AN ADAPTIVE BLIND SIMO IDENTIFICATION APPROACH TO JOINT MULTICHANNEL TIME DELAY ESTIMATION

Jingdong Chen, Yiteng (Arden) Huang

Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, New Jersey 07974, USA
{jingdong, arden}@research.bell-labs.com

Jacob Benesty

Université du Québec, INRS-EMT
800 de la Gauchetière Ouest, Suite 6900
Montréal, Québec, H5A 1K6, Canada
benesty@inrs-emt.uquebec.ca

ABSTRACT

Time delay estimation (TDE) is a difficult problem in a reverberant environment and the traditional generalized cross-correlation (GCC) methods perform poorly. The adaptive eigenvalue decomposition (AED) algorithm recently proposed by the authors exploits blind channel identification (BCI) technique and deals with room reverberation more effectively. The AED algorithm was developed for a two-channel system. It requires that the two channels do not share any common zeros (a necessary condition of system identifiability). In this paper we generalize the AED algorihm to multichannel (more than 2) systems. Compared to the AED algorithm, the generalized method is more robust since it is less likely for all channels to share a common zero when more sensors are used.

1. INTRODUCTION

In many multimedia communication and voice processing systems, the knowledge of sound source location is essential for steering a video camera to facilitate interactive collaboration [1] and for applying advanced array signal processing technologies to acquire high-fidelity speech [2]. After two decades of continuous research, the time delay estimation (TDE) based approach to acoustic source localization has become the technique of choice for implementations of these applications, especially in recent digital systems. A robust TDE method for wideband acoustic signals is apparently the cornerstone of the success of these systems.

Traditionally, time difference of arrival (TDOA) is estimated in light of the cross-correlation function between two received signals. The generalized cross-correlation (GCC) method [3] is the most popular technique for TDE mainly because of its computational simplicity and negligible delay in processing, but is sensitive to noise and room reverberation in practice. Some amendments to the GCC method have been proposed. But unfortunately, the cross-correlation-based algorithms fundamentally cannot cope well with reverberation since they all assume an unrealistic single-path propagation channel model without taking into account the mutlipath effect. To the best of our knowledge, all known TDE methods fail in a highly reverberant room.

In this paper, we consider a real-reverberant model for room acoustics. Suppose that M microphones are used to capture signals $x_i(n), i = 1, 2, \cdots, M$, propagating from a single *unknown* source $s(n)$, which leads to a single-input multiple-output (SIMO) system as illustrated in Fig. 1:

$$x_i(n) = h_{\mathrm{t},i} * s(n) + b_i(n), \quad i = 1, 2, \cdots, M, \tag{1}$$

where $h_{\mathrm{t},i}$ is the true (subscript t) impulse response of the i-th channel and $b_i(n)$ is the additive background noise at the i-th microphone. In vector form, (1) can be expressed as:

$$\mathbf{x}_i(n) = \mathbf{H}_{\mathrm{t},i} \cdot \mathbf{s}(n) + \mathbf{b}_i(n), \tag{2}$$

where

$$\begin{aligned}
\mathbf{x}_i(n) &= [x_i(n) \;\; x_i(n-1) \;\; \cdots \;\; x_i(n-L+1)]^T, \\
\mathbf{H}_{\mathrm{t},i} &= \begin{bmatrix} h_{\mathrm{t},i,0} & \cdots & h_{\mathrm{t},i,L-1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & h_{\mathrm{t},i,0} & \cdots & h_{\mathrm{t},i,L-1} \end{bmatrix}, \\
\mathbf{s}(n) &= [s(n) \;\; s(n-1) \;\; \cdots \;\; s(n-2L+2)]^T, \\
\mathbf{b}_i(n) &= [b_i(n) \;\; b_i(n-1) \;\; \cdots \;\; b_i(n-L+1)]^T,
\end{aligned}$$

L is set to the length of the longest channel impulse response by assumption, and $(\cdot)^T$ denotes vector/matrix transpose. The channel parameter matrix $\mathbf{H}_{\mathrm{t},i}$ is of dimension $L \times (2L-1)$ and is constructed from:

$$\mathbf{h}_{\mathrm{t},i} = [h_{\mathrm{t},i,0} \;\; h_{\mathrm{t},i,1} \;\; \cdots \;\; h_{\mathrm{t},i,L-1}]^T. \tag{3}$$

In an earlier study [4], we tackled the TDE problem from a different point of view and proposed the adaptive eigenvalue decomposition (AED) algorithm based on the blind channel identification (BCI) technique. For an identifiable SIMO system, the following two conditions need to be met [5]:

1. The polynomials formed from $\mathbf{h}_{\mathrm{t},i}$ are co-prime, i.e., the channel transfer functions $H_{\mathrm{t},i}(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\left\{\mathbf{s}(n)\mathbf{s}^T(n)\right\}$ of the source signal is of full rank, where $E\{\cdot\}$ stands for mathematical expectation.

