ROBUST TIME-DELAY ESTIMATION IN HIGHLY ADVERSE ACOUSTIC ENVIRONMENTS

Simon Doclo, Marc Moonen

Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT-SISTA),
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
{simon.doclo,marc.moonen}@esat.kuleuven.ac.be

ABSTRACT

This paper describes an algorithm for robust time-delay estimation (TDE) in situations where a large amount of additive noise and reverberation is present. In [1] an adaptive eigenvalue decomposition algorithm has been developed for TDE between two microphones in highly reverberant acoustic environments. In this paper we extend this algorithm to highly noisy and reverberant environments, by using a generalized eigenvalue decomposition or by prewhitening the noisy microphone signals. It is shown that time-delays can be robustly estimated for SNRs down to -5 dB.

1. INTRODUCTION

In many speech communication applications, such as teleconferencing, hands-free voice-controlled systems and hearing aids, it is desirable to localize the dominant speaker. In teleconferencing applications the speaker position has to be known for correct camera steering, while in other applications the speaker position is needed for microphone array beamforming in order to suppress acoustic noise and reverberation.
By estimating the time-delays between the microphone signals of a microphone array, it is possible to accurately estimate the speaker position [2]. However, time-delay estimation is not an easy task because of the non-stationary character of the speech signals, room reverberation and acoustic background noise. Generally, room reverberation is considered to be the main problem for TDE [3], but also acoustic background noise can considerably decrease the performance of time-delay estimators. While highly noisy situations are not very common in typical teleconferencing applications, they frequently occur in *e.g.* hearing aid applications.
Most TDE methods are based on a generalized cross-correlation (GCC) measure between the microphone signals [4][5]. However, since most of these methods assume an ideal room model without reverberation, *i.e.* only a direct path between the source and the microphone array, they cannot handle reverberation very well. To make TDE more robust to room reverberation, a cepstral prefiltering technique has been proposed [6] and techniques have been developed which use a more realistic room model with reverberation [1][7]. In [1] an eigenvalue decomposition algorithm has been developed to estimate the room impulse responses. This algorithm performs much better for highly reverberant rooms than GCC-based methods. However this algorithm is only optimal if no noise or only spatio-temporally white noise is present. In this paper we extend the eigenvalue decomposition algorithm to the colored noise case, by using a generalized eigenvalue decomposition or by prewhitening the microphone signals.
In section 2 it is shown that if the length of the room impulse responses is known or can be overestimated, the total room impulse responses can be exactly identified from the noise subspace of the speech and noise correlation matrix. In practice this subspace is computed with the generalized singular value decomposition (GSVD) of a speech and noise data matrix or the singular value decomposition (SVD) of a prewhitened speech data matrix.
Since for TDE only the time delay between the first peak (direct path) of the impulse responses is required, it is not necessary to estimate the complete impulse responses. In [1] an adaptive procedure is presented for estimating the time-delay between the peaks of two impulse responses. In section 3 this adaptive procedure will be extended to the colored noise case.
Section 4 describes the simulation results, where the performance and convergence speed of the algorithms is compared for different signal-to-noise ratios (SNR). It is shown that time-delays can be robustly estimated for SNRs down to -5 dB and that the convergence time is dependent on the SNR.

2. ESTIMATION OF ROOM IMPULSE RESPONSES

Consider M microphones where each microphone signal $x_m[k]$, $m = 0 \dots M-1$, consists of a filtered version of the clean speech signal $s[k]$ and some additive noise,

$$x_m[k] = s_m[k] + n_m[k] = h_m[k] \otimes s[k] + n_m[k] \ , \quad (1)$$

with $s_m[k]$ and $n_m[k]$ the speech and noise component received at the mth microphone at time k and $h_m[k]$ the room impulse response between the speech source and the mth microphone. The goal is to estimate $h_m[k]$ from $x_m[k]$ without any knowledge of $s[k]$. Knowing the complete room impulse responses, it is trivial to compute the time-delays between the microphone signals. If the room impulse responses have length L, then

$$\mathbf{s}_{i,L}^T[k]\ \mathbf{h}_j = \mathbf{s}_{j,L}^T[k]\ \mathbf{h}_i \ , \quad (2)$$

with

$$\mathbf{s}_{m,L}^T[k] = \left[\ s_m[k] \quad s_m[k-1] \quad \dots \quad s_m[k-L+1]\ \right] \quad (3)$$

$$\mathbf{h}_m^T = \left[\ h_m[0] \quad h_m[1] \quad \dots \quad h_m[L-1]\ \right] . \quad (4)$$

Although we do not explicitly attribute a time index k to the impulse responses, this does not mean that they cannot be time-variant. In the following we will assume $M = 2$, although it is quite easy to extend the algorithms to the case of more than two microphones.

