EURASIP Journal on Applied Signal Processing 2003:11, 1110–1124

Robust Adaptive Time Delay Estimation for Speaker Localization in Noisy and Reverberant Acoustic Environments

Simon Doclo

Department of Electrical Engineering, Katholieke Universiteit Leuven, ESAT-SISTA, Kasteelpark Arenberg 10,
B-3001 Heverlee, Belgium
Email: simon.doclo@esat.kuleuven.ac.be

Marc Moonen

Department of Electrical Engineering, Katholieke Universiteit Leuven, ESAT-SISTA, Kasteelpark Arenberg 10,
B-3001 Heverlee, Belgium
Email: marc.moonen@esat.kuleuven.ac.be

Received 23 September 2002 and in revised form 2 June 2003

Two adaptive algorithms are presented for robust time delay estimation (TDE) in acoustic environments with a large amount of background noise and reverberation. Recently, an adaptive eigenvalue decomposition (EVD) algorithm has been developed for TDE in highly reverberant acoustic environments. In this paper, we extend the adaptive EVD algorithm to noisy and reverberant acoustic environments, by deriving an adaptive stochastic gradient algorithm for the generalized eigenvalue decomposition (GEVD) or by prewhitening the noisy microphone signals. We have performed simulations using a localized and a diffuse noise source for several SNRs, showing that the time delays can be estimated more accurately using the adaptive GEVD algorithm than using the adaptive EVD algorithm. In addition, we have analyzed the sensitivity of the adaptive GEVD algorithm with respect to the accuracy of the noise correlation matrix estimate, showing that its performance may be quite sensitive, especially for low SNR scenarios.

Keywords and phrases: time delay estimation, acoustic source localization, generalized eigenvalue decomposition, stochastic gradient.

1. INTRODUCTION

In many speech communication applications, such as teleconferencing, hand-free voice-controlled systems, and hearing aids, it is desirable to localize the dominant speaker. By using a microphone array, it is possible to determine the *position* of this speaker such that the microphone array can be electronically steered using a fixed (or adaptive) beamformer in order to provide spatially selective speech acquisition [1, 2]. In multimedia teleconferencing systems, the position of the speaker can be used not only for microphone array beamforming, but also for automatic video camera steering [3, 4] and for determining binaural cues for stereo imaging.

It has been shown that it is possible to calculate the position of a speaker from the *time delays* between the different microphone signals, for example, using maximum likelihood or least-squares methods [5, 6]. However, accurate estimation of the time delays between the different microphone signals is not an easy task because of the room reverberation, the acoustic background noise, and the nonstationary character of the speech signal. Generally, room reverberation is considered to be the main problem for time delay estimation (TDE) [7], but acoustic background noise can also considerably decrease the performance of TDE algorithms. Whereas highly noisy situations are not very common in typical teleconferencing applications, they frequently occur in, for example, hearing aid applications.

Most TDE algorithms are based on the generalized cross-correlation (GCC) or the cross-power spectrum phase (CSP) between the microphone signals [8, 9]. However, since most of these methods assume an ideal room model without reverberation, that is, only a direct path between the signal source and the microphone array, they cannot handle reverberation well. In order to make TDE more robust to room reverberation, a cepstral prefiltering technique has been proposed [10] and there have been developed techniques which use a more realistic room model incorporating reverberation [11, 12]. In [12], an adaptive eigenvalue decomposition

(EVD) algorithm has been developed for (partial) estimation of two *acoustic impulse responses* using a stochastic gradient algorithm that iteratively estimates the eigenvector corresponding to the smallest eigenvalue. From the estimated acoustic impulse responses, the time delay can be calculated as the time difference between the main peak (direct path) of the two impulse responses or as the peak of the correlation function between the two impulse responses. Since only the time difference between the main peak (direct path) of the impulse responses is required, it is therefore not necessary to estimate the complete acoustic impulse responses.

The adaptive EVD algorithm for TDE performs much better in highly reverberant environments than the GCC-based methods. However, the adaptive EVD algorithm is—strictly speaking—only valid if either no noise or if spatiotemporally white noise is present. In this paper, we extend the adaptive EVD algorithm for TDE to the spatiotemporally colored noise case by using an adaptive generalized eigenvalue decomposition (GEVD) algorithm or by prewhitening the noisy microphone signals. Furthermore, we extend all considered TDE algorithms to the case of more than two microphones.

The paper is organized as follows. Section 2 discusses the batch, that is, nonadaptive estimation of the complete acoustic impulse responses from the recorded microphone signals. It is shown that if the length of the impulse responses is either known or can be overestimated, the complete impulse responses can be identified from the EVD of the speech correlation matrix (noiseless case and spatiotemporally white noise case) or from the GEVD of the speech and the noise correlation matrices (colored noise case). These batch impulse response estimation procedures form the basis for deriving stochastic gradient algorithms that iteratively estimate the (generalized) eigenvector corresponding to the smallest (generalized) eigenvalue. These adaptive EVD and GEVD algorithms are discussed in Section 3. In [12], it has been shown that the adaptive EVD algorithm can be used for TDE, remarkably, even when underestimating the length of the acoustic impulse responses. We will show that this result also holds for the spatiotemporally colored noise case when using the adaptive GEVD algorithm (and the adaptive prewhitening algorithm) for TDE. In Section 4, it is shown that all considered batch and adaptive TDE algorithms can easily be extended to the case of more than two microphones. Section 5 describes the simulation results for different reverberation conditions (ideal and realistic), different SNRs, and different noise sources (localized and diffuse noise source). For all conditions, it is shown that the time delays can be estimated more accurately using the adaptive GEVD algorithm and the adaptive prewhitening algorithm than using the adaptive EVD algorithm. Since the adaptive GEVD algorithm requires an estimate of the noise correlation matrix, we also analyze its sensitivity with respect to the accuracy of this noise correlation matrix estimate, showing that the performance of the adaptive GEVD algorithm may be quite sensitive to deviations, especially for low SNR scenarios.

2. BATCH ESTIMATION OF ACOUSTIC IMPULSE RESPONSES

This section discusses the nonadaptive estimation of the complete acoustic impulse responses from the recorded microphone signals, for the noiseless case as well as for the spatiotemporally white and colored noise case. The techniques discussed in this section are based on the subspace method, for example, presented in [13, 14] for different applications. We will briefly review these well-known techniques since they form the basis for deriving the stochastic gradient algorithms that iteratively estimate the (generalized) eigenvector corresponding to the smallest (generalized) eigenvalue, which will be used for TDE in practice (see Section 3).

Consider N microphones, where each microphone signal $y_n[k]$, $n = 0,\ldots,N-1$, at time k, consists of a filtered version of the clean speech signal $s[k]$ and additive noise:

$$y_n[k] = h_n[k] \otimes s[k] + v_n[k] = x_n[k] + v_n[k], \qquad (1)$$

where $x_n[k]$ and $v_n[k]$ are the speech and the noise components received at the nth microphone, respectively, $h_n[k]$ is the acoustic impulse response between the speech source and the nth microphone, and $\otimes$ denotes convolution. The additive noise can be colored and is assumed to be uncorrelated with the clean speech signal. The goal is to estimate the impulse responses $h_n[k]$ from the recorded microphone signals $y_n[k]$ without any a priori knowledge about the clean speech signal $s[k]$. From the estimates of the complete acoustic impulse responses, it is then trivial to compute the time delays between the direct paths.

If we model the acoustic impulse response $h_n[k]$ with an FIR-filter $\mathbf{h}_n$ of length L, that is,

$$\mathbf{h}_n = \begin{bmatrix} h_n[0] & h_n[1] & \cdots & h_n[L-1] \end{bmatrix}^T, \qquad (2)$$

the relation

$$\mathbf{x}_{i,L}^T[k]\mathbf{h}_j = \mathbf{x}_{j,L}^T[k]\mathbf{h}_i, \quad i, j = 0,\ldots,N-1, \qquad (3)$$

holds [12], with the L-dimensional data vector

$$\mathbf{x}_{n,L}[k] = \begin{bmatrix} x_n[k] & x_n[k-1] & \cdots & x_n[k-L+1] \end{bmatrix}^T \qquad (4)$$

since $h_j[k] \otimes x_i[k] = h_j[k] \otimes h_i[k] \otimes s[k] = h_i[k] \otimes x_j[k]$. Although we do not explicitly attribute a time index k to the impulse responses, this does not imply that they cannot be time variant. In the remainder of this section, we will assume $N = 2$, although all considered algorithms can be straightforwardly extended to the case of more than two microphones (see Section 4).

2.1. Noiseless case

The $(2K \times 2K)$-dimensional correlation matrix $\mathbf{R}_K^x$ is defined as

$$\mathbf{R}_K^x = \begin{bmatrix} \mathbf{R}_{11,K}^x & -\mathbf{R}_{10,K}^x \\ -\mathbf{R}_{01,K}^x & \mathbf{R}_{00,K}^x \end{bmatrix}, \qquad (5)$$

with the $(K \times K)$-dimensional submatrix

$$\mathbf{R}^x_{ij,K} = \mathscr{E}\{\mathbf{x}_{i,K}[k]\mathbf{x}^T_{j,K}[k]\}, \tag{6}$$

and $\mathscr{E}\{\cdot\}$ denoting the expected value operator. If $K \geq L$, that is, when the true impulse response length L is overestimated, the correlation matrix $\mathbf{R}^x_K$ has rank $K + L - 1$, and hence, its null space has dimension $K - L + 1$ under the condition that [15]

(1) the impulse responses $\mathbf{h}_0$ and $\mathbf{h}_1$ do not have common zeros;
(2) the $((K + L - 1) \times (K + L - 1))$-dimensional autocorrelation matrix of the clean speech signal $s[k]$ has full rank.

If $K = L$, the null space of $\mathbf{R}^x_K$ has dimension 1, and the $2L$-dimensional vector

$$\mathbf{u} = \begin{bmatrix}\mathbf{h}_0\\ \mathbf{h}_1\end{bmatrix} \tag{7}$$

belongs to this null space since, using (3), $\mathbf{R}^x_K\mathbf{u} = \mathbf{0}$. Consider the EVD of $\mathbf{R}^x_K$,

$$\mathbf{R}^x_K = \mathbf{V}_x\mathbf{\Delta}_x\mathbf{V}^T_x, \tag{8}$$

with $\mathbf{V}_x$ a $(2K \times 2K)$-dimensional orthogonal matrix, containing the eigenvectors, and $\mathbf{\Delta}_x$ a diagonal matrix, containing the eigenvalues. Hence, the unit-norm eigenvector, corresponding to the only zero eigenvalue of $\mathbf{R}^x_K$, contains a scaled version of the two impulse responses $\mathbf{h}_0$ and $\mathbf{h}_1$.

