ROBUST MULTICHANNEL TDOA ESTIMATION FOR SPEAKER LOCALIZATION USING THE IMPULSIVE CHARACTERISTICS OF SPEECH SPECTRUM

Hongsen He[†], Jingdong Chen[‡], Jacob Benesty[♯], Yingyue Zhou[†], and Tao Yang[†]

[†]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

[‡]Center of Immersive and
Intelligent Acoustics
Northwestern Polytechnical University
127 Youyi West Road
Xi'an 710072, China

[♯]INRS-EMT
University of Quebec
800 de la Gauchetiere Ouest
Suite 6900, Montreal
QC H5A 1K6, Canada

ABSTRACT

Time delay estimation (TDE) plays an important role in localizing and tracking radiating acoustic sources. Although many efforts have been devoted to this problem in the literature, the robustness of TDE with respect to noise and reverberation remains a great challenge for practical systems. In this paper, we investigate the TDE problem in acoustic single-input/multiple-output (SIMO) systems in reverberant and noisy environments. We first define a Cauchy estimator in the frequency domain, which is robust in dealing with speech as the SIMO system's excitation. This robust estimator is then used to construct a cost function, from which a robust multichannel frequency-domain adaptive filter is deduced. This adaptive algorithm is subsequently employed to blindly identify the acoustic impulse responses between the source and the microphones. Finally, the time difference of arrival is determined from the identified channel responses.

Index Terms—Acoustic source localization, time delay estimation, microphone arrays, multichannel frequency-domain adaptive filter.

1. INTRODUCTION

Time delay estimation (TDE), which aims at measuring the time difference of arrival (TDOA) based on the signals captured by an array of sensors, plays a crucial role in hands-free speech communications for localizing and tracking radiating acoustic sources [1], [2]. A great deal of efforts have been devoted to this problem in the literature and many methods have been developed including the well-known generalized cross-correlation (GCC) method [3], [4], the blind channel identification based techniques [5]–[9], the multichannel linear prediction based approaches [10]–[12], the information theory based algorithms [13]–[15], etc. While most of these approaches can achieve reasonable accurate estimates in favorable environments, the robustness of TDE remains a challenging problem. There are three major sources that affect significantly the performance of TDE: noise, reverberation, and nonstationarity and nonwhiteness of the excitation signals. To deal with the three factors and improve the robustness of TDE in practical systems, we develop in this paper a robust multichannel approach to TDE in acoustic single-input/multiple-output (SIMO) systems. First, we use the Cauchy estimator to define a frequency-domain cost function, which is subsequently used to deduce an adaptive multichannel algorithm to blindly identify the acoustic SIMO system. This adaptive algorithm is subsequently employed to blindly identify the acoustic impulse responses between the source and the microphones. Finally, TDOA is determined from the identified channel responses.

This work was supported in part by the NSFC (Grant Nos. 61571376, 61401379) and the NSFC "Distinguished Young Scientists Fund" (Grant No. 61425005), also supported in part by the Open Foundation of the Key Laboratory of Modern Acoustics of Nanjing University (Grant No. 1302).

2. ROBUST TDE VIA FREQUENCY-DOMAIN BLIND SYSTEM IDENTIFICATION

2.1. Robust Adaptive Blind Multichannel Identification

Assume that an acoustic SIMO system is composed of one acoustic source and M microphones. Usually, this multichannel system can be blindly identified based on the well-known cross relation that the output of any one channel convolved with the impulse response of another channel is equal to the output of that other channel convolved with the impulse response of this channel if the additive noise is neglected [16], [17]. With a system of M channels and in the presence of additive noise or/and modeling errors, one can write the following error signals [18], [19]:

$$e_{ij}(n) = \mathbf{x}_i^T(n)\widehat{\mathbf{h}}_j(n) - \mathbf{x}_j^T(n)\widehat{\mathbf{h}}_i(n), \tag{1}$$

where $i, j = 1, 2, \ldots, M$, $i \neq j$, $e_{ij}(n)$ is the *a priori* error signal between the ith and jth channels, $\mathbf{x}_i$ is the observation signal vector of the ith channel of length L, and $\widehat{\mathbf{h}}_i(n)$ is an estimate of the impulse response vector $\mathbf{h}_i$ of the ith channel of length L at time n. The error signal can be used to define a conditional cost function via a square function or a robust cost function based on an M-estimator in the time domain [19].