Simon Doclo is a Research Assistant supported by I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). This research work was carried out in the frame of the Interuniversity Attraction Pole IUAP-P4/02 *Modeling, Identification, Simulation and Control of Complex Systems*, the F.W.O. Research Project *Signal Processing and Automatic Patient-Adaptation for Advanced Hearing Aids* (G.0233.01) and the Concerted Research Action *Mathematical Engineering Techniques for Information and Communication Systems* (GOA-MEFISTO-666).

2.1. Noiseless case

The $2K \times 2K$ clean speech correlation matrix $\mathbf{R}_K^s$ is defined as

$$\mathbf{R}_K^s = \begin{bmatrix} \mathbf{R}_{22,K}^s & -\mathbf{R}_{21,K}^s \\ -\mathbf{R}_{12,K}^s & \mathbf{R}_{11,K}^s \end{bmatrix} , \quad (5)$$

with

$$\mathbf{R}_{ij,K}^s = \mathcal{E}\{\mathbf{s}_{i,K}[k]\ \mathbf{s}_{j,K}^T[k]\} . \quad (6)$$

If $K \geq L$ and the impulse responses $\mathbf{h}_1$ and $\mathbf{h}_2$ do not have common zeros and the autocorrelation matrix of the input signal $s[k]$ has full rank [8], then the clean speech correlation matrix $\mathbf{R}_K^s$ has rank $K+L-1$, such that it is rank-deficient and its null-space has dimension $K-L+1$.
If $K = L$, the null-space of $\mathbf{R}_L^s$ has dimension 1, and the vector

$$\mathbf{v} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} \quad (7)$$

belongs to this null-space because $\mathbf{R}_L^s \mathbf{v} = \mathbf{0}$. If we take the eigenvalue decomposition, $\mathbf{R}_L^s = \mathbf{V}_s \mathbf{\Delta}_s \mathbf{V}_s^T$, the unit-norm eigenvector, belonging to the only zero eigenvalue of $\mathbf{R}_L^s$, contains a scaled version of the two impulse responses, such that the time-delays can be exactly estimated.
If $K > L$, the null-space of $\mathbf{R}_K^s$ contains $K-L+1$ eigenvectors, which all contain a different filtered version of the impulse responses. By computing the QR-decomposition of the full null-space or by applying successive QR-decompositions to two eigenvectors in the null-space, the correct impulse responses of length L can be identified [9].

2.2. Spatio-temporally white noise

If additive noise is present, we can define the noisy speech correlation matrix $\mathbf{R}_L^x$ and the noise correlation matrix $\mathbf{R}_L^n$ similar as in (5). Assuming that the clean speech signal $s[k]$ and the noise components $n_m[k]$ are uncorrelated,

$$\mathbf{R}_L^x = \mathbf{R}_L^s + \mathbf{R}_L^n . \quad (8)$$

From the eigenvalue decomposition of the noisy speech correlation matrix, $\mathbf{R}_L^x = \mathbf{V}_x \mathbf{\Delta}_x \mathbf{V}_x^T$, the impulse responses can only be identified if the noise is spatio-temporally white, *i.e.* $\mathbf{R}_L^n = \sigma_n^2 \mathbf{I}$. Because in that case $\mathbf{R}_L^x = \mathbf{V}_x(\mathbf{\Delta}_s + \sigma_n^2 \mathbf{I})\mathbf{V}_x^T$, the eigenvector corresponding to the smallest eigenvalue σ_n^2 contains a scaled version of the impulse responses. Also for $K > L$, the procedure is similar to the procedure in the noiseless case.