The AED algorithm performs blind identification of only two channels at a time. For such a single-input two-output system, the zeros of the two channels can be close to each other if not common, especially when their impulse responses are long. This leads to an ill-conditioned system that is difficult to identify. This problem can be alleviated by using more microphones in the system. When more channels are involved, it is less likely for all channels to share a common zero and therefore the BCI algorithm can perform more robustly. As such, when a microphone array is used for locating a sound source, the relative TDOAs are no longer estimated pair by

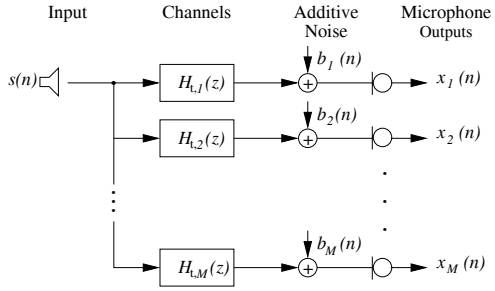


Figure 1: Illustration of the relationships between the input source $s(n)$ and the microphone outputs $x_i(n)$, $i = 1, 2, \cdots, M$, in a single-input multiple-output FIR system.

pair. The multichannel system will be treated as a whole and the BCI algorithm can be globally optimized.

While using more microphones can make a system more feasible to identify, more parameters should be estimated at a time and the BCI algorithm is more complicate to develop. To make the estimation process efficient, in this paper, we will apply the frequency-domain adaptive BCI method [6] and will propose a generalized joint multichannel (JMC) TDE algorithm.

2. THE FREQUENCY-DOMAIN BCI ALGORITHM

Basically, a multichannel system can be blindly identified because of the channel diversity, which makes the outputs of different channels distinct though related. By following the fact that

$$x_i(n) * h_{\mathrm{t},j} = s(n) * h_{\mathrm{t},i} * h_{\mathrm{t},j} = x_j(n) * h_{\mathrm{t},i}, \qquad (4)$$

a cross-relation between the i-th and j-th channel outputs, in the absence of noise, can be formulated as

$$\mathbf{x}_i^T(n)\mathbf{h}_{\mathrm{t},j} = \mathbf{x}_j^T(n)\mathbf{h}_{\mathrm{t},i}, \;\; i, j = 1, 2, ..., M, \; i \neq j. \qquad (5)$$

When noise is present or the channel impulse responses are improperly modeled, the left and right hand sides of (5) are generally not equal and the inequality can be used to define an *a priori* error signal as follows:

$$e_{ij}(n+1) = \frac{\mathbf{x}_i^T(n+1)\mathbf{h}_j(n) - \mathbf{x}_j^T(n+1)\mathbf{h}_i(n)}{\|\mathbf{h}(n)\|}, \qquad (6)$$

where $\mathbf{h}_i(n)$ is the model filter for the i-th channel at time n and

$$\mathbf{h}(n) = \left[\begin{array}{cccc} \mathbf{h}_1^T(n) & \mathbf{h}_2^T(n) & \cdots & \mathbf{h}_M^T(n) \end{array} \right]^T.$$

The model filter is normalized in order to avoid a trivial solution whose elements are all zeros. Based on the error signal defined here, a cost function at time $n+1$ is given by

$$J(n+1) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} e_{ij}^2(n+1). \qquad (7)$$

An adaptive algorithm is then derived to efficiently determine the model filters $\mathbf{h}_i$ that minimize this cost function and therefore would be good estimates of $\mathbf{h}_{\mathrm{t},i}/\|\mathbf{h}_{\mathrm{t}}\|$. This can be done in either the time or the frequency domain. In the following, we present a frequency-domain algorithm that is more computationally efficient and converges much faster.