If $K > L$, the null space of $\mathbf{R}^x_K$ is spanned by $K - L + 1$ eigenvectors, corresponding to the $K - L + 1$ zero eigenvalues, which all contain a different filtered version of the impulse responses. By extracting the common part of the eigenvectors, which can be done, for example, by performing a QR decomposition of the full null space or by using a least squares approach [14], the correct impulse responses of length L can be identified. If $K < L$, the null space of $\mathbf{R}^x_K$ is empty and the impulse responses cannot be correctly identified.

2.2. Spatiotemporally white noise

If additive noise is present, we define the $(2K \times 2K)$-dimensional speech correlation matrix $\mathbf{R}^y_K$ and the $(2K \times 2K)$-dimensional noise correlation matrix $\mathbf{R}^v_K$, similar to (5), as

$$\mathbf{R}^y_K = \begin{bmatrix}\mathbf{R}^y_{11,K} & -\mathbf{R}^y_{10,K}\\ -\mathbf{R}^y_{01,K} & \mathbf{R}^y_{00,K}\end{bmatrix}, \qquad \mathbf{R}^v_K = \begin{bmatrix}\mathbf{R}^v_{11,K} & -\mathbf{R}^v_{10,K}\\ -\mathbf{R}^v_{01,K} & \mathbf{R}^v_{00,K}\end{bmatrix}, \tag{9}$$

with the $(K \times K)$-dimensional submatrices

$$\mathbf{R}^y_{ij,K} = \mathscr{E}\{\mathbf{y}_{i,K}[k]\mathbf{y}^T_{j,K}[k]\}, \qquad \mathbf{R}^v_{ij,K} = \mathscr{E}\{\mathbf{v}_{i,K}[k]\mathbf{v}^T_{j,K}[k]\}, \tag{10}$$

and the K-dimensional vectors $\mathbf{y}_{n,K}[k]$ and $\mathbf{v}_{n,K}[k]$ defined similarly as in (4). Assuming that the clean speech signal $s[k]$ and the noise components $v_n[k]$ are uncorrelated, we can write

$$\mathbf{R}^y_K = \mathbf{R}^x_K + \mathbf{R}^v_K. \tag{11}$$

If the noise is spatiotemporally white, that is, $\mathbf{R}^v_K = \sigma^2_v\mathbf{I}$, with σ^2_v the noise power and $\mathbf{I}$ the identity matrix, the impulse responses can be identified from the EVD of the speech correlation matrix

$$\mathbf{R}^y_K = \mathbf{V}_y\mathbf{\Delta}_y\mathbf{V}^T_y. \tag{12}$$

In this case, we can write (12) using (8) and (11) as

$$\mathbf{R}^y_K = \mathbf{V}_x(\mathbf{\Delta}_x + \sigma^2_v\mathbf{I})\mathbf{V}^T_x, \tag{13}$$

such that $\mathbf{V}_y = \mathbf{V}_x$ and $\mathbf{\Delta}_y = \mathbf{\Delta}_x + \sigma^2_v\mathbf{I}$. If $K = L$, only one of the diagonal elements of $\mathbf{\Delta}_y$ is equal to σ^2_v (smallest eigenvalue), and the eigenvector in $\mathbf{V}_y$, corresponding to this eigenvalue, again contains a scaled version of the impulse responses. If $K > L$, the procedure for estimating the impulse responses of length L is similar to the procedure in the noiseless case, now based on the $K - L + 1$ eigenvectors in $\mathbf{V}_y$ corresponding to eigenvalues which are equal to σ^2_v.

2.3. Spatiotemporally colored noise

If spatiotemporally colored noise is present, the acoustic impulse responses cannot be identified from the EVD of $\mathbf{R}^y_K$, but they can still be identified from the GEVD of $\mathbf{R}^y_K$ and $\mathbf{R}^v_K$ or from the EVD of the prewhitened speech correlation matrix. In both cases, the noise correlation matrix $\mathbf{R}^v_K$ needs to be known in advance or we have to estimate it during noise-only periods, requiring the use of a voice activity detector which determines when speech is present.

(1) *GEVD procedure*. The GEVD of $\mathbf{R}^y_K$ and $\mathbf{R}^v_K$ is defined as [16]

$$\mathbf{R}^y_K = \mathbf{Q}\mathbf{\Lambda}_y\mathbf{Q}^T, \qquad \mathbf{R}^v_K = \mathbf{Q}\mathbf{\Lambda}_v\mathbf{Q}^T, \tag{14}$$

with $\mathbf{Q}$ a $(2K \times 2K)$-dimensional invertible, but not necessarily orthogonal, matrix, and $\mathbf{\Lambda}_y$ and $\mathbf{\Lambda}_v$ diagonal matrices. From (11) and (14), it follows that

$$\begin{aligned}(\mathbf{R}^v_K)^{-1}\mathbf{R}^x_K &= (\mathbf{R}^v_K)^{-1}(\mathbf{R}^y_K - \mathbf{R}^v_K)\\ &= \mathbf{Q}^{-T}(\mathbf{\Lambda}^{-1}_v\mathbf{\Lambda}_y - \mathbf{I})\mathbf{Q}^T.\end{aligned} \tag{15}$$

Since $(\mathbf{R}^v_K)^{-1}\mathbf{R}^x_K$ has rank $K + L - 1$ ($\mathbf{R}^v_K$ is assumed to be of full rank), $K - L + 1$ diagonal elements of the diagonal matrix $\mathbf{\Lambda}^{-1}_v\mathbf{\Lambda}_y$ are equal to 1. Therefore, $K - L + 1$ columns $\mathbf{q}$ of $\mathbf{Q}^{-T}$ exist for which

$$(\mathbf{R}^v_K)^{-1}\mathbf{R}^x_K\mathbf{q} = \mathbf{0}, \tag{16}$$

such that $\mathbf{R}^x_K\mathbf{q} = \mathbf{0}$. If $K = L$, the null space of $\mathbf{R}^x_K$ has dimension 1, and the $2L$-dimensional vector $\mathbf{q}$

contains a scaled version of the impulse responses. If $K > L$, the $K - L + 1$ vectors $\mathbf{q}$ contain different filtered versions of the impulse responses, and the procedure for estimating the correct impulse responses of length L is similar to the procedure in the noiseless case.

(2) *Prewhitening procedure.* The $(2K \times 2K)$-dimensional prewhitened speech correlation matrix $\bar{\mathbf{R}}_K^y$ is defined as

$$\bar{\mathbf{R}}_K^y \triangleq (\mathbf{R}_K^v)^{-T/2}\mathbf{R}_K^y(\mathbf{R}_K^v)^{-1/2}, \tag{17}$$

with $(\mathbf{R}_K^v)^{1/2}$ the $(2K \times 2K)$-dimensional (upper-triangular) Cholesky factor of the noise correlation matrix $\mathbf{R}_K^v$, that is, $\mathbf{R}_K^v = (\mathbf{R}_K^v)^{T/2}(\mathbf{R}_K^v)^{1/2}$ [16]. From the EVD of $\bar{\mathbf{R}}_K^y$,

$$\bar{\mathbf{R}}_K^y = \bar{\mathbf{V}}_y\bar{\mathbf{\Lambda}}_y\bar{\mathbf{V}}_y^T, \tag{18}$$

it follows, using (11), that $\bar{\mathbf{R}}_K^x$ can be written as

$$\bar{\mathbf{R}}_K^x \triangleq (\mathbf{R}_K^v)^{-T/2}\mathbf{R}_K^x(\mathbf{R}_K^v)^{-1/2} = \bar{\mathbf{V}}_y(\bar{\mathbf{\Lambda}}_y - \mathbf{I})\bar{\mathbf{V}}_y^T. \tag{19}$$

Since $\bar{\mathbf{R}}_K^x$ has rank $K+L-1$, $K-L+1$ diagonal elements of the diagonal matrix $\bar{\mathbf{\Lambda}}_y$ have to be equal to 1, and hence, $K - L + 1$ columns $\bar{\mathbf{u}}$ of $\bar{\mathbf{V}}_y$ exist, for which

$$\bar{\mathbf{R}}_K^x\bar{\mathbf{u}} = (\mathbf{R}_K^v)^{-T/2}\mathbf{R}_K^x(\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{u}} = \mathbf{0} \tag{20}$$

such that $\mathbf{R}_K^x(\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{u}} = \mathbf{0}$. If $K = L$, the null space of $\mathbf{R}_K^x$ has dimension 1, and the vector $(\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{u}}$ contains a scaled version of the impulse responses. If $K > L$, the $K - L + 1$ vectors $(\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{u}}$ contain different filtered versions of the impulse responses, and the procedure for estimating the correct impulse responses of length L is similar to the procedure in the noiseless case.

It is readily verified that the GEVD procedure and the prewhitening procedure are in fact equivalent since

$$\bar{\mathbf{\Lambda}}_y = \mathbf{\Lambda}_v^{-1}\mathbf{\Lambda}_y, \qquad \mathbf{Q}^{-T} = (\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{V}}_y. \tag{21}$$

However, the adaptive versions of both algorithms, which are presented in Section 3 and which will be used for TDE in practice, can produce different results.

2.4. Practical computation

In practice, we will not work with correlation matrices, but with data matrices. The $(p \times 2K)$-dimensional speech data matrix $\mathbf{Y}_K[k]$ is defined as

$$\mathbf{Y}_K[k] = \begin{bmatrix} \mathbf{y}_K^T[k] \\ \mathbf{y}_K^T[k+1] \\ \vdots \\ \mathbf{y}_K^T[k+p-1] \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{1,K}^T[k] & -\mathbf{y}_{0,K}^T[k] \\ \mathbf{y}_{1,K}^T[k+1] & -\mathbf{y}_{0,K}^T[k+1] \\ \vdots & \vdots \\ \mathbf{y}_{1,K}^T[k+p-1] & -\mathbf{y}_{0,K}^T[k+p-1] \end{bmatrix}, \tag{22}$$

with p typically much larger than K, such that the empirical speech correlation matrix can be computed as $\mathbf{R}_K^y = \mathbf{Y}_K^T[k]\mathbf{Y}_K[k]/p$. The noise data matrix $\mathbf{V}_K[k]$ is defined similarly.