2.3. Spatio-temporally colored noise

If spatio-temporally colored noise is present, the room impulse responses can still be identified from the generalized eigenvalue decomposition of $\mathbf{R}_L^x$ and $\mathbf{R}_L^n$ or from the eigenvalue decomposition of the prewhitened noisy speech correlation matrix. In both cases, $\mathbf{R}_L^n$ needs to be known or we have to able to estimate $\mathbf{R}_L^n$ from noise-only periods, *e.g.* using a voice activity detector (VAD). In the following we will also assume that $\mathbf{R}_L^n$ has full rank.

1. The generalized eigenvalue decomposition (GEVD) of $\mathbf{R}_L^x$ and $\mathbf{R}_L^n$ is defined as [10]

$$\begin{cases} \mathbf{R}_L^x &= \mathbf{Q}\ \mathbf{\Lambda}_x \mathbf{Q}^{\mathbf{T}} \\ \mathbf{R}_L^n &= \mathbf{Q}\ \mathbf{\Lambda}_n \mathbf{Q}^{\mathbf{T}} , \end{cases} \quad (9)$$

with $\mathbf{Q}$ an invertible, but not necessarily orthogonal matrix. From (8) and (9) it follows that

$$(\mathbf{R}_L^n)^{-1}\mathbf{R}_L^s = (\mathbf{R}_L^n)^{-1}(\mathbf{R}_L^x - \mathbf{R}_L^n) \quad (10)$$
$$= \mathbf{Q}^{-T}(\mathbf{\Lambda}_n^{-1}\mathbf{\Lambda}_x - \mathbf{I})\mathbf{Q}^T . \quad (11)$$

Since $(\mathbf{R}_L^n)^{-1}\mathbf{R}_L^s$ has rank $2L-1$, one (and only one) of the values of the diagonal matrix $\mathbf{\Lambda}_n^{-1}\mathbf{\Lambda}_x$ is equal to 1. Therefore a column $\mathbf{q}$ of $\mathbf{Q}^{-T}$ exists for which

$$(\mathbf{R}_L^n)^{-1}\mathbf{R}_L^s \mathbf{q} = \mathbf{0} , \quad (12)$$

such that $\mathbf{R}_L^s \mathbf{q} = \mathbf{0}$. Since the dimension of the null-space of $\mathbf{R}_L^s$ is 1, the vector $\mathbf{q}$ contains a scaled version of the impulse responses.

2. The prewhitened correlation matrix $\bar{\mathbf{R}}_L^x$ is defined as

$$\bar{\mathbf{R}}_L^x \triangleq (\mathbf{R}_L^n)^{-T/2}\mathbf{R}_L^x(\mathbf{R}_L^n)^{-1/2} , \quad (13)$$

with $(\mathbf{R}_L^n)^{1/2}$ the Cholesky-factor of the noise correlation matrix $\mathbf{R}_L^n$, such that $\mathbf{R}_L^n = (\mathbf{R}_L^n)^{T/2}(\mathbf{R}_L^n)^{1/2}$. From the eigenvalue decomposition of $\bar{\mathbf{R}}_L^x$,

$$\bar{\mathbf{R}}_L^x = \bar{\mathbf{V}}_x \bar{\mathbf{\Lambda}}_x \bar{\mathbf{V}}_x^T , \quad (14)$$

it follows that $\bar{\mathbf{R}}_L^s$ can be written as

$$\bar{\mathbf{R}}_L^s \triangleq (\mathbf{R}_L^n)^{-T/2}\mathbf{R}_L^s(\mathbf{R}_L^n)^{-1/2} = \bar{\mathbf{V}}_x(\bar{\mathbf{\Lambda}}_x - \mathbf{I})\bar{\mathbf{V}}_x^T .$$

Since $\bar{\mathbf{R}}_L^s$ has rank $2L-1$, one of the values of the diagonal matrix $\bar{\mathbf{\Lambda}}_x$ is 1 and a column $\bar{\mathbf{v}}$ of $\bar{\mathbf{V}}_x$ exists for which

$$\bar{\mathbf{R}}_L^s \bar{\mathbf{v}} = (\mathbf{R}_L^n)^{-T/2}\mathbf{R}_L^s(\mathbf{R}_L^n)^{-1/2}\bar{\mathbf{v}} = \mathbf{0} , \quad (15)$$

such that $\mathbf{R}_L^s(\mathbf{R}_L^n)^{-1/2}\bar{\mathbf{v}} = \mathbf{0}$. Since the dimension of the null-space of $\mathbf{R}_L^s$ is 1, the vector $(\mathbf{R}_L^n)^{-1/2}\bar{\mathbf{v}}$ contains a scaled version of the impulse responses.