To begin, we define an intermediate signal $y_{ij} \triangleq x_i * h_j$. In vector form, a block of such a signal can be expressed in the frequency domain as

$$\underline{\boldsymbol{y}}_{ij}(m+1) = \boldsymbol{\mathcal{W}}_{L\times 2L}^{01} \boldsymbol{\mathcal{D}}_{x_i}(m+1) \boldsymbol{\mathcal{W}}_{2L\times L}^{10} \underline{\boldsymbol{h}}_j(m), \qquad (8)$$

where

$$\begin{aligned} \boldsymbol{\mathcal{W}}_{L\times 2L}^{01} &= \mathbf{F}_{L\times L} \left[\begin{array}{cc} \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L} \end{array} \right] \mathbf{F}_{2L\times 2L}^{-1}, \\ \boldsymbol{\mathcal{D}}_{x_i}(m+1) &= \mathrm{diag}\left\{ \mathbf{F}_{2L\times 2L} \cdot \mathbf{x}_i(m+1)_{2L\times 1} \right\}, \\ \boldsymbol{\mathcal{W}}_{2L\times L}^{10} &= \mathbf{F}_{2L\times 2L} \left[\begin{array}{cc} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times L} \end{array} \right]^T \mathbf{F}_{L\times L}^{-1}, \\ \underline{\boldsymbol{h}}_j(m) &= \mathbf{F}_{L\times L} \mathbf{h}_j(m), \\ \mathbf{x}_i(m+1)_{2L\times 1} &= \left[\begin{array}{cccc} x_i(mL) & x_i(mL+1) & \cdots & \\ & & & x_i(mL+2L-1) \end{array} \right]^T, \end{aligned} \qquad (9)$$

$\mathbf{F}_{L\times L}$ and $\mathbf{F}_{L\times L}^{-1}$ are respectively the Fourier and inverse Fourier matrices of size $L \times L$, and m is the block time index. Then a block of the error signal based on the cross-relation between the i-th and the j-th channel in the frequency domain is determined as:

$$\begin{aligned} \underline{\boldsymbol{e}}_{ij}(m+1) &= \underline{\boldsymbol{y}}_{ij}(m+1) - \underline{\boldsymbol{y}}_{ji}(m+1) \\ &= \boldsymbol{\mathcal{W}}_{L\times 2L}^{01} \left[\boldsymbol{\mathcal{D}}_{x_i}(m+1) \boldsymbol{\mathcal{W}}_{2L\times L}^{10} \underline{\boldsymbol{h}}_j(m) - \right. \\ &\qquad \left. \boldsymbol{\mathcal{D}}_{x_j}(m+1) \boldsymbol{\mathcal{W}}_{2L\times L}^{10} \underline{\boldsymbol{h}}_i(m) \right]. \end{aligned} \qquad (10)$$

Continuing, we construct a frequency-domain cost function at the $(m+1)$-th block as follows:

$$J_{\mathrm{f}}(m+1) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \underline{\boldsymbol{e}}_{ij}^H(m+1) \underline{\boldsymbol{e}}_{ij}(m+1), \qquad (11)$$

where $(\cdot)^H$ denotes Hermitian transpose. By using Newton's method, we update the model filter coefficients according to:

$$\begin{aligned} \underline{\boldsymbol{h}}_k(m+1) = \underline{\boldsymbol{h}}_k(m) - \\ \mu_{\mathrm{f}} E^{-1} \left\{ \frac{\partial}{\partial \underline{\boldsymbol{h}}_k^T(m)} \left[\frac{\partial J_{\mathrm{f}}(m+1)}{\partial \underline{\boldsymbol{h}}_k^*(m)} \right] \right\} \frac{\partial J_{\mathrm{f}}(m+1)}{\partial \underline{\boldsymbol{h}}_k^*(m)}, \end{aligned} \qquad (12)$$

where $(\cdot)^*$ stands for complex conjugate and μ_{f} is a small positive step size. It can be shown that

$$\begin{aligned} &\frac{\partial J_{\mathrm{f}}(m+1)}{\partial \underline{\boldsymbol{h}}_k^*(m)} = \\ &\qquad \sum_{i=1}^{M} \left[\boldsymbol{\mathcal{W}}_{L\times 2L}^{01} \boldsymbol{\mathcal{D}}_{x_i}(m+1) \boldsymbol{\mathcal{W}}_{2L\times L}^{10} \right]^H \underline{\boldsymbol{e}}_{ik}(m+1), \end{aligned} \qquad (13)$$

