned into nonoverlapping frames with a frame length of 64 ms. Each frame is windowed with a Hamming window, and a time delay estimate is then obtained. Two performance metrics, namely the probability of anomalous estimates and the root mean square error (RMSE) of nonanomalous estimates, are used to evaluate the performance of the TDE algorithms (see [22], [27], and [28] for the definition of these two metrics and how to classify an estimate as an anomaly or a nonanomaly). The total number of speech frames used for statistics is 936 (the frame length is 1024 samples). The true time delay from the sound source to the first two microphones is 2.0 samples.

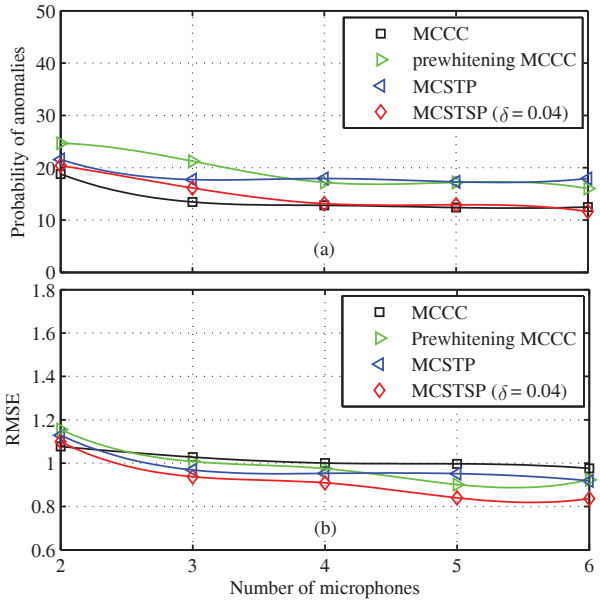

Fig. 2. TDE performance versus the number of microphones in a noisy (SNR=10 dB) and moderately reverberant (T_{60}=300 ms) environment: (a) the probability of anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates.

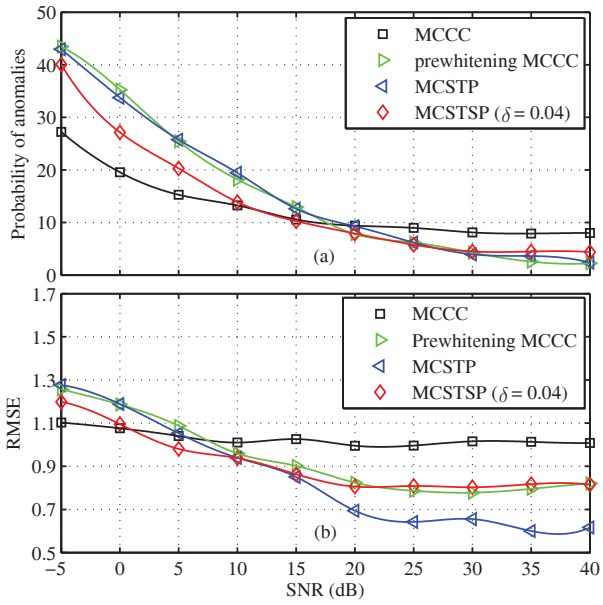

Fig. 3. TDE performance versus SNR in a moderately reverberant (T_{60}=300 ms) environment: (a) the probability of anomalous time delay estimates and (b) RMSE of nonanomalous time delay estimates.

3.2. Results

Figure 2 shows the the TDE results versus the number of microphones in a noisy (SNR=10 dB) and moderately reverberant (T_{60}=300 ms) environment. It is seen that the probability of anomalous time delay estimates and RMSEs of nonanomalous time delay estimates of the four algorithms generally decrease as the number of microphones increases, which shows the effectiveness of the developed method in improving the TDE robustness by taking advantage of the spatial and temporal information provided by the multiple sensors. For the case of two microphones, all the pre-whitening TDE algorithms do not achieve obvious superiority, while the original multichannel cross-correlation coefficient (MCCC) algorithm [22], [23] obtains slightly better performance. When multiple microphones are employed, the MCCC algorithm has small probability of anomalous estimates, but the corresponding RMSE is large. Although MCCC with pre-whitening [24] and multichannel spatiotemporal prediction (MCSTP) [25] algorithms obtain moderate RMSEs, the probability of anomalous estimates is large. The MCSTSP algorithm, in comparison, achieves a good performance in terms of both the anomalous estimates and the RMSE, which implies that the multichannel sparse prediction is effective for TDE in dealing with both noise and reverberation.