In fact, both algorithms are equivalent, since

$$\bar{\mathbf{\Lambda}}_x = \mathbf{\Lambda}_n^{-1}\mathbf{\Lambda}_x, \quad \mathbf{Q}^{-T} = (\mathbf{R}_L^n)^{-1/2}\bar{\mathbf{V}}_x . \quad (16)$$

However the adaptive versions of the algorithms, which will be used for practical TDE and which are defined in section 3, can produce different results.
Also if $K > L$, the procedure for estimating the impulse responses of length L is similar to the procedure in the noiseless case.

2.4. Practical computation

In practice we do not work with correlation matrices, but with data matrices. The $p \times 2L$ speech data matrix $\mathbf{X}_L[k]$ is defined as

$$\mathbf{X}_L[k] = \begin{bmatrix} \mathbf{x}_{2,L}^T[k] & -\mathbf{x}_{1,L}^T[k] \\ \mathbf{x}_{2,L}^T[k+1] & -\mathbf{x}_{1,L}^T[k+1] \\ \vdots & \vdots \\ \mathbf{x}_{2,L}^T[k+p-1] & -\mathbf{x}_{1,L}^T[k+p-1] \end{bmatrix} , \quad (17)$$

such that the empirical correlation matrix $\mathbf{R}_L^x \simeq \mathbf{X}_L[k]^T\mathbf{X}_L[k]/p$. The noise data matrix $\mathbf{N}_L[k]$ is similarly defined.

1. *GSVD-procedure.* Instead of computing the GEVD of $\mathbf{R}_L^x$ and $\mathbf{R}_L^n$, we compute the generalized singular value decomposition (GSVD) of $\mathbf{X}_L[k]$ and $\mathbf{N}_L[k]$, defined as [10]

$$\begin{cases} \mathbf{X}_L[k] &= \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{Q}^T \\ \mathbf{N}_L[k] &= \mathbf{U}_n \mathbf{\Sigma}_n \mathbf{Q}^T . \end{cases} \quad (18)$$

2. *Prewhitening-procedure.* The prewhitened matrix $\bar{\mathbf{X}}_L[k]$ is

$$\bar{\mathbf{X}}_L[k] = \mathbf{X}_L[k](\mathbf{R}_L^n)^{-1/2} \quad (19)$$

where the Cholesky-factor $(\mathbf{R}_L^n)^{1/2}$ is computed by the QR-decomposition of the noise matrix, $\mathbf{N}_L[k] = \mathbf{Q}_n(\mathbf{R}_L^n)^{1/2}$. The singular value decomposition of $\bar{\mathbf{X}}_L[k]$ is defined as

$$\bar{\mathbf{X}}_L[k] = \bar{\mathbf{U}}_x \bar{\mathbf{\Sigma}}_x \bar{\mathbf{V}}_x^T . \quad (20)$$

2.5. Simulation results

In our simulations we have filtered a 16 kHz speech segment of 160000 samples with 2 impulse responses ($L = 20$), which are depicted in figure 1a. A stationary speech-like, *i.e.* with the same long-term spectral characteristics as speech, noise signal has been added and the SNR of the microphone signals is 10 dB.
Figure 1b and 1c show the estimated impulse responses ($K = 20$), for the SVD-procedure (assuming no noise) and the GSVD (or prewhitening) procedure. As can be clearly seen, the impulse responses are almost correctly estimated with the GSVD-procedure, unlike the SVD-procedure. Because the assumption of uncorrelated speech and noise segments is not always completely satisfied, *i.e.* $\mathbf{X}_L^T[k]\mathbf{N}_L[k] \simeq \mathbf{0}$, small estimation errors occur in the GSVD-procedure. In our simulations we noticed that the better this assumption is satisfied, *i.e.* the higher the SNR and the longer the speech and noise segments, the smaller the estimation error is.