For typical acoustic channels, the magnitude of their transfer functions is often flat. However, the amplitude spectra of the excitation signals, which are speech most of the time, typically have an impulse-like structure [20]. In the course of adaptive iteration, the impulse responses $h_i(n)$ and $h_j(n)$ are often estimated crudely, which implies that $[h_i(n) * \widehat{h}_j(n) - h_j(n) * \widehat{h}_i(n)]$ ($*$ denotes linear convolution) is such far away from zero that it emphasizes the presence of the impulsive spectrum of the excitation signal $s(n)$ in the amplitude spectrum of the error signal $e_{ij}(n)$. As a result, the frequency-domain cost function based on the square error is dominated by the large peaks in the spectra of the excitation speech signals, which seriously affects the accuracy and robustness of channel identification. In this paper, we consider to transform the error signal $e_{ij}(n)$ into the frequency domain denoted as $\underline{e}_{ij}(n)$ to develop an adaptive frequency-domain algorithm for the improvement of channel identification with speech excitations. Then, we propose to define a cost function based on an M-estimator:

$$\mathcal{J}_\rho(m) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \sum_{n=mL}^{mL+L-1} \rho\left[|\underline{e}_{ij}(n)|\right], \tag{2}$$

where

$$\begin{aligned}\left[\underline{e}_{ij}(mL)\ \ \underline{e}_{ij}(mL+1)\ \cdots\ \underline{e}_{ij}(mL+L-1)\right]^T &= \underline{\boldsymbol{e}}_{ij}(m) \\ &= \boldsymbol{\mathcal{G}}_{L\times 2L}^{01}\left[\boldsymbol{\mathcal{D}}_{x_i}(m)\boldsymbol{\mathcal{G}}_{2L\times L}^{10}\underline{\widehat{\boldsymbol{h}}}_j(m) - \boldsymbol{\mathcal{D}}_{x_j}(m)\boldsymbol{\mathcal{G}}_{2L\times L}^{10}\underline{\widehat{\boldsymbol{h}}}_i(m)\right],\end{aligned} \tag{3}$$

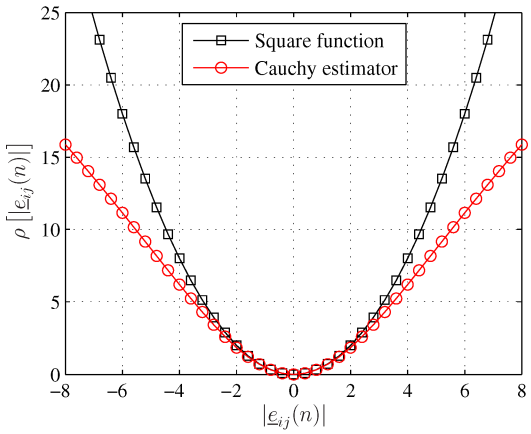

Fig. 1. Comparison between the square function and the Cauchy estimator (with $c = 5.0$) in the frequency domain.

$$\boldsymbol{\mathcal{D}}_{x_i}(m) = \text{diag}\{\mathbf{F}_{2L\times 2L}\mathbf{x}_{i,2L}(m)\}, \tag{4}$$

$$\mathbf{x}_{i,2L}(m) = [x_i(mL-L)\;\; x_i(mL-L+1) \;\; \cdots \;\; x_i(mL+L-1)]^T, \tag{5}$$

$$\boldsymbol{\mathcal{G}}^{01}_{L\times 2L} = \mathbf{F}_{L\times L}\left[\begin{array}{cc}\mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{array}\right]\mathbf{F}^{-1}_{2L\times 2L}, \tag{6}$$

$$\boldsymbol{\mathcal{G}}^{10}_{2L\times L} = \mathbf{F}_{2L\times 2L}\left[\begin{array}{cc}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\end{array}\right]^T\mathbf{F}^{-1}_{L\times L}, \tag{7}$$

$$\underline{\widehat{\boldsymbol{h}}}_i(m) = \mathbf{F}_{L\times L}\widehat{\mathbf{h}}_i(m), \tag{8}$$