(1) *GSVD procedure.* Instead of computing the GEVD of $\mathbf{R}_K^y$ and $\mathbf{R}_K^v$, we compute the generalized singular value decomposition (GSVD) of the data matrices $\mathbf{Y}_K[k]$ and $\mathbf{V}_K[k]$, defined as

$$\mathbf{Y}_K[k] = \mathbf{U}_y\mathbf{\Sigma}_y\mathbf{Q}^T, \qquad \mathbf{V}_K[k] = \mathbf{U}_v\mathbf{\Sigma}_v\mathbf{Q}^T, \tag{23}$$

with $\mathbf{U}_y$ and $\mathbf{U}_v$ orthogonal matrices, $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_v$ diagonal matrices, and $\mathbf{Q}$ a $(2K \times 2K)$-dimensional invertible, but not necessarily orthogonal, matrix [16, 17]. Again, the impulse responses are estimated from the columns $\mathbf{q}$ of the matrix $\mathbf{Q}^{-T}$.

(2) *Prewhitening procedure.* The prewhitened speech data matrix $\bar{\mathbf{Y}}_K[k]$ is defined as

$$\bar{\mathbf{Y}}_K[k] = \mathbf{Y}_K[k](\mathbf{R}_K^v)^{-1/2}, \tag{24}$$

where the $(2K \times 2K)$-dimensional (upper-triangular) Cholesky factor $(\mathbf{R}_K^v)^{1/2}$ can be computed using the QR decomposition of the noise data matrix, that is,

$$\mathbf{V}_K[k] = \mathbf{Q}_v(\mathbf{R}_K^v)^{1/2}. \tag{25}$$

The singular value decomposition (SVD) of $\bar{\mathbf{Y}}_K[k]$ is defined as

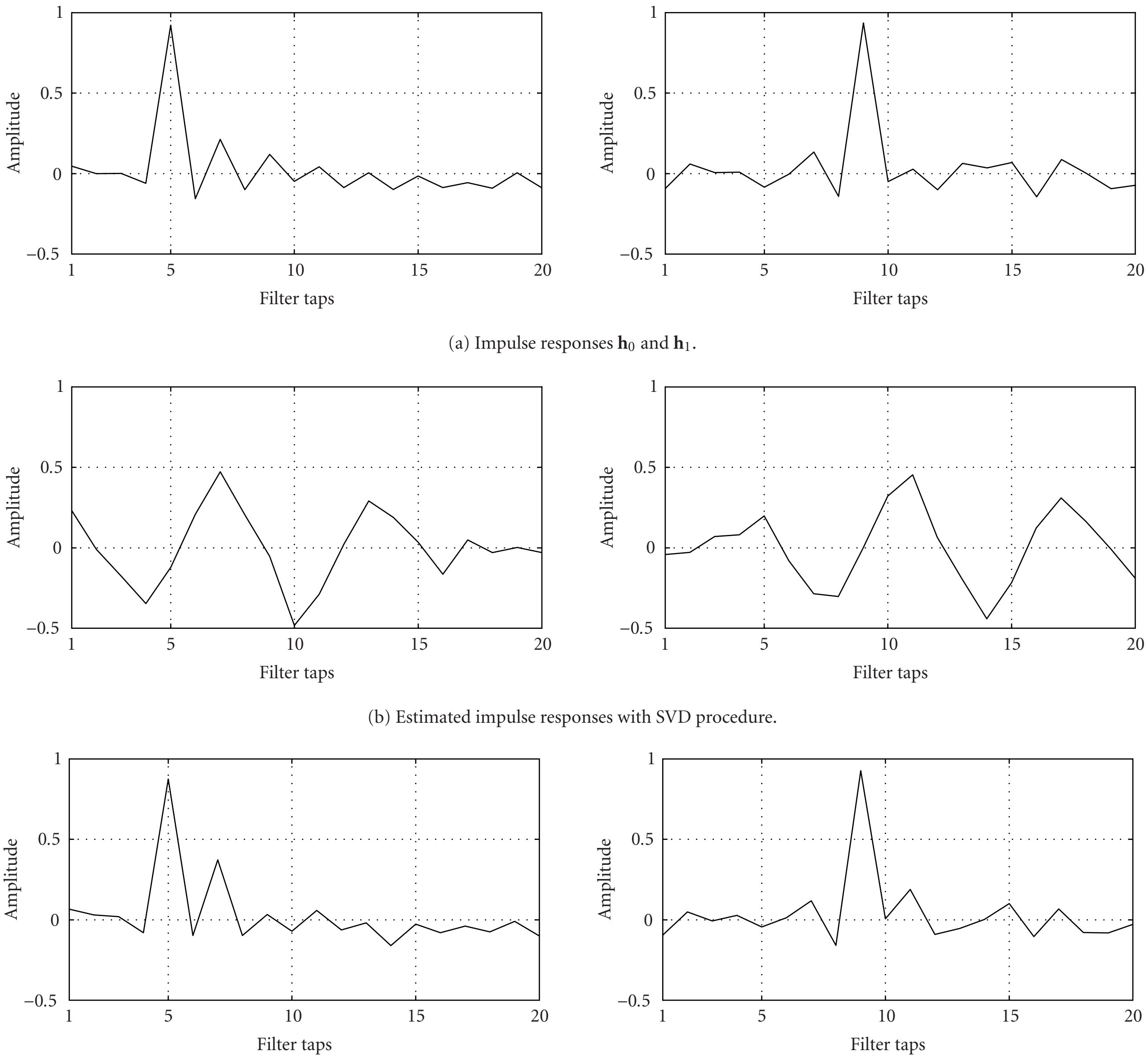
$$\bar{\mathbf{Y}}_K[k] = \bar{\mathbf{U}}_y\bar{\mathbf{\Sigma}}_y\bar{\mathbf{V}}_y^T, \tag{26}$$

with $\bar{\mathbf{U}}_y$ and $\bar{\mathbf{V}}_y$ orthogonal matrices and $\bar{\mathbf{\Sigma}}_y$ a diagonal matrix. Again, the impulse responses are estimated from the columns $\bar{\mathbf{u}}$ of the matrix $\bar{\mathbf{V}}_y$.

2.5. Simulation results

We have filtered a 16-kHz speech segment of 160000 samples (10 seconds) with 2 impulse responses ($L = 20$), which are depicted in Figure 1a. A stationary colored speech-like noise signal, having the same long-term spectrum as speech [18], has been added, and the SNR of the microphone signals is 10 dB.

Figures 1a and 1b show the estimated impulse responses ($K = L$), for the SVD procedure and the GSVD procedure, using all microphone samples. As can be clearly seen, the impulse responses are almost correctly estimated with the GSVD procedure, which is not the case for the SVD procedure. Because the assumption of uncorrelated speech and noise segments is not always perfectly satisfied, that is, $\mathbf{X}_K^T[k]\mathbf{V}_K[k] \approx \mathbf{0}$, small estimation errors occur in the GSVD procedure. In our simulations, we have noticed that the better this assumption is satisfied, that is, the higher the SNR and the longer the speech and the noise segments, the smaller the estimation error becomes. This fact has also been observed in [14].



(a) Impulse responses $\mathbf{h}_0$ and $\mathbf{h}_1$.

(b) Estimated impulse responses with SVD procedure.

(c) Estimated impulse responses with GSVD procedure.

Figure 1

3. ADAPTIVE PROCEDURE FOR TIME DELAY ESTIMATION

In practice, acoustic impulse responses may have thousands of taps,depending on the room reverberation. Because of the correlated nature of speech, correspondingly large autocorrelation matrices of the clean speech signal $s[k]$ can be rank deficient or at least ill conditioned [19]. Therefore, it is quite difficult to identify the complete impulse responses, especially when a large amount of background noise is present [14]. If we underestimate the length of the impulse responses ($K < L$), the acoustic impulse responses estimated with the batch procedures are biased. This makes it difficult to calculate the correct time delays from these estimated acoustic impulse responses.

In [12], an adaptive EVD algorithm has been presented, which iteratively estimates the eigenvector corresponding to the smallest eigenvalue. Remarkably, even when underestimating the length of the impulse responses ($K < L$), simulations show that this adaptive EVD algorithm is still able to identify the main peak (direct path) of the impulse responses. Obviously, only the time difference between the main peak of the impulse responses is required for TDE.

Strictly speaking, the adaptive EVD algorithm is only valid when no noise or when spatiotemporally white noise is present. In this section, we therefore extend the adaptive EVD algorithm to the colored noise case by deriving stochastic gradient algorithms for the procedures presented in Section 2.3, that is, algorithms which iteratively estimate the generalized eigenvector corresponding to the smallest generalized eigenvalue. Using simulations with spatiotemporally colored noise, it will be shown that—just as for the adaptive EVD algorithm—it is possible to correctly estimate the time delays with the adaptive GEVD algorithm, even when underestimating the length of the acoustic impulse responses (see Section 5).

In the remainder of the text, we will assume that the length of the acoustic impulse responses is underestimated ($K < L$), and hence, we will derive algorithms that estimate the one-dimensional subspace corresponding to the smallest (generalized) eigenvalue.