Figure 3 illustrates the TDE results versus SNR in a moderately reverberant (T_{60}=300 ms) environment. It is seen that the original MCCC algorithm yields the best performance in low SNR environments, but it is most sensitive to reverberation when SNR is high. The pre-whitening MCCC algorithm achieves good robustness to reverberation as compared to MCCC. The MCSTP and pre-whitening MCCC algorithms obtain comparable TDE performance in low SNR environments; the former, however, has a better robustness to reverberation due to its optimal pre-whitening ability [25]. Although the two TDE algorithms with pre-whitening obtain considerable performance improvement under reverberation conditions, they suffer from performance degradation when noise is strong. The developed MCSTSP algorithm achieves a good compromise between MCCC and MCSTP.

4. CONCLUSIONS

In this paper, a new time delay estimator based on MCSTP is developed from a multichannel sparse linear prediction perspective. This algorithm uses the sparsity of the spatiotemporal prediction coefficient matrix to improve the TDE performance. The MCSTSP algorithm is effectively solved by ADMM. Experimental results show that the MCSTSP algorithm offers an effective compromise between the MCCC and MCSTP algorithms. A proper value of δ needs to be found in practical applications, depending on the level of noise and reverberation.

5. RELATION TO PRIOR WORK

It is a difficult and challenging problem to make TDE robust to both noise and reverberation in room acoustic environments. The MCCC algorithm extends the traditional cross-correlation method from two to multiple channels. It exploits the spatial information among multiple microphones to improve the robustness of TDE with respect to noise [22], [23]; but the MCCC is found sensitive to reverberation. The MCSTP algorithm, which exploits both spatial and temporal information, improves the robustness of TDE to reverberation due to its optimal pre-whitening ability [25]; this algorithm, however, is sensitive to additive noise. In an earlier study [29], we developed a two-channel sparse linear prediction algorithm for TDE, which transforms the TDE problem into one of ℓ_2/ℓ_1-norm based optimization by introducing a sparse regularization term to the least squares criterion. This algorithm can improve TDE performance in both noise and reverberation. This paper generalizes the work in [29] from the two-channel to the multichannel case.

6. REFERENCES

[1] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin, Germany: Springer, 2006.

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

[3] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 1082–1095, Jun. 2014.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] J. Chen, Y. Huang, and J. Benesty, "An adaptive blind SIMO identification approach to joint multichannel time delay estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2004, pp. IV-53–IV-56.

[10] 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.

[11] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, pp. 157–160, Mar. 2007.

[12] H. He, J. Lu, L. Wu, X. Qiu, "Time delay estimation via non-mutual information among multiple microphones," *Appl. Acoust.*, vol. 74, pp. 1033–1036, Aug. 2013.

[13] 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.

[14] 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.

[15] J. F. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[16] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-37, no.4, pp. 467–478, Apr. 1989.

[17] E. D. Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. New York: Springer, 2001.

[18] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, 2012.

[19] X. Jiang, T. Kirubarajan, and W. J. Zeng, "Robust sparse channel estimation and equalization in impulsive noise using linear programming," *Elsevier Signal Process.*, vol. 93, no. 5, pp. 1095–1105, May, 2013.

[20] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Select. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[21] S. J. Wright, *Primal-Dual Interior-Point Methods*. Philadelphia, PA: SIAM, 1997.

[22] 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.

[23] 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.

[24] 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.

[25] H. He, L. Wu, J. Lu, X. Qiu, and J. Chen, "Time difference of arrival estimation exploiting multichannel spatio-temporal prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, pp. 463–475, Mar. 2013.

[26] 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.

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

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

[29] H. He, T. Yang, and J. Chen, "On time delay estimation from a sparse linear prediction perspective," *J. Acoust. Soc. Amer.*, vol. 137, no. 2, pp. 1044–1047, Feb. 2015.