and the Hessian matrix can be approximated as

$$\begin{aligned} &E \left\{ \frac{\partial}{\partial \underline{\boldsymbol{h}}_k^T(m)} \left[\frac{\partial J_{\mathrm{f}}(m+1)}{\partial \underline{\boldsymbol{h}}_k^*(m)} \right] \right\} \\ &\quad \approx \frac{1}{2} \boldsymbol{\mathcal{W}}_{L\times 2L}^{10} \boldsymbol{\mathcal{P}}_k(m+1) \boldsymbol{\mathcal{W}}_{2L\times L}^{10}, \end{aligned} \qquad (14)$$

where

$$\boldsymbol{\mathcal{P}}_k(m+1) = E \left\{ \sum_{i=1, i\neq k}^{M} \boldsymbol{\mathcal{D}}_{x_i}^*(m+1) \boldsymbol{\mathcal{D}}_{x_i}(m+1) \right\}.$$

Substituting (13) and (14) into (12) and multiplying by $\boldsymbol{\mathcal{W}}^{10}_{2L\times L}$ produces the *constrained* frequency-domain BCI algorithm:

$$\underline{\boldsymbol{h}}^{10}_k(m+1) = \underline{\boldsymbol{h}}^{10}_k(m) - 2\mu_{\rm f}\boldsymbol{\mathcal{W}}^{10}_{2L\times 2L}\boldsymbol{\mathcal{P}}^{-1}_k(m+1)\sum_{i=1}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m+1)\underline{\boldsymbol{e}}^{01}_{ik}(m+1), \quad (15)$$

where

$$\begin{aligned}\underline{\boldsymbol{h}}^{10}_k(m) &= \boldsymbol{\mathcal{W}}^{10}_{2L\times L}\underline{\boldsymbol{h}}_k(m),\\ \underline{\boldsymbol{e}}^{01}_{ik}(m+1) &= \boldsymbol{\mathcal{W}}^{01}_{2L\times L}\underline{\boldsymbol{e}}_{ik}(m+1),\\ \boldsymbol{\mathcal{W}}^{10}_{2L\times 2L} &= \boldsymbol{\mathcal{W}}^{10}_{2L\times L}\boldsymbol{\mathcal{W}}^{10}_{L\times 2L}.\end{aligned}$$

By approximating $2\boldsymbol{\mathcal{W}}^{10}_{2L\times 2L}$ by the identity matrix, we deduce the *unconstrained* frequency-domain BCI algorithm:

$$\underline{\boldsymbol{h}}^{10}_k(m+1) = \underline{\boldsymbol{h}}^{10}_k(m) - \mu_{\rm f}\boldsymbol{\mathcal{P}}^{-1}_k(m+1)\sum_{i=1}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m+1)\underline{\boldsymbol{e}}^{01}_{ik}(m+1), \quad (16)$$

where $\boldsymbol{\mathcal{P}}_k(m+1)$ is diagonal and easy to be inverted. Note that the unit-norm constraint will be enforced on the model filter coefficients after every step of update.

3. JOINT MULTICHANNEL TDE

When the channel impulse responses are long as in the multichannel acoustic systems of interest, blind identification is not easy. For the adaptive algorithms developed in this paper, it takes a long time to determine the filter coefficients in reverberant paths. However, in the application of time delay estimation, the goal is not to accurately estimate the system impulse responses. As long as the direct path of each channel is located, the time delay can be found and the problem is successfully solved. Even though the proposed adaptive algorithms would converge to the desired system impulse responses with an arbitrary initialization, deliberately selected initial model filter coefficients will make the direct path in each channel become dominant earlier during adaption. In this paper, we place a peak at tap $L/2$ of the first channel and initialize all other model filter coefficients to zeros, i.e.

$$\begin{aligned}\mathbf{h}_1 &= \Big[\underbrace{0\ \cdots\ 0}_{L/2-1}\ \ 1\ \ \underbrace{0\ \cdots\ 0}_{L/2}\Big]^T,\\ \mathbf{h}_i &= \mathbf{0},\ \ i=2,3,\cdots,M.\end{aligned} \quad (17)$$