3.1. Adaptive EVD algorithm [12]

Instead of updating the full EVD of $\mathbf{R}_K^y$ [20] and then using the eigenvector corresponding to the smallest eigenvalue, it is possible to iteratively estimate this eigenvector by minimizing the cost function $\mathbf{u}^T\mathbf{R}_K^y\mathbf{u}$ subject to the constraint $\mathbf{u}^T\mathbf{u} = 1$. A cheap procedure consists in minimizing the mean square value of the error signal $e[k]$, defined as

$$e[k] = \frac{\mathbf{u}^T[k]\mathbf{y}_K[k]}{||\mathbf{u}[k]||}, \tag{27}$$

with $\mathbf{y}_K[k] = [\mathbf{y}_{1,K}^T[k] \;\; -\mathbf{y}_{0,K}^T[k]]^T$. This expression in fact is a Rayleigh quotient, where $\lambda_y^{\max} \geq \mathscr{E}\{e^2[k]\} \geq \lambda_y^{\min}$, with $\lambda_y^{\max}$ and $\lambda_y^{\min}$, respectively, the largest and the smallest eigenvalues of the correlation matrix $\mathbf{R}_K^y$. Minimizing (27) can be done, for example, using a gradient-descent LMS procedure, where normalization is included in each iteration step in order to avoid roundoff error propagation [21],

$$\mathbf{u}[k+1] = \frac{\mathbf{u}[k] - \mu e[k](\partial e[k]/\partial \mathbf{u}[k])}{||\mathbf{u}[k] - \mu e[k](\partial e[k]/\partial \mathbf{u}[k])||}, \tag{28}$$

with μ the step size of the adaptive algorithm. The gradient of $e[k]$ is equal to

$$\frac{\partial e[k]}{\partial \mathbf{u}[k]} = \frac{1}{||\mathbf{u}[k]||}\left(\mathbf{y}_K[k] - e[k]\frac{\mathbf{u}[k]}{||\mathbf{u}[k]||}\right). \tag{29}$$

In [12], it has been assumed that the smallest eigenvalue of $\mathbf{R}_K^y$ is very small (in the noiseless case) such that the gradient eventually reduces to $\partial e[k]/\partial\mathbf{u}[k] \approx \mathbf{y}_K[k]$, and the update formulas become

$$\begin{aligned} e[k] &= \mathbf{u}^T[k]\mathbf{y}_K[k], \\ \mathbf{u}[k+1] &= \frac{\mathbf{u}[k] - \mu e[k]\mathbf{y}_K[k]}{||\mathbf{u}[k] - \mu e[k]\mathbf{y}_K[k]||}. \end{aligned} \tag{30}$$

In [12], it has been indicated that a good initialization of $\mathbf{u}$ and a proper choice of the parameters K and μ are essential for a good convergence behavior. It has also been shown by simulations that the adaptive EVD algorithm performs more robustly in highly reverberant environments than the GCC-based methods.

3.2. Adaptive GEVD and prewhitening algorithm

For the noise-robust GEVD and prewhitening procedures, described in Section 2.3, it is also possible to derive stochastic gradient algorithms which iteratively estimate the generalized eigenvector corresponding to the smallest generalized eigenvalue of $\mathbf{R}_K^y$ and $\mathbf{R}_K^v$. It will be assumed that the noise correlation matrix $\mathbf{R}_K^v$ (or its Cholesky factor) is either known or updated during noise-only periods. Since the noise correlation matrix cannot be updated during speech-and-noise periods, we have to assume that the noise is stationary enough such that the noise correlation matrix computed during noise-only periods can be used in the update formulas during subsequent speech-and-noise periods.

Adaptive GEVD algorithm

Instead of updating the full GEVD of $\mathbf{R}_K^y$ and $\mathbf{R}_K^v$ [22] and then using the generalized eigenvector corresponding to the smallest generalized eigenvalue, it is possible to iteratively estimate this generalized eigenvector by minimizing the cost function $\mathbf{q}^T\mathbf{R}_K^y\mathbf{q}$ subject to the constraint $\mathbf{q}^T\mathbf{R}_K^v\mathbf{q} = 1$. A cheap procedure consists in minimizing the mean square value of the error signal $e[k]$, defined as the generalized Rayleigh quotient

$$e[k] = \frac{\mathbf{q}^T[k]\mathbf{y}_K[k]}{\sqrt{\mathbf{q}^T[k]\mathbf{R}_K^v\mathbf{q}[k]}} = \frac{\mathbf{q}^T[k]\mathbf{y}_K[k]}{||(\mathbf{R}_K^v)^{1/2}\mathbf{q}[k]||}, \tag{31}$$

which can be done, for example, using a gradient-descent LMS procedure

$$\mathbf{q}[k+1] = \mathbf{q}[k] - \mu e[k]\frac{\partial e[k]}{\partial \mathbf{q}[k]}, \tag{32}$$

with μ the step size of the adaptive algorithm. The gradient of $e[k]$ now is equal to

$$\frac{\partial e[k]}{\partial \mathbf{q}[k]} = \frac{1}{\sqrt{\mathbf{q}^T[k]\mathbf{R}_K^v\mathbf{q}[k]}}\left(\mathbf{y}_K[k] - e[k]\frac{\mathbf{R}_K^v\mathbf{q}[k]}{\sqrt{\mathbf{q}^T[k]\mathbf{R}_K^v\mathbf{q}[k]}}\right). \tag{33}$$

Substituting (31) and (33) into (32) gives

$$\begin{aligned} &\mathbf{q}[k+1] \\ &\quad = \mathbf{q}[k] - \frac{\mu}{\mathbf{q}^T[k]\mathbf{R}_K^v\mathbf{q}[k]}\left(\mathbf{y}_K[k]\mathbf{y}_K^T[k]\mathbf{q}[k] - e^2[k]\mathbf{R}_K^v\mathbf{q}[k]\right) \end{aligned} \tag{34}$$

such that, when taking mathematical expectation after convergence, we get

$$\mathbf{R}_K^y\mathbf{q}[\infty] = \mathscr{E}\{e^2[k]\}\mathbf{R}_K^v\mathbf{q}[\infty]. \tag{35}$$

This is exactly what is desired, that is, $\mathbf{q}[\infty]$ is the generalized eigenvector which corresponds to the smallest generalized eigenvalue of $\mathbf{R}_K^y$ and $\mathbf{R}_K^v$. Since the smallest generalized eigenvalue is equal to 1 (see Section 2.3), we cannot further simplify the expression in (34). In order to avoid roundoff error propagation, we include an additional normalization in each iteration step such that the update formulas can be written as

$$\begin{aligned} e[k] &= \mathbf{q}^T[k]\mathbf{y}_K[k], \\ \tilde{\mathbf{q}}[k+1] &= \mathbf{q}[k] - \mu e[k]\{\mathbf{y}_K[k] - e[k]\mathbf{R}_K^v\mathbf{q}[k]\}, \\ \mathbf{q}[k+1] &= \frac{\tilde{\mathbf{q}}[k+1]}{\sqrt{\tilde{\mathbf{q}}^T[k+1]\mathbf{R}_K^v\tilde{\mathbf{q}}[k+1]}}. \end{aligned} \tag{36}$$

Adaptive prewhitening algorithm

The prewhitening procedure can be made adaptive by using prewhitened speech data vectors $\bar{\mathbf{y}}_K[k] = (\mathbf{R}_K^v)^{-T/2}\mathbf{y}_K[k]$ in the adaptive EVD procedure of Section 3.1. The update formulas then become

$$\begin{aligned} e[k] &= \bar{\mathbf{u}}^T[k]\bar{\mathbf{y}}_K[k], \\ \bar{\mathbf{u}}[k+1] &= \frac{\bar{\mathbf{u}}[k] - \mu e[k](\bar{\mathbf{y}}_K[k] - e[k]\bar{\mathbf{u}}[k])}{\|\bar{\mathbf{u}}[k] - \mu e[k](\bar{\mathbf{y}}_K[k] - e[k]\bar{\mathbf{u}}[k])\|}. \end{aligned} \tag{37}$$

Note that the gradient $\partial e[k]/\partial\bar{\mathbf{u}}[k]$ cannot now be approximated by $\bar{\mathbf{y}}_K[k]$ (as is the case for the adaptive EVD algorithm) since the smallest eigenvalue of $\bar{\mathbf{R}}_K^y$ is not equal to zero (see Section 2.3). The impulse response at time k is estimated as $(\mathbf{R}_K^v)^{-1/2}\bar{\mathbf{u}}[k]$. If the noise correlation matrix $\mathbf{R}_K^v$ is not known in advance, the Cholesky factor $(\mathbf{R}_K^v)^{-1/2}$ can be updated by inverse QR updating during noise-only periods.

The computational complexity of the adaptive GEVD and the adaptive prewhitening algorithm is higher than that of the adaptive EVD algorithm since in each iteration step two additional matrix-vector multiplications (either with the noise correlation matrix or with the inverse Cholesky factor) have to be performed. Reducing the computational complexity of these algorithms is a topic of further research. The noise correlation matrix $\mathbf{R}_K^v$ in the adaptive GEVD algorithm could be replaced, for example, by its instantaneous estimate $\mathbf{v}[k']\mathbf{v}^T[k']$, where $\mathbf{v}[k']$ is a noise data vector which is stored in a buffer during noise-only periods and which is used in the update equations during subsequent speech-and-noise periods. Similarly as in the momentum LMS algorithm [23], it could then also be advantageous to perform an averaging operation on (part of) the gradient $\partial e[k]/\partial\mathbf{q}[k]$.

In addition, the computational complexity of all presented adaptive TDE algorithms can be reduced by using subsampling, that is, the estimated impulse response vectors are not updated for every time step at the expense of a slower convergence and tracking behavior.

4. EXTENSION TO MORE THAN TWO MICROPHONES

All presented (batch and adaptive) algorithms can easily be extended to the case of more than two microphones, either by constructing $(p(N-1) \times NK)$-dimensional data matrices, considering the time delays between every microphone and the first microphone, or by constructing $(pC_N^2 \times NK)$-dimensional data matrices (with C_N^2 all possible combinations of two out of N), considering the time delays between every combination of two microphones. For example, if $N = 3$, the speech data matrix $\mathbf{Y}_K[k]$ in (22) can be redefined by replacing each vector $\mathbf{y}_K^T[k]$ by the matrix

$$\begin{bmatrix} \mathbf{y}_{1,K}^T[k] & -\mathbf{y}_{0,K}^T[k] & \mathbf{0} \\ \mathbf{y}_{2,K}^T[k] & \mathbf{0} & -\mathbf{y}_{0,K}^T[k] \end{bmatrix}, \tag{38}$$

considering time delays between every microphone and the first microphone, or by the matrix

$$\begin{bmatrix} \mathbf{y}_{1,K}^T[k] & -\mathbf{y}_{0,K}^T[k] & \mathbf{0} \\ \mathbf{y}_{2,K}^T[k] & \mathbf{0} & -\mathbf{y}_{0,K}^T[k] \\ \mathbf{0} & \mathbf{y}_{2,K}^T[k] & -\mathbf{y}_{1,K}^T[k] \end{bmatrix}, \tag{39}$$

considering time delays between every combination of two microphones. The noise data matrix $\mathbf{V}_K[k]$ is constructed similarly. It can easily be verified that, if $K = L$ and for the noiseless case, the NL-dimensional vector consisting of the impulse responses

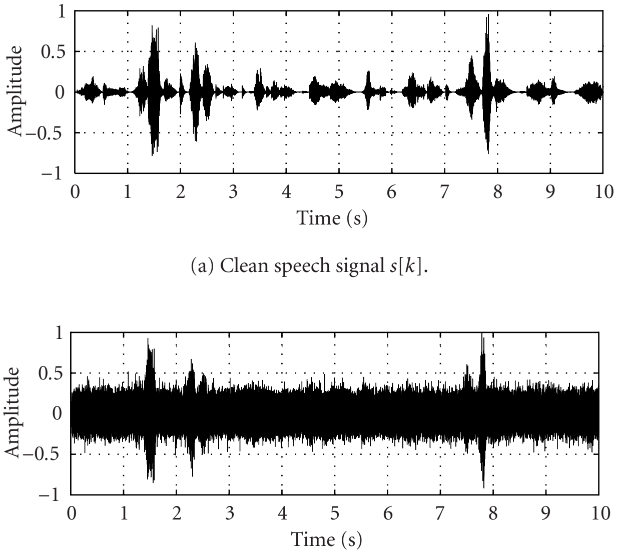
$$\mathbf{u} = \begin{bmatrix} \mathbf{h}_0 \\ \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{N-1} \end{bmatrix} \tag{40}$$

belongs to the null space of the speech data matrix. Therefore all presented (batch and adaptive) algorithms can be used with the redefined data matrices and data vectors. For the adaptive algorithms, several updates now have to be performed in each iteration step, either with $N-1$ or C_N^2 data vectors. However, the computational complexity can be reduced, for example, by only performing an update with one data vector in each iteration step, that is, by using consecutive rows of the matrices (38) or (39) in each iteration step.