$\mathbf{F}_{L\times L}$ and $\mathbf{F}^{-1}_{L\times L}$ are, respectively, the Fourier and inverse Fourier matrices of size $L\times L$, $\mathbf{0}_{L\times L}$ is the null matrix of size $L\times L$, $\mathbf{I}_{L\times L}$ is the identity matrix of size $L\times L$, diag$[\cdot]$ denotes a diagonal matrix with diagonal entries from the indicated vector, m is the block index, and $\rho[\cdot]$ is a robust frequency-domain M-estimator. In this work, we use the Cauchy estimator [21], which is written as

$$\rho\left[|\underline{e}_{ij}(n)|\right] = \frac{c^2}{2}\log\left[1+\left(\frac{|\underline{e}_{ij}(n)|}{c}\right)^2\right], \tag{9}$$

where the parameter c is a positive constant. The comparison between the Cauchy estimator (with $c = 5.0$) and the square function is illustrated in Fig. 1. As can be seen, if the *a priori* errors are small, the two cost functions are of similar change rate, indicating that both of them can yield similar adaption performance. In comparison, the Cauchy estimator deemphasizes the large errors caused by large spectral peaks of speech. This property of the Cauchy estimator can help improve the performance of the frequency-domain adaptive filter with speech excitation. Note that one may use other types of M-estimators, such as the Huber estimator [22]. The advantage of the Cauchy estimator is that it is continuously differentiable. So, it is mathematically easier to derive rigorous robust adaptive filters.

In this work, we use the iterative Newton's method to derive the adaptive algorithm that minimizes the cost function $\mathcal{J}_\rho(m)$ [23]. To do so, we need to calculate the gradient of $\mathcal{J}_\rho(m)$ with respect to $\underline{\widehat{\boldsymbol{h}}}^*_k(m)$ (where the superscript * denotes the complex conjugate) and the corresponding Hessian matrix.

First, the gradient of $\mathcal{J}_\rho(m)$ with respect to $\underline{\widehat{\boldsymbol{h}}}^*_k(m)$ is deduced as follows:

$$\begin{aligned}\nabla\mathcal{J}_\rho(m) &= 2\frac{\partial\mathcal{J}_\rho(m)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)} = 2\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\sum_{n=mL}^{mL+L-1}\frac{\partial\rho\left[|\underline{e}_{ij}(n)|\right]}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\\ &= \sum_{i=1}^{M}\sum_{n=mL}^{mL+L-1}\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^*_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\\ &\quad\times\mathbf{u}_{n-mL+1}\rho'\left[|\underline{e}_{ik}(n)|\right]\exp\{\jmath\arg[\underline{e}_{ik}(n)]\}\\ &= \sum_{i=1}^{M}\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^*_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\underline{\boldsymbol{\varphi}}\left[\underline{\boldsymbol{e}}_{ik}(m)\right],\end{aligned} \tag{10}$$

where $\mathbf{u}_i$ $(i = 1, 2, \ldots, L)$ is the ith column of the identity matrix $\mathbf{I}_{L\times L}$, $\rho'(\cdot)$ is the first-order derivative of $\rho(\cdot)$, $\jmath = \sqrt{-1}$ is the imaginary unit,

$$\boldsymbol{\mathcal{G}}^{10}_{L\times 2L} = \mathbf{F}_{L\times L}\left[\begin{array}{cc}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\end{array}\right]\mathbf{F}^{-1}_{2L\times 2L}, \tag{11}$$

$$\boldsymbol{\mathcal{G}}^{01}_{2L\times L} = \mathbf{F}_{2L\times 2L}\left[\begin{array}{cc}\mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{array}\right]^T\mathbf{F}^{-1}_{L\times L}, \tag{12}$$

$$\underline{\boldsymbol{\varphi}}\left[\underline{\boldsymbol{e}}_{ik}(m)\right] = \left[\begin{array}{c}\rho'\left[|\underline{e}_{ik}(mL)|\right]\exp\{\jmath\arg[\underline{e}_{ik}(mL)]\}\\ \rho'\left[|\underline{e}_{ik}(mL+1)|\right]\exp\{\jmath\arg[\underline{e}_{ik}(mL+1)]\}\\ \vdots\\ \rho'\left[|\underline{e}_{ik}(mL+L-1)|\right]\exp\{\jmath\arg[\underline{e}_{ik}(mL+L-1)]\}\end{array}\right]. \tag{13}$$