3. ADAPTIVE PROCEDURE FOR TDE

In practice, room impulse responses can have thousands of taps. Because of the correlated nature of speech, autocorrelation matrices of the input signal $s[k]$ of these dimensions will be rank-deficient. Therefore it is impossible to identify the complete room impulse responses in practice. If we underestimate the length of the impulse responses ($K < L$), the estimated impulse responses are biased and do not necessarily exhibit any resemblance to the actual impulse responses, making it difficult (and practically impossible) to estimate the correct time-delays.
However, in [1] it has been shown that by using an adaptive eigenvalue decomposition algorithm, it is still possible to identify the main peak in the impulse responses, even when underestimating the length of the impulse responses. For TDE only this time-delay between the first peak of the impulse responses is required.
The procedure iteratively estimates the eigenvector of $\mathbf{R}_K^x$ corresponding to the smallest eigenvalue by minimizing $\mathbf{v}^T\mathbf{R}_K^x\mathbf{v}$, subject to the constraint $\mathbf{v}^T\mathbf{v} = 1$. The problem is solved by minimizing the mean square value of the error signal $e[k]$,

$$e[k] = \frac{\mathbf{v}^T\mathbf{x}_K[k]}{\|\mathbf{v}\|} , \tag{21}$$

with $\mathbf{x}_K[k] = \begin{bmatrix} \mathbf{x}_{2,K}^T[k] & -\mathbf{x}_{1,K}^T[k] \end{bmatrix}^T$. This can be done using a gradient-descent constrained LMS-procedure:

$$\mathbf{v}[k+1] = \frac{\mathbf{v}[k] - \mu e[k]\frac{\partial e[k]}{\partial \mathbf{v}[k]}}{\|\mathbf{v}[k] - \mu e[k]\frac{\partial e[k]}{\partial \mathbf{v}[k]}\|} , \tag{22}$$

$$\frac{\partial e[k]}{\partial \mathbf{v}[k]} = \frac{1}{\|\mathbf{v}[k]\|}\left\{\mathbf{x}_K[k] - e[k]\frac{\mathbf{v}[k]}{\|\mathbf{v}[k]\|}\right\} . \tag{23}$$

Since the smallest eigenvalue of $\mathbf{R}_K^x$ is assumed to be zero and normalization is included in each iteration, the gradient eventually reduces to $\frac{\partial e[k]}{\partial \mathbf{v}[k]} \simeq \mathbf{x}_K[k]$, such that the update formula becomes

$$\mathbf{v}[k+1] = \frac{\mathbf{v}[k] - \mu e[k]\mathbf{x}_K[k]}{\|\mathbf{v}[k] - \mu e[k]\mathbf{x}_K[k]\|} . \tag{24}$$

In [1] it is also indicated that initialization of $\mathbf{v}$ and the choice of the parameters K and μ are quite important for a good convergence behavior. The time-delay is calculated as the difference between the main peaks in the two estimated impulse responses or as the peak of the correlation function between the two impulse responses. It is also shown that this algorithm performs more robustly in highly reverberant rooms than GCC-based methods.

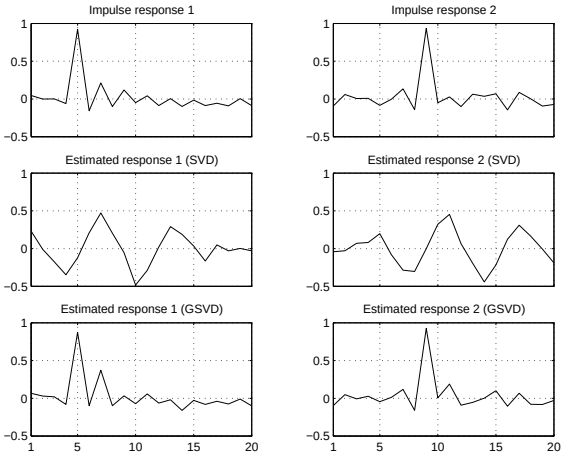

Figure 1: (**a**) Impulse responses $\mathbf{h}_1$ and $\mathbf{h}_2$, (**b**) Estimated impulse responses with SVD-procedure and (**c**) GSVD-procedure

For the noise-robust algorithms, described in sections 2.3 and 2.4, it is also possible to derive adaptive versions. In the simulations it will be shown that the adaptive version of the GSVD-procedure and the prewhitening-procedure can produce different results.