From the view point of system identification, stationary white noise would be a good source signal to fully excite the system's impulse responses. However, in time delay estimation for acoustic source localization, the source signal is speech, which is neither white nor stationary. The power spectrum of the multichannel outputs changes considerably with time. Therefore a recursive scheme is employed for a more stable estimate of power spectrum:

$$\boldsymbol{\mathcal{P}}_k(m+1) = \lambda\boldsymbol{\mathcal{P}}_k(m) + (1-\lambda)\sum_{i=1,i\neq k}^{M}\boldsymbol{\mathcal{D}}^{*}_{x_i}(m+1)\boldsymbol{\mathcal{D}}_{x_i}(m+1), \quad (18)$$
$$k=1,2,\cdots,M,$$

where λ is a forgetting factor set as $\lambda = [1-1/(3L)]^L$. In addition, a small positive number δ is inserted into the normalization to avoid a numerically small denominator during silent period.

After a multichannel system is blindly identified, the direct path can be determined by examining the channel's impulse response. For a multi-path channel, reverberation components are usually weaker than the signal component propagating through the direct path. This is particularly true for an acoustic channel where waveform energy would be absorbed by room surfaces and waveform magnitude would be attenuated by wall reflection. However, this doesn't imply that the tap corresponding to the direct path is always dominant in the channel's impulse response. When two or more reverberant signals via multiple paths happen to arrive at the microphone at the same time, the component of that particular delay might have a larger magnitude. Therefore, the channel propagation delay τ_i in samples can be more robustly determined as the smallest TDOA of the Q largest components in the channel's impulse response:

$$\hat{\tau}_i = \min\left\{\operatorname{argmax}^q_l |h_{i,l}|\right\},\ \ q=1,2,\cdots,Q. \quad (19)$$

where $\max^q$ computes the q-th largest element. The relative TDOA between the i-th and the j-th channel is then obtained as:

$$\hat{\tau}_{ij} = \hat{\tau}_i - \hat{\tau}_j,\ \ i,j=1,2,\cdots,M. \quad (20)$$

4. SIMULATIONS

To evaluate the performance of the proposed algorithm, we carried out a number of experiments for multichannel TDE in an actual reverberant room of dimensions 16.6 ft long by 11.2 ft wide by 8 ft long. For comparison, the phase transform (PHAT) [3] and the AED algorithms are also studied. Two speech sources were used, one male and the other female. The equally-spaced linear array that consists of four omni-directional microphones was mounted 6 ft above the floor. The spacing between adjacent microphones is 1.6 ft. The room geometry as well as the positions of the sources and the microphones are illustrated in Fig. 2. The speech signals as shown in Fig. 3 were sampled at 16 kHz and of duration about 42 seconds.

The results are enumerated in Table 1 and the result for $T_{60} = 623$ ms is also ploted in Fig. 4. For a multichannel system, a TDE algorithm is not satisfactory if it is not accurate over all sensor pairs. Therefore we present the percentage of successful overall TDOA estimates in addition to that of successful individual TDOA estimates. A set of TDOA estimates for all sensor pairs at a given time is deemed successful if all of the individual estimates are successful. It is clear that all algorithms perform well when room reverberation is low. But when room reverberation is significant, the proposed JMC method is more robust than the PHAT and AED algorithms.

5. CONCLUSIONS

Channel diversities including relative time delay of arrival for a single-input multiple-output system can be easily determined after all channel impulse responses are found. Adaptive blind channel identification techniques can be used for the problem of time delay estimation and it is demonstrated that they can more effectively

T_{60} (ms)	TDE	Percent Successful Estimate (%)			
		τ_{12}	τ_{13}	τ_{14}	Overall
370	PHAT	96.87	97.04	97.04	96.87
	AED	95.83	95.83	95.83	95.83
	JMC	97.22	97.22	97.22	97.22
533	PHAT	96.87	87.83	96.87	87.83
	AED	95.30	94.96	95.65	94.78
	JMC	97.04	97.04	97.04	96.87
623	PHAT	92.70	78.96	96.35	76.35
	AED	90.26	85.39	94.78	82.26
	JMC	96.35	95.48	96.87	94.78

Table 1: Experimental results evaluating the accuracy of the three investigated TDE algorithms in a room with different amount of reverberation.