In [24], another adaptive algorithm has been proposed for extending these TDE procedures to more than two microphones. This algorithm is based on the minimization of an error signal constructed using all cross-correlations between the different microphone signals, either using a stochastic gradient (MCLMS) or a Newton (MCN) method, and requires only one update in each iteration step. It has been shown that this class of algorithms can be efficiently implemented in the frequency domain [25].

5. SIMULATIONS

We have performed several simulations analyzing the performance of the different adaptive TDE algorithms (EVD, GEVD, and prewhitening) for different reverberation conditions (ideal and realistic), different SNRs, and different noise sources (localized and diffuse noise source). In all simulations, the sampling frequency $f_s = 16$ kHz and the length of the used signals is 160000 samples (10 seconds). We have used a continuous clean speech signal $s[k]$ (plotted in



(a) Clean speech signal $s[k]$.

(b) Noisy speech signal $y_0[k]$ (SNR = −5 dB).

Figure 2

Figure 2a), such that no voice activity detector is required and we continuously estimate the time delays. For the simulations in Sections 5.1, 5.2, and 5.3, we have calculated the (exact) noise correlation matrix estimate $\mathbf{R}_K^v$ in advance using the noise components $v_n[k]$, whereas, in Section 5.4, the sensitivity of the adaptive GEVD algorithm with respect to the accuracy of this noise correlation matrix estimate is analyzed. The time delay between the microphone signals is computed using the peak of the correlation function between the different estimated acoustic impulse responses.

5.1. *No reverberation,* $N = 2$

In a first simulation, we have assumed no reverberation and $N = 2$ microphones. We have used a colored noise signal constructed by filtering white noise with the five-tap FIR filter $[1\ \ -4\ \ 6\ \ 4\ \ 0.5]$. The microphone signals are constructed such that the time delay between the speech components is −8 samples, whereas the time delay between the noise components is 5 samples. We have performed simulations using the adaptive EVD, prewhitening, and GEVD algorithms for different SNRs (−5 dB, 0 dB, 5 dB). The used filter length $K = 40$, the subsampling factor for the update formulas is 10, and the step size μ of the adaptive algorithms is chosen such that the optimal performance is obtained, that is, most of the estimated time delays are close to the correct time delay (in this case, $\mu = 1e - 7$ for all algorithms).

Figure 3 shows the TDE convergence plots for the different adaptive algorithms for different SNRs. The correct time delay is indicated by the dashed line. As can be seen, the adaptive EVD algorithm converges to the correct time delay for SNR = 5 dB, but converges to the wrong time delay of the noise source for lower SNRs. Both the adaptive prewhitening and the adaptive GEVD algorithm converge to the correct time delay for all SNRs. The adaptive GEVD algorithm converges faster than the adaptive prewhitening algorithm.

5.2. *Realistic conditions,* $N = 2$

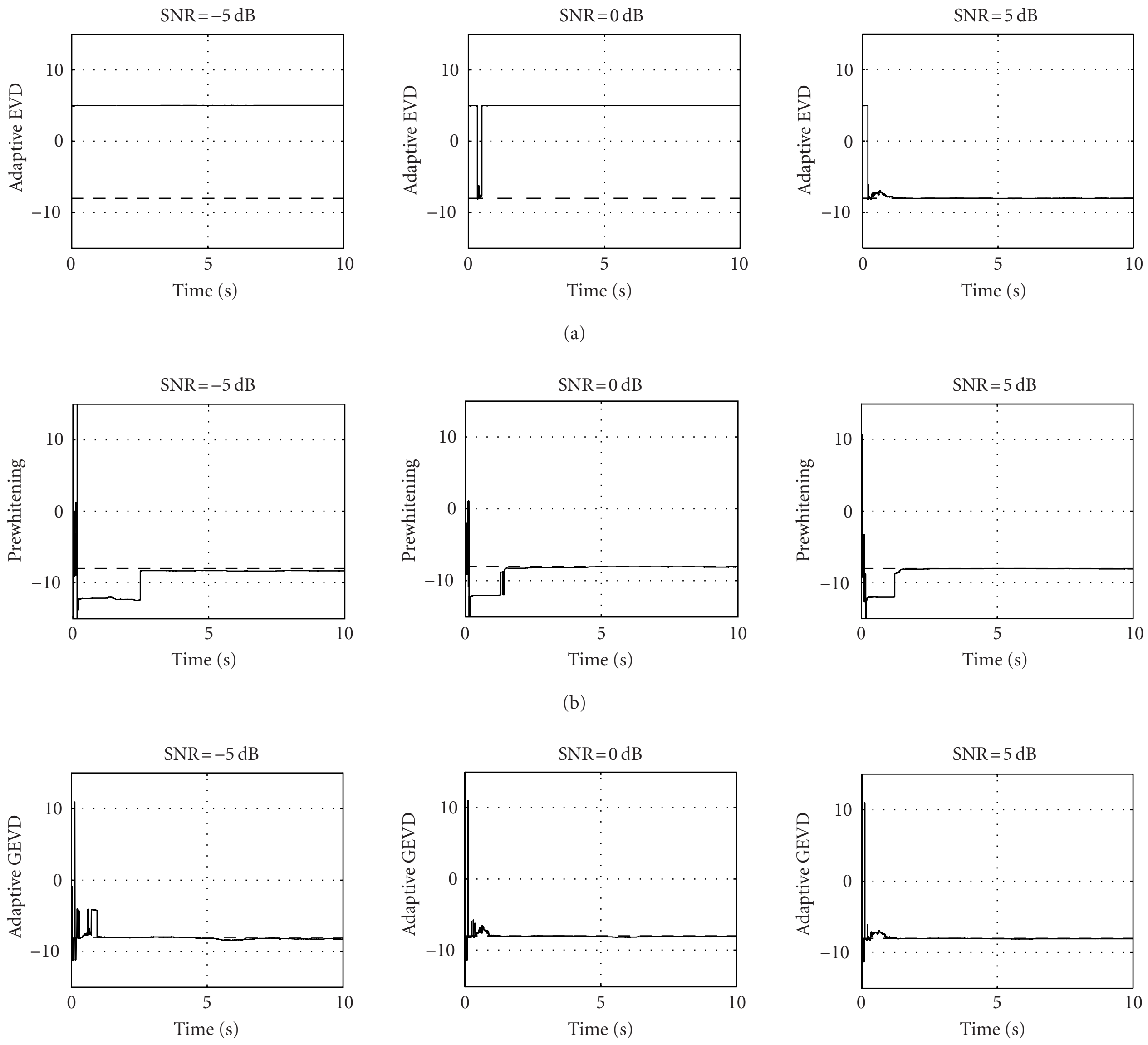
In order to simulate realistic reverberation conditions, we have simulated a room with dimensions $5m \times 4m \times 2m$, having a reverberation time $T_{60} = 250$ milliseconds. The reverberation time T_{60} can be expressed as a function of the absorption coefficient γ of the walls, according to Eyring's formula [26]

$$T_{60} = \frac{0.163V}{-S\log(1-\gamma)}, \tag{41}$$

with V the volume of the room and S the total surface of the room. The room consists of a microphone array, with $N = 2$ omnidirectional microphones at positions $[1\ \ 1\ \ 1]$ and $[1.5\ \ 1\ \ 1]$, and a speech source at position $[2\ \ 2\ \ 1.7]$. The speech components $x_n[k]$ received at the microphone array are filtered versions of the clean speech signal using simulated acoustic impulse responses, which are constructed using the image method [27, 28] with a filter length $L = 1000$. Figure 4 depicts the acoustic impulse responses $h_0[k]$ and $h_1[k]$ for the speech source. The exact time delay between the speech components is −12.18 samples, which has been obtained by a simple geometrical calculation. We will perform simulations for a localized noise source at position $[4\ \ 1.5\ \ 1]$ and for a diffuse, that is, isotropic, noise source. For the localized noise source, we have used a stationary colored speech-like noise signal having the same long-term spectrum as speech [18], and the noise components $v_n[k]$ received at the microphone array are filtered versions using simulated acoustic impulse responses. The diffuse noise source has been generated by considering 1000 uncorrelated white noise sources equally distributed over all directions.

We have performed simulations using the adaptive EVD, prewhitening, and GEVD algorithms for different SNRs (ranging from −10 dB to 10 dB) and for subsampling factor 1, that is, no subsampling. The noisy microphone signal $y_0[k]$ with SNR = −5 dB is plotted in Figure 2b. We have used $K = 40$ and, for each algorithm, we have chosen the step size μ which gives the best performance, that is the smallest percentage of anomalous estimates. An anomalous estimate is defined as a time delay estimate which corresponds to an angle outside a 5° error region from the correct angle of incidence.