The Hessian matrix is then derived as follows:

$$\begin{aligned}\boldsymbol{\mathcal{S}}_k(m) &= 2\frac{\partial}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\left[\nabla\mathcal{J}_\rho(m)\right]^H\\ &= 2\frac{\partial}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\left\{\sum_{i=1,i\neq k}^{M}\underline{\boldsymbol{\varphi}}^H\left[\underline{\boldsymbol{e}}_{ik}(m)\right]\boldsymbol{\mathcal{G}}^{01}_{L\times 2L}\boldsymbol{\mathcal{D}}_{x_i}(m)\boldsymbol{\mathcal{G}}^{10}_{2L\times L}\right\}\\ &= 2\sum_{i=1,i\neq k}^{M}\frac{\partial\underline{\boldsymbol{\varphi}}^H\left[\underline{\boldsymbol{e}}_{ik}(m)\right]}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\boldsymbol{\mathcal{G}}^{01}_{L\times 2L}\boldsymbol{\mathcal{D}}_{x_i}(m)\boldsymbol{\mathcal{G}}^{10}_{2L\times L}.\end{aligned} \tag{14}$$

It can be checked that

$$\begin{aligned}&\frac{\partial}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\left\{\rho'\left[|\underline{e}_{ik}(n)|\right]\exp\{\jmath\arg[\underline{e}_{ik}(n)]\}\right\}^*\\ &= \frac{\partial\left\{\rho'\left[|\underline{e}_{ik}(n)|\right]\exp\{-\jmath\arg[\underline{e}_{ik}(n)]\}\right\}}{\partial\underline{e}^*_{ik}(n)}\cdot\frac{\partial\underline{e}^*_{ik}(n)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\\ &= \left\{\rho''\left[|\underline{e}_{ik}(n)|\right]\frac{\partial|\underline{e}_{ik}(n)|}{\partial\underline{e}^*_{ik}(n)}\exp\{-\jmath\arg[\underline{e}_{ik}(n)]\}\right.\\ &\quad\left.+\rho'\left[|\underline{e}_{ik}(n)|\right]\frac{\partial\exp\{-\jmath\arg[\underline{e}_{ik}(n)]\}}{\partial\underline{e}^*_{ik}(n)}\right\}\frac{\partial\underline{e}^*_{ik}(n)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\\ &= \frac{1}{2}\underline{\eta}_{ik}(n)\frac{\partial\underline{e}^*_{ik}(n)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)},\end{aligned} \tag{15}$$

where $\rho''(\cdot)$ is the second-order derivative of $\rho(\cdot)$,

$$\underline{\eta}_{ik}(n) = \rho''\left[|\underline{e}_{ik}(n)|\right] + \frac{\rho'\left[|\underline{e}_{ik}(n)|\right]}{|\underline{e}_{ik}(n)|}, \tag{16}$$

and

$$\begin{aligned}\frac{\partial\underline{\boldsymbol{\varphi}}^H\left[\underline{\boldsymbol{e}}_{ik}(m)\right]}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)} = \frac{1}{2}\Bigg[&\underline{\eta}_{ik}(mL)\frac{\partial\underline{e}^*_{ik}(mL)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)} \quad \underline{\eta}_{ik}(mL+1)\\ &\times\frac{\partial\underline{e}^*_{ik}(mL+1)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)} \quad \cdots \quad \underline{\eta}_{ik}(mL+L-1)\\ &\times\frac{\partial\underline{e}^*_{ik}(mL+L-1)}{\partial\underline{\widehat{\boldsymbol{h}}}^*_k(m)}\Bigg]\end{aligned}$$

$$= \frac{1}{2}\Big[\underline{\eta}_{ik}(mL)\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\mathbf{u}_1 \quad \underline{\eta}_{ik}(mL+1)\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\mathbf{u}_2 \quad \cdots \quad \underline{\eta}_{ik}(mL+L-1)\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\mathbf{u}_L\Big]$$
$$= \frac{1}{2}\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\boldsymbol{\mathcal{T}}_{ik}(m), \tag{17}$$

where

$$\boldsymbol{\mathcal{T}}_{ik}(m) = \operatorname{diag}\big\{\underline{\eta}_{ik}(mL) \;\; \underline{\eta}_{ik}(mL+1) \;\; \cdots \;\; \underline{\eta}_{ik}(mL+L-1)\big\}. \tag{18}$$