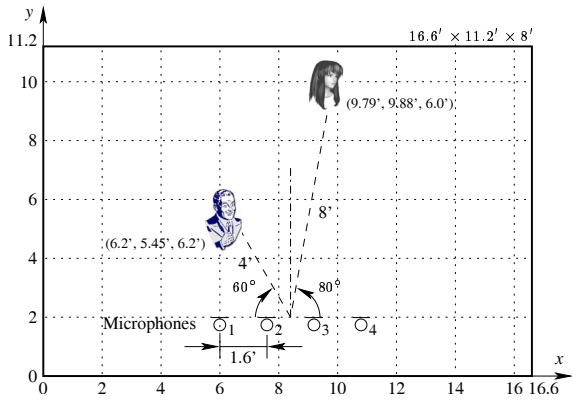


Figure 2: Room geometry (coordinate units in ft).

deal with room reverberation than traditional generalized cross-correlation methods. In this paper, the blind channel identification-based approach was generalized from a two-channel system to a multichannel (greater than 2) system and a joint multichannel TDE algorithm was proposed. The experimental results showed some promise of the robustness of the proposed algorithm in reverberant environments.

6. REFERENCES

[1] Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds., Boston, MA: Kluwer Academic, 2000.

[2] J. L. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, Jan. 1993.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[4] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, vol. 2, pp. 937–940.

[5] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identifi cation," *IEEE Trans. Signal Processing*, vol. 43, pp. 2982–2993, Dec. 1995.

[6] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multi-channel identifi cation," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11-24, Jan. 2003.

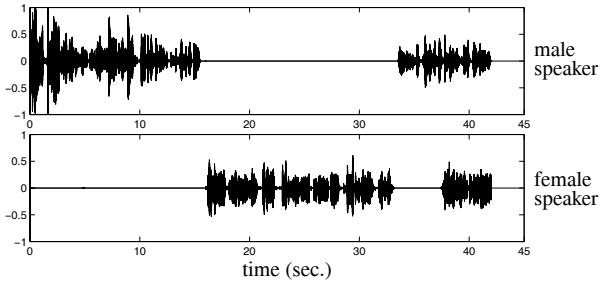
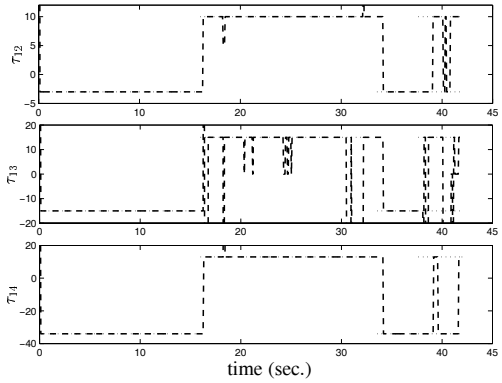
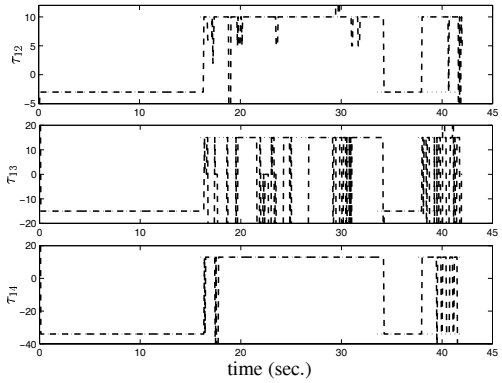


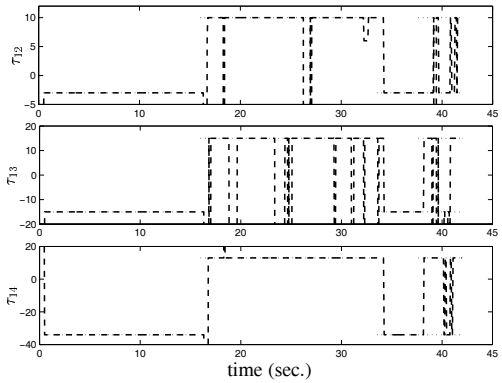
Figure 3: Speech signals used in the simulation.



(a) Joint Multichannel TDE



(b) Phase Transform



(c) Adaptive Eigenvalue Decomposition

Figure 4: Comparison of performance among the three investigated algorithms for $T_{60} = 623$ ms.