Figure 5 shows the TDE convergence plots for SNR = −5 dB. The correct time delay is indicated by the dashed line. As can be seen, the adaptive EVD algorithm does not converge to the correct time delay (except for the signal segment between 1.5 and 3 seconds, where the segmental SNR is quite high, see Figure 2b), whereas both the adaptive prewhitening and GEVD algorithms converge to the correct time delay. Figure 6 shows the TDE convergence plots for SNR = 0 dB. In this case, all algorithms converge to the correct time delay, but both the adaptive prewhitening and the adaptive GEVD algorithm converge faster than the adaptive EVD algorithm. Note that it is quite remarkable that the adaptive EVD



(a)

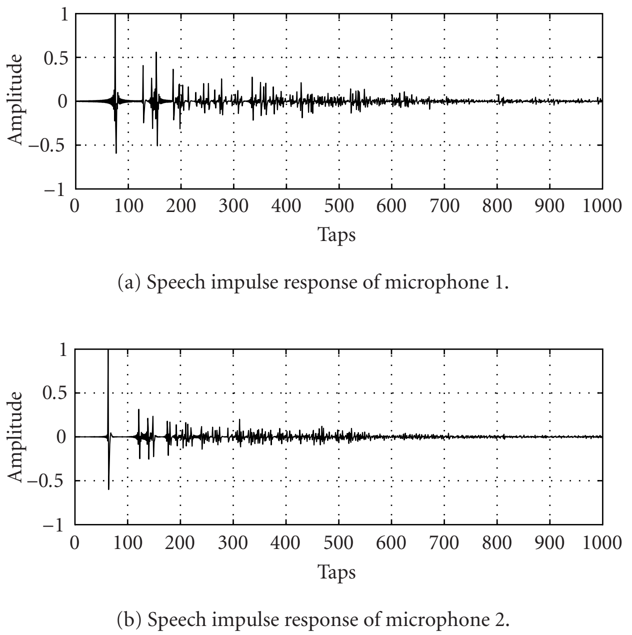
(b)

(c)

Figure 3: TDE convergence plots of (a) adaptive EVD, (b) prewhitening, and (c) GEVD algorithms for different SNRs without reverberation ($N = 2$, $K = 40$, subsampling = 10, and $\mu = 1e - 7$).

algorithm converges to the correct time delay for SNR = 0 dB without any knowledge of the noise characteristics.

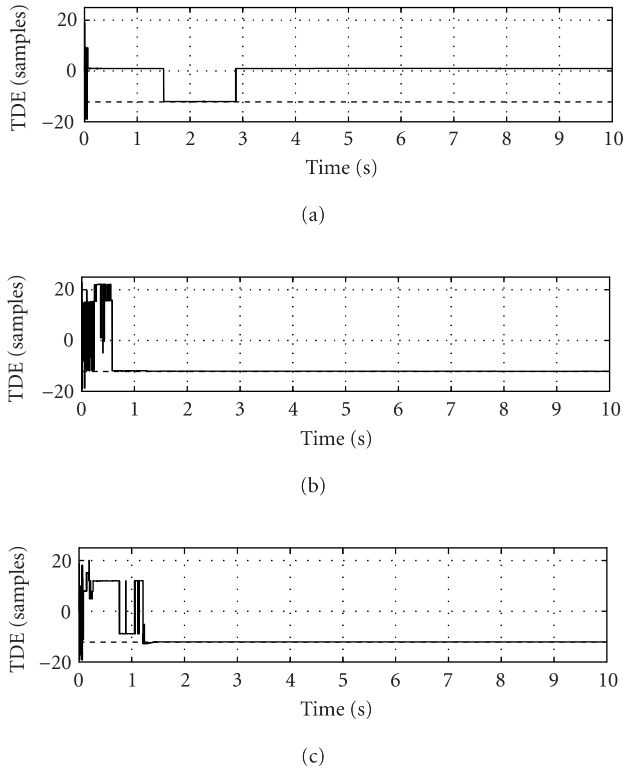
For the different adaptive TDE algorithms and for different SNRs, Figure 7a shows the percentage of anomalous time delay estimates for the localized noise source, whereas Figure 7b shows the percentage of anomalous estimates for the diffuse noise source. As can be seen from both figures, the performance of the adaptive prewhitening and the adaptive GEVD algorithms is better than the performance of the adaptive EVD algorithm for all scenarios. For the localized noise source, the performance of the adaptive EVD algorithm decreases dramatically when the SNR is smaller than 0 dB, whereas the performance of both the adaptive prewhitening and the adaptive GEVD algorithms only slightly decreases with decreasing SNR. However, the difference in performance between the adaptive EVD and GEVD algorithms is negligible when the SNR is higher than 5 dB. For a diffuse noise source, the difference in performance between all TDE algorithms is small for all SNRs, and hence, there is no real advantage in using the adaptive prewhitening or GEVD algorithms. For a diffuse noise source, the adaptive EVD algorithm has a remarkably good performance for low SNRs. This can be partly explained by the fact that, for a large microphone distance, the noise correlation matrix $\mathbf{R}_K^v$ for a diffuse noise source is approximately equal to the identity matrix.



(a) Speech impulse response of microphone 1.

(b) Speech impulse response of microphone 2.

Figure 4: Acoustic impulse responses $h_0[k]$ and $h_1[k]$ for the speech source.

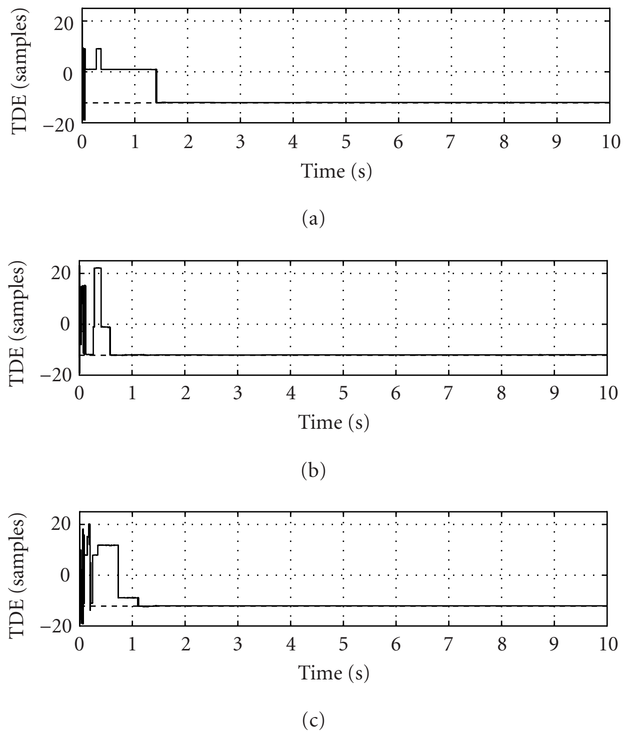


(a)

(b)

(c)

Figure 5: TDE convergence plots of (a) adaptive EVD algorithm ($\mu = 1e - 3$), (b) adaptive prewhitening algorithm ($\mu = 1e - 5$), and (c) adaptive GEVD algorithm ($\mu = 1e - 3$) with $N = 2$, $K = 40$, SNR $= -5$ dB, $T_{60} = 250$ milliseconds, and subsampling = 1. The correct time delay is indicated by the dashed line.



(a)

(b)

(c)

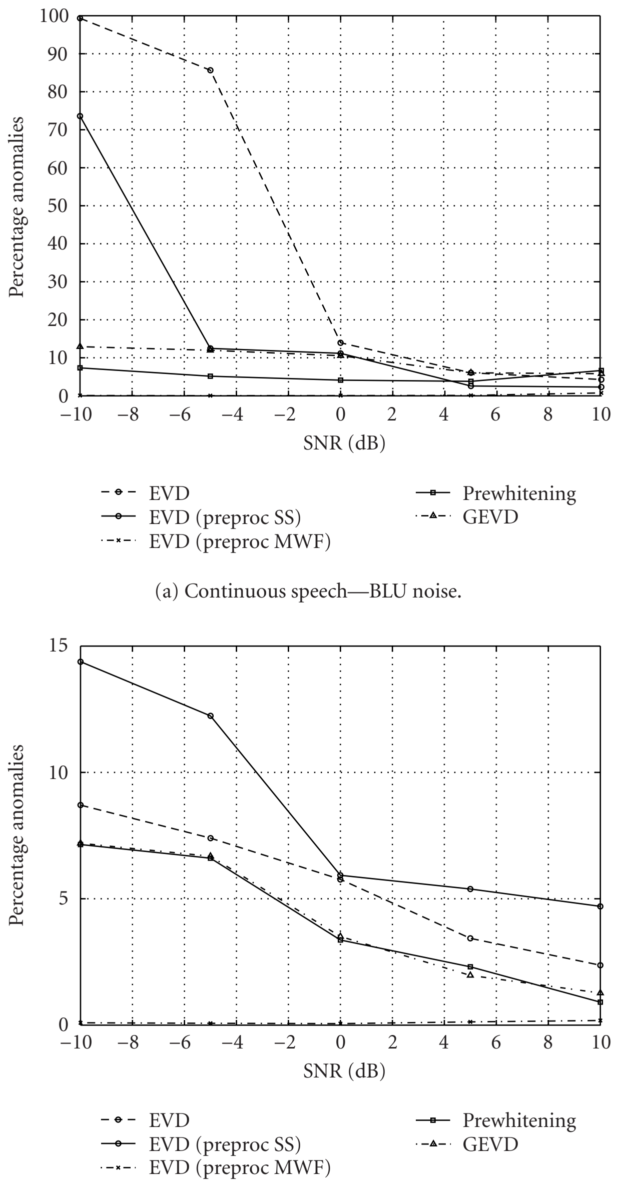
Figure 6: TDE convergence plots of (a) adaptive EVD algorithm ($\mu = 1e - 3$), (b) adaptive prewhitening algorithm ($\mu = 1e - 5$), and (c) adaptive GEVD algorithm ($\mu = 1e - 3$) with $N = 2$, $K = 40$, SNR $= 0$ dB, $T_{60} = 250$ milliseconds, and subsampling = 1. The correct time delay is indicated by the dashed line.

Instead of using the adaptive prewhitening or the adaptive GEVD algorithm in highly noisy acoustic environments, it is also possible to first perform a noise reduction procedure as a preprocessing step for the adaptive EVD algorithm. We have considered two noise reduction algorithms.

(i) A *spectral subtraction* (SS) technique on each microphone signal independently [29]. We have calculated the average noise spectrum for each microphone signal in advance and have used a simple magnitude subtraction weighting function [30] (FFT size = 512, half-wave rectification, no noise overestimation, and no magnitude averaging).

(ii) A *multichannel Wiener filtering* (MWF) technique, making an optimal (MMSE) estimate of the speech components in each microphone signal using knowledge about the spatiotemporal correlation properties of the noise components. We have used a GSVD based implementation [31] with a filter length $K = 40$ on each microphone signal. Other implementations having a lower computational complexity, such as a subband implementation [32] or a QRD-based implementation [33], could have also been used.