Substituting (17) into (14), we obtain the Hessian matrix as

$$\boldsymbol{\mathcal{S}}_k(m) = \boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{P}}_k(m)\boldsymbol{\mathcal{G}}^{10}_{2L\times L}, \tag{19}$$

where

$$\boldsymbol{\mathcal{P}}_k(m) = \sum_{i=1,i\neq k}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\boldsymbol{\mathcal{T}}_{ik}(m)\boldsymbol{\mathcal{G}}^{01}_{L\times 2L}\boldsymbol{\mathcal{D}}_{x_i}(m). \tag{20}$$

Using Newton's method, we can write the update equations of the channel estimates as

$$\widehat{\underline{\boldsymbol{h}}}_k(m+1) = \widehat{\underline{\boldsymbol{h}}}_k(m) - \mu\boldsymbol{\mathcal{S}}_k^{-1}(m)\nabla\mathcal{J}_\rho(m), \quad k = 1,2,\ldots,M, \tag{21}$$

where μ is the step size. Now, substituting (10) and (19) into (21) and pre-multiplying both sides by $\boldsymbol{\mathcal{G}}^{10}_{2L\times L}$, we then obtain the update equations:

$$\boldsymbol{\mathcal{G}}^{10}_{2L\times L}\widehat{\underline{\boldsymbol{h}}}_k(m+1) = \boldsymbol{\mathcal{G}}^{10}_{2L\times L}\widehat{\underline{\boldsymbol{h}}}_k(m) - \mu\boldsymbol{\mathcal{G}}^{10}_{2L\times L} \times\left[\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{P}}_k(m)\boldsymbol{\mathcal{G}}^{10}_{2L\times L}\right]^{-1} \times\sum_{i=1}^{M}\boldsymbol{\mathcal{G}}^{10}_{L\times 2L}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\underline{\boldsymbol{\varphi}}\left[\underline{\boldsymbol{e}}_{ik}(m)\right], \quad k = 1,2,\ldots,M. \tag{22}$$

After some simple mathematical manipulation, we obtain the simplified update equations:

$$\widehat{\underline{\boldsymbol{h}}}^{10}_k(m+1) = \widehat{\underline{\boldsymbol{h}}}^{10}_k(m) - \mu_{\mathrm{f}}\boldsymbol{\mathcal{P}}_k^{-1}(m)\sum_{i=1}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m) \times\underline{\boldsymbol{\varphi}}^{01}\left[\underline{\boldsymbol{e}}_{ik}(m)\right], k = 1,2,\ldots,M, \tag{23}$$

where $\mu_{\mathrm{f}} = \mu/2$ is a new step size and

$$\widehat{\underline{\boldsymbol{h}}}^{10}_k(m) = \boldsymbol{\mathcal{G}}^{10}_{2L\times L}\widehat{\underline{\boldsymbol{h}}}_k(m), \tag{24}$$

$$\underline{\boldsymbol{\varphi}}^{01}\left[\underline{\boldsymbol{e}}_{ik}(m)\right] = \boldsymbol{\mathcal{G}}^{01}_{2L\times L}\underline{\boldsymbol{\varphi}}\left[\underline{\boldsymbol{e}}_{ik}(m)\right]. \tag{25}$$

To simplify the expression of the matrix $\boldsymbol{\mathcal{P}}_k(m)$ in (20), let us approximate (18) with

$$\boldsymbol{\mathcal{T}}_{ik}(m) = \phi_{ik}(m)\mathbf{I}_{L\times L}, \tag{26}$$

where

$$\phi_{ik}(m) = \max_{0\le l\le L-1}\big\{\underline{\eta}_{ik}(mL+l)\big\}. \tag{27}$$

Note that

$$\boldsymbol{\mathcal{G}}^{01}_{2L\times L}\boldsymbol{\mathcal{T}}_{ik}(m)\boldsymbol{\mathcal{G}}^{01}_{L\times 2L} = \phi_{ik}(m)\boldsymbol{\mathcal{G}}^{01}_{2L\times 2L} \approx \frac{1}{2}\phi_{ik}(m)\mathbf{I}_{2L\times 2L}, \tag{28}$$

where

$$\boldsymbol{\mathcal{G}}^{01}_{2L\times 2L} = \mathbf{F}_{2L\times 2L}\begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}_{2L\times 2L}. \tag{29}$$