1. For the *GEVD-procedure*, we need to iteratively estimate the generalized eigenvector of $\mathbf{R}_K^x$ and $\mathbf{R}_K^n$ corresponding to the smallest generalized eigenvalue by minimizing the cost function $\mathbf{q}^T\mathbf{R}_K^x\mathbf{q}$, subject to $\mathbf{q}^T\mathbf{R}_K^n\mathbf{q} = 1$. This problem can be solved by minimizing the mean square value of the error signal $e[k]$,

$$e[k] = \frac{\mathbf{q}^T\mathbf{x}_K[k]}{\sqrt{\mathbf{q}^T\mathbf{R}_K^n\mathbf{q}}} = \frac{\mathbf{q}^T\mathbf{x}_K[k]}{\|(\mathbf{R}_K^n)^{1/2}\mathbf{q}\|} . \tag{25}$$

   The gradient now becomes

$$\frac{\partial e[k]}{\partial \mathbf{q}[k]} = \frac{1}{\sqrt{\mathbf{q}^T[k]\mathbf{R}_K^n\mathbf{q}[k]}}\left\{\mathbf{x}_K[k] - e[k]\frac{\mathbf{R}_K^n\mathbf{q}[k]}{\sqrt{\mathbf{q}^T[k]\mathbf{R}_K^n\mathbf{q}[k]}}\right\} .$$

   Since the smallest generalized eigenvalue is 1, we cannot further simplify this expression. To avoid roundoff error propagation, we include a normalization step in each iteration, such that the update formula can be written as

$$\begin{aligned} \tilde{\mathbf{q}}[k+1] &= \mathbf{q}[k] - \mu e[k]\{\mathbf{x}_K[k] - e[k]\mathbf{R}_K^n\mathbf{q}[k]\} \\ \mathbf{q}[k+1] &= \frac{\tilde{\mathbf{q}}[k+1]}{\sqrt{\tilde{\mathbf{q}}^T[k+1]\mathbf{R}_K^n\tilde{\mathbf{q}}[k+1]}} \end{aligned} \tag{26}$$

2. The *prewhitening-procedure* can be made adaptive by using prewhitened speech data vectors $\bar{\mathbf{x}}_K[k] = \mathbf{x}_K[k](\mathbf{R}_K^n)^{-1/2}$. The update formula now becomes

$$\bar{\mathbf{v}}[k+1] = \frac{\bar{\mathbf{v}}[k] - \mu e[k]\{\bar{\mathbf{x}}_K[k] - e[k]\bar{\mathbf{v}}[k]\}}{\|\bar{\mathbf{v}}[k] - \mu e[k]\{\bar{\mathbf{x}}_K[k] - e[k]\bar{\mathbf{v}}[k]\}\|} , \tag{27}$$

   and the actual impulse response is $\mathbf{v}[k] = (\mathbf{R}_K^n)^{-1/2}\bar{\mathbf{v}}[k]$. As indicated in section 2.4, the Cholesky-factor $(\mathbf{R}_K^n)^{-1/2}$ can be updated during noise periods by inverse QR-updating.

4. SIMULATIONS

In our simulations we have filtered a 16 kHz speech signal of 80000 samples and a speech-like noise signal with two room impulse responses ($L = 2000$), constructed with the image method

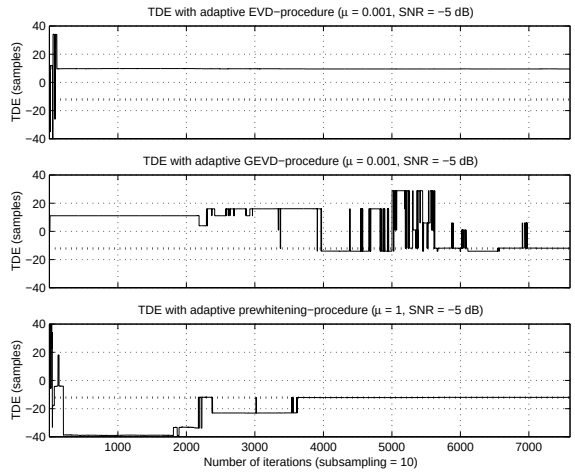

Figure 2: Convergence plots of adaptive EVD, GEVD and prewhitening-procedure (SNR = -5 dB, subsampling = 10)