From Figure 7, it can be seen that, for a localized noise source, the SS preprocessing gives rise to a significant per-



(a) Continuous speech—BLU noise.

(b) Continuous speech—diffuse noise.

Figure 7: Percentage of anomalous estimates versus SNR for adaptive EVD (no preprocessing, SS and MWF preprocessing), adaptive prewhitening, and adaptive GEVD algorithms for (a) localized noise source and (b) diffuse noise source ($N = 2$, $K = 40$, $T_{60} = 250$ milliseconds, and subsampling = 1).

formance improvement, certainly for low SNR scenarios, whereas, for the diffuse noise source, the SS preprocessing apparently does not give rise to a performance improvement. For both the localized and the diffuse noise source, the MWF preprocessing reduces the percentage of anomalous estimates to below 1% for all SNRs. However, the computational complexity of the adaptive EVD algorithm combined with MWF preprocessing is still higher than the computational complexity of the adaptive GEVD algorithm.

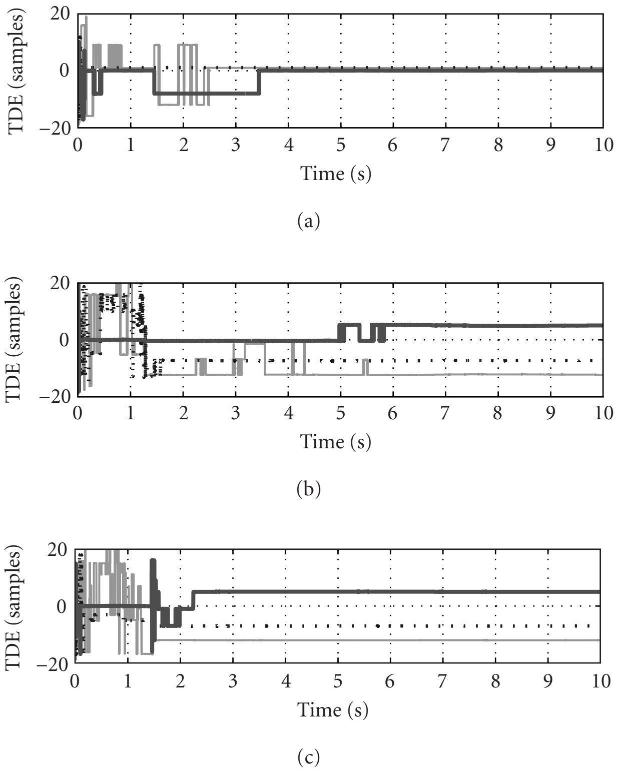
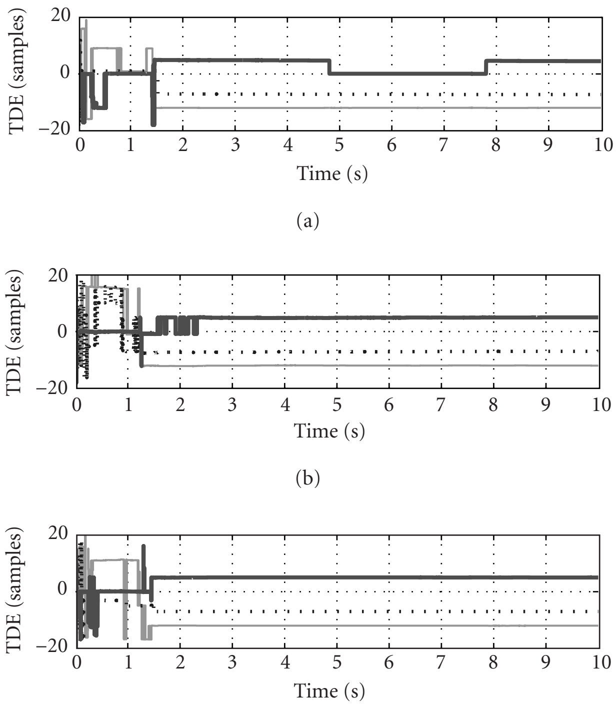


Figure 8: TDE convergence plots of (a) adaptive EVD algorithm ($\mu = 1e-3$), (b) adaptive prewhitening algorithm ($\mu = 1e-4$), and (c) adaptive GEVD algorithm ($\mu = 1e-2$) with $N = 3$, $K = 40$, SNR = −5 dB, $T_{60} = 250$ milliseconds, and subsampling = 10. The TDE between microphones 1 and 2 is denoted by the solid line, the TDE between microphones 1 and 3 by the dotted line, and the TDE between microphones 2 and 3 by the thick solid line.

5.3. *Realistic conditions,* $N = 3$

For the same acoustical conditions as in Section 5.2, we have performed simulations using $N = 3$ microphones, where the position of the third microphone is $[1\ \ 1\ \ 1.5]$. We have considered the time delays between every combination of 2 microphones and, in each iteration step, we have performed updates using all three data vectors from (39). The exact time delay between the speech components of the first and the second microphone signal is −12.18 samples, between the first and the third microphone signal −7.04 samples, and between the second and the third microphone signal 5.14 samples. We have performed simulations for different SNRs (−5 dB, 0 dB), the used filter length $K = 40$, the subsampling factor is 10, and, for each algorithm, we have chosen the step size μ which gives rise to the best performance.

Figure 8 shows the TDE convergence plots for SNR = −5 dB. As can be seen, the adaptive EVD algorithm does not converge to the correct time delays, whereas both the adaptive prewhitening and the adaptive GEVD algorithm converge to the correct time delays. The adaptive GEVD algorithm exhibits a better and faster convergence than the adaptive prewhitening algorithm. Figure 9 shows the TDE



(a)

(b)

(c)

Figure 9: TDE convergence plots of (a) adaptive EVD algorithm ($\mu = 1e - 2$), (b) adaptive prewhitening algorithm ($\mu = 1e - 4$), and (c) adaptive GEVD algorithm ($\mu = 1e - 2$) with $N = 3$, $K = 40$, SNR $= 0$ dB, $T_{60} = 250$ milliseconds, and subsampling = 10. The TDE between microphones 1 and 2 is denoted by the solid line, the TDE between microphones 1 and 3 by the dotted line, and the TDE between microphones 2 and 3 by the thick solid line.

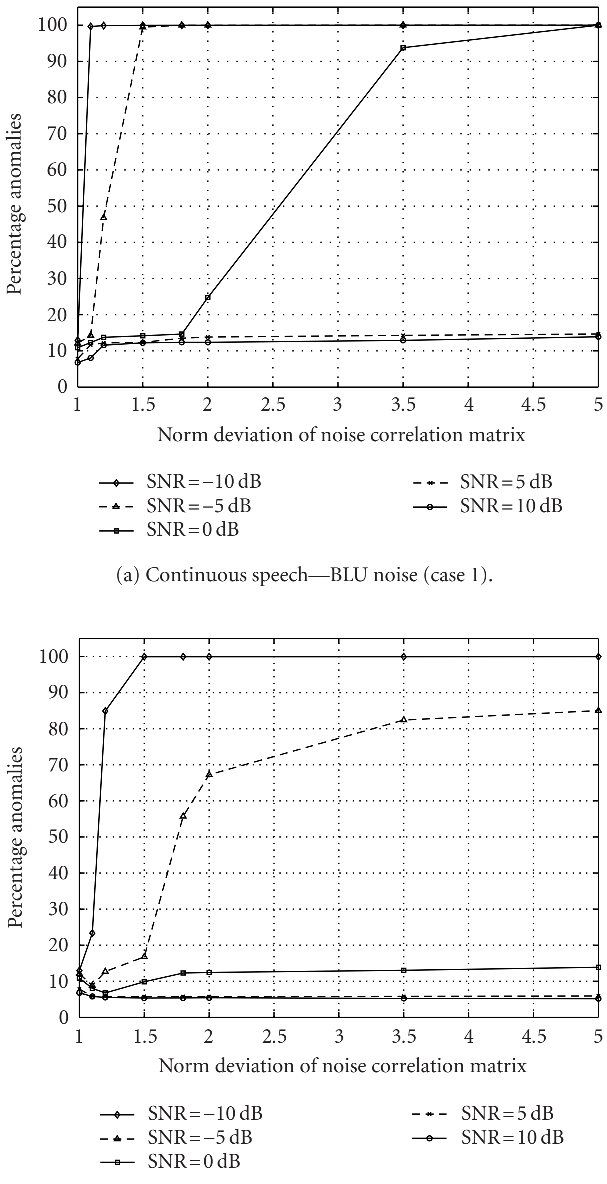
convergence plots for SNR $= 0$ dB. In this case, all algorithms converge to the correct time delays, although the time delay between the second and the third microphone signal is only correctly estimated by the adaptive EVD algorithm in signal segments with a high segmental SNR.

From these simulations, we can conclude that, for all SNRs and microphone configurations, the adaptive prewhitening and the adaptive GEVD algorithms converge more robustly to the correct time delays than the adaptive EVD algorithm, certainly in low SNR scenarios.

5.4. Sensitivity to the accuracy of the noise correlation matrix estimate

In the previous simulations, we have always assumed that an accurate estimate of the noise correlation matrix $\mathbf{R}_K^v$ is available. Since it is well known that GEVD-based algorithms may be sensitive to the accuracy of this noise correlation matrix estimate, we will analyze the sensitivity of the adaptive GEVD algorithm in this section. Instead of using the (correct) noise correlation matrix estimate $\mathbf{R}_K^v$, we will use

$$\tilde{\mathbf{R}}_K^v = \mathbf{R}_K^v + \alpha \mathbf{R}_K^e, \qquad (42)$$



(a) Continuous speech—BLU noise (case 1).

(b) Continuous speech—BLU noise (case 2).