It follows then that we can write the matrix $\boldsymbol{\mathcal{P}}_k(m)$ in (20) into a diagonal matrix

$$\boldsymbol{\mathcal{P}}_k(m) = \frac{1}{2}\sum_{i=1,i\neq k}^{M}\phi_{ik}(m)\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\boldsymbol{\mathcal{D}}_{x_i}(m). \tag{30}$$

This simplification would considerably reduce the complexity for computing the inverse of the matrix $\boldsymbol{\mathcal{P}}_k(m)$. In implementation, a more smoothed power spectrum matrix $\boldsymbol{\mathcal{P}}_k(m)$ can be obtained by the widely used recursive method.

Same as in [19], [24], a spectral constraint on the channel impulse responses is introduced into the above algorithm to improve its robustness with noise and reverberation. Then, the final frequency-domain adaptive filter algorithm is as follows:

$$\widehat{\underline{\boldsymbol{h}}}^{10}_k(m+1) = \widehat{\underline{\boldsymbol{h}}}^{10}_k(m) - \mu_{\mathrm{f}}\nabla\mathcal{J}^{01}_{\mathrm{NFM},k}(m) + \mu_{\mathrm{f}}\beta(m) \times\nabla\mathcal{J}^{10}_{\mathrm{SC},k}(m), \quad k = 1,2,\ldots,M, \tag{31}$$

where

$$\nabla\mathcal{J}^{01}_{\mathrm{NFM},k}(m) = \boldsymbol{\mathcal{P}}_k^{-1}(m)\sum_{i=1}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m)\underline{\boldsymbol{\varphi}}^{01}\left[\underline{\boldsymbol{e}}_{ik}(m)\right], \tag{32}$$

$$\nabla\mathcal{J}^{10}_{\mathrm{SC},k}(m) = 2\widehat{\underline{\boldsymbol{h}}}^{10}_k(m)\oslash\left(\mathbf{1}_{2L\times 1} + \left|\widehat{\underline{\boldsymbol{h}}}^{10}_k(m)\right|^2\right), \tag{33}$$

$\beta(m)$ is the Lagrange multiplier, $\oslash$ denotes element-by-element division of two vectors, and $\mathbf{1}_{2L\times 1}$ is a vector of length $2L$ with all the elements being 1.

2.2. TDE Based on the Robust Adaptive Blind Multichannel Identification

After the channel impulse responses are adaptively estimated by the aforementioned blind multichannel identification algorithm, we can then determine the TDOAs by comparing the time differences of the direct-path components between different channels. The TDOA between two different channels can then be obtained as [1], [25]

$$\widehat{\tau}_{ij} = \arg\max_l\,|\widehat{h}_{j,l}| - \arg\max_l\,|\widehat{h}_{i,l}|. \tag{34}$$

3. EXPERIMENTS

In this section, we study the performance of the proposed algorithm in noisy and reverberant acoustic environments. For the purpose of comparison, the performance of the phase transform (PHAT) [1], normalized multichannel frequency-domain least-mean-square (NMCFLMS) [8], robust normalized multichannel frequency-domain least-mean-square (RNMCFLMS) [24], and robust normalized multichannel frequency-domain least-mean-M-estimate (RNMCFLMM) [26] algorithms will also be presented.

3.1. Experimental Setup

The impulses responses used in this study were made in the Varechoic Chamber at Bell Labs [27]. The dimension of the Chamber is 6.7 m×6.1 m×2.9 m. For convenience, positions in the room are designated by (x, y, z) coordinates with reference to the northwest corner of the Chamber floor. We select three microphones from the

measuring system in [27] to construct our linear microphone array system. The three microphones are located at (2.437, 0.500, 1.400), (3.137, 0.500, 1.400), and (3.837, 0.500, 1.400), respectively. A sound source is placed at (0.337, 3.938, 1.600). The impulse responses of the acoustic channels between the source and microphones were measured at a 48 kHz sampling rate when 89% panels on the Chamber's walls were open (the corresponding reverberation time is 280 ms). Then the obtained channel impulse responses are downsampled to a 16 kHz sampling rate and truncated to 1024 samples. The measured impulse responses are treated as the true impulse responses in our blind multichannel identification experiments.