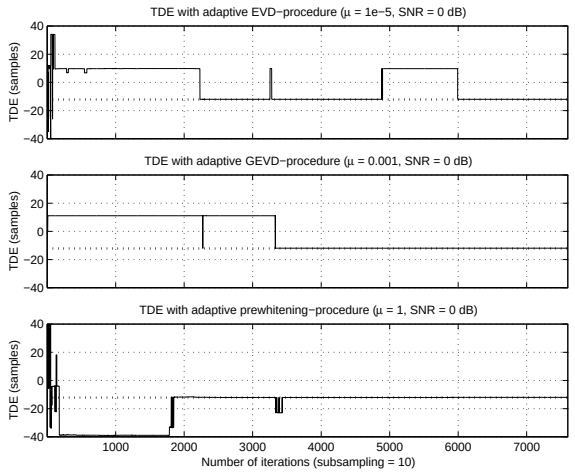

Figure 3: Convergence plots of adaptive EVD, GEVD and prewhitening-procedure (SNR = 0 dB, subsampling = 10)

[11]. The room dimensions are $5\,m \times 4\,m \times 2\,m$, the positions of the 2 omni-directional microphones are $\left[\begin{array}{ccc} 1 & 1 & 1 \end{array}\right]$ and $\left[\begin{array}{ccc} 1 & 1 & 1.5 \end{array}\right]$, the speech source position is $\left[\begin{array}{ccc} 2 & 2 & 1.7 \end{array}\right]$ and the noise source position is $\left[\begin{array}{ccc} 1 & 1.5 & 1 \end{array}\right]$. The correct time-delay for the speech source is -12.183 samples (and 9.746 samples for the noise source) and is indicated as a dotted line in the figures. The reverberation time T_{60} of the room is $250\,\mathrm{ms}$. We have performed simulations at different SNRs for the different algorithms (adaptive EVD, GEVD and prewhitening-procedure). The filterlength K is 100 and for each algorithm we have chosen the stepsize μ which gave the best results. The subsampling factor is 10, *i.e.* only every 10 samples a new iteration is performed.

Figure 2 shows the convergence plots of the time-delays if the SNR is -5 dB. As can be seen the adaptive EVD-procedure does not converge to the speech time-delay, but to the noise time-delay. The adaptive GEVD-procedure has an irregular behavior, but converges to the correct time-delay. The prewhitening-procedure is the procedure which converges fastest to the correct time-delay.

Figure 3 shows the convergence plots of the time-delays if the SNR is 0 dB. All procedures now converge to the correct time-delay, but the adaptive prewhitening and adaptive GEVD-procedure are somewhat faster than the adaptive EVD-procedure. Note that is quite remarkable that the adaptive EVD-procedure converges to the correct time-delay at an SNR of 0 dB, without any knowledge of the noise characteristics.

From our simulations, we conclude that the adaptive prewhitening-procedure is the most robust procedure in additive noise and that it converges to the correct time-delay for SNRs down to -5 dB. The convergence time is dependent of the SNR (and of the parameters of the algorithm), but for most low-SNR scenarios the convergence time lies between 1 and 2 sec.

5. CONCLUSION

In this paper we have described an algorithm for robust time-delay estimation in adverse acoustic situations with a large amount of reverberation and additive noise. We have extended an adaptive EVD-algorithm for time-delay estimation to noisy environments, by using a GEVD or by prewhitening the microphone signals. Simulations show that the adaptive prewhitening-algorithm is the most robust algorithm for time-delay estimation in additive noise.

6. REFERENCES

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

[2] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech, Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

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

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

[5] P. G. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 1997.

[6] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for time delay estimation under reverberant conditions," *Signal Processing*, vol. 59, pp. 253–266, 1997.

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

[8] 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, Feb. 1995.

[9] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, MD : John Hopkins University Press, Baltimore, 3rd edition, 1996.

[11] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoust. Soc. of America*, vol. 65, pp. 943–950, Apr. 1979.