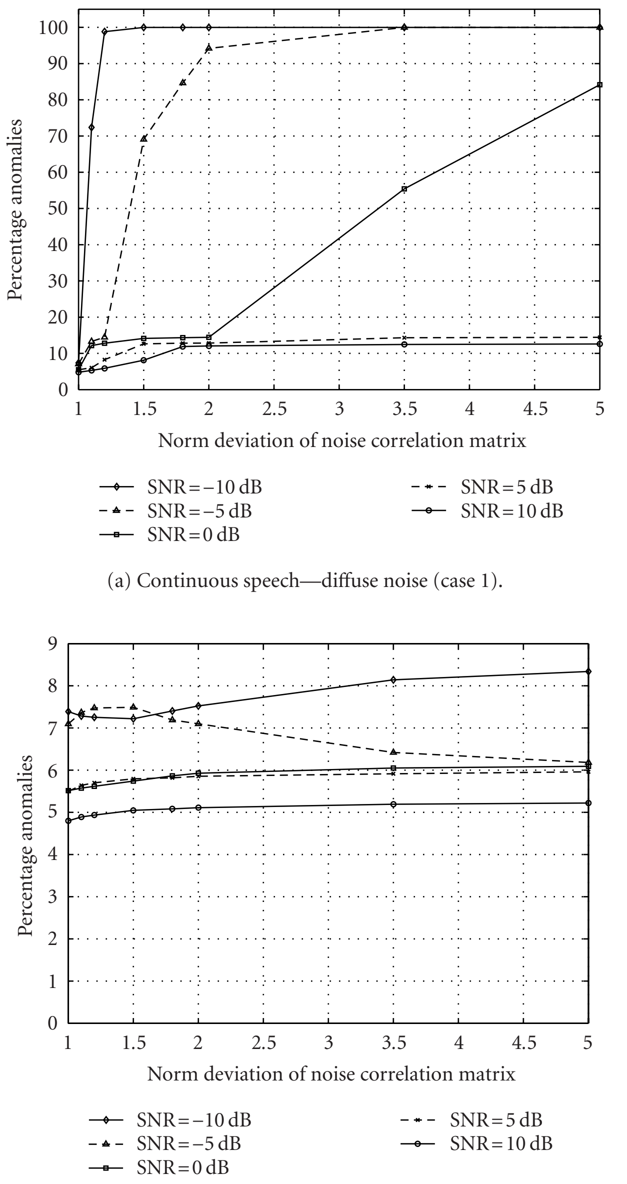
Figure 10: Sensitivity of adaptive GEVD algorithm with respect to noise correlation matrix estimate for (a) random deviation and (b) uncorrelated white noise deviation (localized noise source, $N = 2$, $K = 40$, $T_{60} = 250$ milliseconds, and subsampling = 1).

with $\mathbf{R}_K^e$ the deviation correlation matrix. We will consider two cases for $\mathbf{R}_K^e$:

(1) $\mathbf{R}_K^e$ is a random (symmetric) matrix corresponding to random errors on all correlation coefficients;

(2) $\mathbf{R}_K^e$ is equal to the identity matrix corresponding to uncorrelated white noise on the microphones.

The degree of deviation is determined by the norm deviation factor β, which is defined as



(a) Continuous speech—diffuse noise (case 1).

(b) Continuous speech—diffuse noise (case 2).

Figure 11: Sensitivity of adaptive GEVD algorithm with respect to noise correlation matrix estimate for (a) random deviation and (b) uncorrelated white noise deviation (diffuse noise source, $N = 2$, $K = 40$, $T_{60} = 250$ milliseconds, and subsampling = 1).

$$\beta = \frac{\|\tilde{\mathbf{R}}_K^v\|_2}{\|\mathbf{R}_K^v\|_2}. \tag{43}$$

For the localized noise source, Figure 10a shows the sensitivity of the adaptive GEVD algorithm for different SNRs when $\mathbf{R}_K^e$ is a random matrix, whereas Figure 10b shows the sensitivity when $\mathbf{R}_K^e$ is equal to the identity matrix. As can be seen, the adaptive GEVD algorithm is more sensitive to the accuracy of the noise correlation matrix estimate for low SNR scenarios and when $\mathbf{R}_K^e$ is a random matrix.

Figure 11 shows the sensitivity of the adaptive GEVD algorithm for a diffuse noise source. As can be seen from Figure 11a, when $\mathbf{R}_K^e$ is a random matrix, the sensitivity for a diffuse noise source is comparable to the sensitivity for a localized noise source. However, as can be seen from Figure 11b, for a diffuse noise source, the adaptive GEVD algorithm is not very sensitive when $\mathbf{R}_K^e$ is equal to the identity matrix. This can be explained by the fact that, for a large microphone distance, the noise correlation matrix $\mathbf{R}_K^v$ for a diffuse noise source is approximately equal to the identity matrix.

6. CONCLUSION

In this paper, we have presented two adaptive algorithms for robust TDE in adverse acoustic environments where a large amount of reverberation and additive noise is present. We have extended a recently developed adaptive EVD algorithm for TDE to noisy environments by using an adaptive GEVD or by prewhitening the microphone signals. For the adaptive GEVD, we have derived a stochastic gradient algorithm which iteratively estimates the generalized eigenvector corresponding to the smallest generalized eigenvalue. In addition, we have extended all presented TDE algorithms to the case of more than two microphones. It has been shown by simulations that, for all considered scenarios, the time delays can be estimated more accurately using the adaptive prewhitening and the adaptive GEVD algorithms than using the adaptive EVD algorithm. However, the difference in performance between the adaptive EVD and GEVD algorithms is negligible for SNRs higher than 5 dB and for a diffuse noise source, and the adaptive GEVD algorithm is quite sensitive to the accuracy of the noise correlation matrix estimate for low SNR scenarios.

ACKNOWLEDGMENTS

This work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, and was supported in part by the FWO Research Project G.0233.01, *Signal Processing and Automatic Patient Fitting for Advanced Auditory Prostheses*, the IWT Project 020540, *Performance Improvement of Cochlear Implants by Innovative Speech Processing Algorithms*, the IWT Project 020476, *Sound Management System for Public Address systems (SMS4PA)*, the Concerted Research Action, *Mathematical Engineering Techniques for Information and Communication Systems* (GOA-MEFISTO-666) of the Flemish Government, and the Interuniversity Attraction Pole IUAP P5-22 (2002-2007), *Dynamical Systems and Control: Computation, Identification, and Modelling*, initiated by the Belgian State, Prime Minister's Office, Federal Office for Scientific, Technical, and Cultural Affairs, and was partially sponsored by Cochlear. The authors would like to thank the reviewers for their valuable comments and suggestions.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. Signal Processing*, vol. 51, no. 10, pp. 2511–2526, 2003.

[3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 187–190, Munich, Germany, April 1997.

[4] 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., vol. 551 of *Kluwer International Series in Engineering and Computer Science*, chapter 11, pp. 239–259, Kluwer Academic Press, Boston, Mass, USA, March 2000.

[5] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1210–1218, 1983.

[6] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[7] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

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

[9] D. V. Rabinkin, R. J. Renomeron, A. J. Dahl, J. C. French, J. L. Flanagan, and M. H. Bianchi, "A DSP implementation of source location using microphone arrays," in *Proc. SPIE*, vol. 2846, pp. 88–99, Denver, Colo, USA, August 1996.

[10] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[11] P. C. Ching, Y. T. Chan, and K. C. Ho, "Constrained adaptation for time delay estimation with multipath propagation," *IEE Proceedings Part F: Radar and Signal Processing*, vol. 138, no. 5, pp. 453–458, 1991.

[12] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[13] L. Tong, G. Xu, and T. Kailath, "Fast blind equalization via antenna arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, pp. 272–275, Minneapolis, Minn, USA, April 1993.

[14] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC '01)*, pp. 47–50, Darmstadt, Germany, September 2001.

[15] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing*, vol. 43, no. 2, pp. 516–525, 1995.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.

[17] F. T. Luk, "A parallel method for computing the generalized singular value decomposition," *Journal of Parallel and Distributed Computing*, vol. 2, no. 3, pp. 250–260, 1985.

[18] J. Wouters, W. Damman, and A. J. Bosman, "Vlaamse opname van woordenlijsten voor spraakaudiometrie," *Logopedie*, vol. 7, no. 6, pp. 28–34, 1994.

[19] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[20] M. Moonen, P. Van Dooren, and J. Vandewalle, "A singular value decomposition updating algorithm for subspace tracking," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 4, pp. 1015–1038, 1992.

[21] P. A. Regalia, "An adaptive unit norm filter with applications to signal analysis and Karhunen-Loève transformations," *IEEE Trans. Circuits and Systems*, vol. 37, no. 5, pp. 646–649, 1990.

[22] M. Moonen, P. Van Dooren, and J. Vandewalle, "A systolic algorithm for QSVD updating," *Signal Processing*, vol. 25, no. 2, pp. 203–213, 1991.

[23] S. Roy and J. J. Shynk, "Analysis of the momentum LMS algorithm," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2088–2098, 1990.

[24] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, no. 8, pp. 1127–1138, 2002.

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

[26] F. A. Everest, *The Master Handbook of Acoustics*, McGraw-Hill, New York, NY, USA, 4th edition, 2001.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[28] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, 1986.

[29] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds., vol. 551 of *Kluwer International Series in Engineering and Computer Science*, chapter 9, pp. 155–178, Kluwer Academic, Boston, Mass, USA, March 2000.

[30] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[31] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[32] A. Spriet, M. Moonen, and J. Wouters, "A multichannel subband generalized singular value decomposition approach to speech enhancement," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 149–158, 2002, Special Issue on Acoustic Echo and Noise Control.

[33] G. Rombouts and M. Moonen, "QRD-based unconstrained optimal filtering for acoustic noise reduction," *Signal Processing*, vol. 83, no. 9, pp. 1889–1904, 2003.

Simon Doclo was born in Wilrijk, Belgium, in 1974. He received the M.S. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. Currently he is a Postdoctoral Researcher at the Electrical Engineering Department, KU Leuven. His research interests are in microphone array processing for acoustic noise reduction, dereverberation and sound localization, adaptive filtering, speech enhancement, and hearing aid technology. Dr. Doclo received the first prize "KVIV-Studentenprijzen" (with Erik De Clippel) for his

M.S. thesis in 1997 and he received a Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001. He has been the Secretary of the IEEE Benelux Signal Processing Chapter (1997–2002).

Marc Moonen received the Electrical Engineering degree and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1986 and 1990, respectively. Since 2000, he has been an Associate Professor at the Electrical Engineering Department of Katholieke Universiteit Leuven, where he is currently heading a research team of sixteen Ph.D. candidates and postdocs, working in the area of signal processing for digital communications, wireless communications, DSL, and audio signal processing. He received the 1994 KULeuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), and was a 1997 "Laureate of the Belgium Royal Academy of Science." He was the Chairman of the IEEE Benelux Signal Processing Chapter (1998–2002), and is currently a EURASIP AdCom Member (European Association for Signal, Speech, and Image Processing, 2000). He is Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003), and a member of the editorial board of Integration, the VLSI Journal, IEEE Transactions on Circuits and Systems II, and IEEE Signal Processing Magazine.