The source signal is pre-recorded from a male and a female speakers. The sampling rate is 16 kHz and the overall length is approximately 2 min: the former half is from the male speaker while the latter half is from the female speaker. The multichannel system outputs are computed by convolving the source signal with the corresponding measured channel impulse responses and noise is then added to the results at a specified signal-to-noise ratio (SNR) value. The additive noise used in this work is white Gaussian noise. All the parameters are set to be the same as those experiments in [19]. For the proposed algorithm, the length of the adaptive filter is 1024, and the parameter c is set to 5.0.

In the experiments, an estimate is yielded every frame with a frame size of 64 ms (1024 samples). The total number of frames is 1886. Two performance metrics, namely the probability of anomalous estimates and the root mean-squared error (RMSE) of nonanomalous estimates [10], [12], are used to evaluate the performance of the proposed algorithm. The true time delays from the sound source to the three microphone pairs are respectively $\tau_{12} = 19$ samples, $\tau_{13} = 42$ samples, and $\tau_{23} = 23$ samples.

3.2. Results

The TDE results of the five studied algorithms are presented in Table 1. As seen, the NMCFLMS algorithm performs better than the PHAT algorithm. The RNMCFLMS algorithm is more robust to moderate noise than the PHAT and NMCFLMS algorithms due to the use of a spectral energy constraint; but it suffers from significant performance degradation when the noise is strong. The RNMCFLMM algorithm almost outperforms the previous three algorithms mainly thanks to the use of the Huber estimator and the alternate employment of the mean-squared error (MSE) and mean-absolute error (MAE) criteria in the time-domain Huber estimator [19]. Among the five studied TDE algorithms, the proposed algorithm obtains the best performance, especially in the environments with low SNRs. This enhancement comes from the fact that the global frequency-domain adaptive filter uses the frequency-domain Cauchy estimator, which is robust to deal with the speech excitation signals with impulsive spectra. This new algorithm can be viewed as an improved version of the RNMCFLMS and RNMCFLMM algorithms.

4. CONCLUSIONS

In this paper, we proposed a global frequency-domain adaptive filter algorithm for TDE in acoustic SIMO systems. The Cauchy estimator is used to define a frequency-domain cost function, from which a robust frequency-domain adaptive filter is derived to blindly identify an acoustic SIMO system. This Cauchy estimator is insensitive to the impulse-like structure of speech spectra while retains the approximate adaption ability of the square cost function if the spectra of the excitation signals are flat. Moreover, the Cauchy estimator is continuously differentiable as compared to the Huber estimator, which yields a mathematically rigorous adaptive filter. Experiments conducted in noisy and reverberant environments validate the robustness of the developed TDE approach.

Table 1. The probability of anomalous time delay estimates and RMSE of nonanomalous time delay estimates of the five studied TDE algorithms under different levels of the SNR.

SNR (dB)	TDE algorithms	$\%_{\text{anomalies}}(\%)$			RMSE (samples)		
		$\widehat{\tau}_{12}$	$\widehat{\tau}_{13}$	$\widehat{\tau}_{23}$	$\widehat{\tau}_{12}$	$\widehat{\tau}_{13}$	$\widehat{\tau}_{23}$
-10	PHAT	96.2	96.8	96.0	1.3	1.2	1.3
	NMCFLMS	71.5	86.3	94.1	0.9	0.9	0.9
	RNMCFLMS	62.1	84.2	75.3	1.4	1.5	0.1
	RNMCFLMM	16.8	66.3	81.0	1.5	0.5	0.8
	Proposed	3.7	5.9	8.1	0.7	0.2	0.7
-5	PHAT	90.0	92.0	92.1	1.0	1.2	1.3
	NMCFLMS	10.6	9.3	59.6	0.5	0.6	0.7
	RNMCFLMS	11.1	1.4	44.3	1.0	0.5	0.9
	RNMCFLMM	1.1	1.6	22.8	0.1	0.9	0.1
	Proposed	1.6	1.0	1.6	0.1	0.2	0.2
0	PHAT	68.7	77.8	81.6	0.9	0.9	1.2
	NMCFLMS	1.2	1.1	2.0	0.5	0.6	0.6
	RNMCFLMS	0.1	1.6	0.7	0.6	1.2	1.2
	RNMCFLMM	0.3	0.4	0.4	0.3	0.7	1.0
	Proposed	0.2	0.6	0.7	0.3	0.3	0.1
5	PHAT	44.6	56.2	69.7	0.7	0.8	1.1
	NMCFLMS	0.5	0.6	0.5	0.1	0.3	0.3
	RNMCFLMS	0.0	0.1	0.1	0.0	0.5	0.6
	RNMCFLMM	0.2	0.3	0.2	0.1	0.4	0.1
	Proposed	0.0	0.2	0.1	0.0	0.5	0.1
10	PHAT	29.4	39.4	60.7	0.5	0.7	0.9
	NMCFLMS	0.4	0.5	0.6	0.2	0.1	0.2
	RNMCFLMS	0.2	0.0	0.1	0.1	0.1	0.8
	RNMCFLMM	0.1	0.1	0.2	0.1	0.1	0.4
	Proposed	0.0	0.0	0.0	0.0	0.2	0.1

5. RELATION TO PRIOR WORK

TDE has attracted a significant amount of attention in the literature [1], [2]. Many methods for TDE have been developed, including the well-known generalized cross-correlation (GCC) method [3], [4], the blind channel identification based approach [5]–[9], multichannel linear prediction algorithm [10]–[12], the information theory based methods [13]–[15], etc. Among those methods, the blind multichannel identification approach based on the NMCFLMS algorithm is very attractive for single source TDE [8], [18]. The underlying core idea is that the channel impulse response from the source to each microphone is first blindly estimated, and the time delays are then determined by comparing the time differences of the direct-path components between different channels [1], [8], [18]. This algorithm is robust to reverberation since reverberation is well modeled in the algorithmic formulation; but it is found sensitive to noise. It was extended to an RNMCFLMS method by introducing a flatness constraint on the channel transfer functions [24], [26]; but the robustness with respect to noise is still a great challenge particularly when SNR is low. In an early work, we developed an RNMCFLMM algorithm [19], where a Huber estimator [22] was used to construct a robust time-domain cost function, from which we obtain a multichannel frequency-domain adaptive filter to blindly identify a SIMO system. The RNMCFLMM algorithm is more robust to both non-Gaussian and Gaussian noise than RNMCFLMS [26]. However, its performance suffers from degradations if the excitation signals are speech. To improve performance in noisy and reverberant environments with speech excitation signals, we followed the good properties in the RNMCFLMS and RNMCFLMM algorithms and meanwhile adopted a Cauchy estimator to define a frequency-domain cost function, which deemphasizes the large errors caused by large spectral peaks of speech. From this new cost function, we developed a robust adaptive multichannel algorithm to blindly identify acoustic SIMO systems from which a multichannel TDE algorithm is obtained.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] K. Kowalczyk, E. Habets, W. Kellermann, and P. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Process. Lett.*, vol. 20, pp. 653–656, Jul. 2013.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.

[17] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no.12, pp. 2982–2993, Dec. 1995.

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

[19] H. He, J. Lu, J. Chen, X. Qiu, and J. Benesty, "Robust blind identification of room acoustic channels in symmetric alpha-stable distributed noise environments," *J. Acoust. Soc. Amer.*, vol. 136, no. 8, pp. 693–704, Aug. 2014.

[20] S. Makino, S. Araki, and H. Sawada, "Underdetermined blind source separation using acoustic arrays,"in *Handbook on array processing and sensor networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, USA: John Wiley & Sons, 2010.

[21] Z. Zhang, "Parameter estimation techniques: A tutorial with application on conic fitting," *Image Vis. Computing*, vol. 15, no. 1, pp. 59–76, 1997.

[22] P. J. Huber, *Robust Statistics*. New York, NY: Wiley, 1981.

[23] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, "Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1340–1351, May 2007.

[24] M. A. Haque and M. K. Hasan, "Noise robust multichannel frequency-domain LMS algorithms for blind channel identification," *IEEE Signal Process. Lett.*, vol. 15, pp. 305–308, 2008.

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

[26] H. He, J. Chen, J. Benesty, and T. Yang, "Multichannel time delay estimation for acoustic source localization via robust adaptive blind system identification," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC)*, 2016.

[27] A. Härmä, "Acoustic measurement data from the varechoic chamber," *Tech. Memorandum* 110101, Agere Systems, Allentown, PA, USA, 2001.