COMBINED ACOUSTIC ECHO AND NOISE REDUCTION USING GSVD-BASED OPTIMAL FILTERING

Simon Doclo, Marc Moonen

K.U.Leuven, Dept. of Elec. Engineering, SISTA
K. Mercierlaan 94, 3001 Leuven, Belgium
{doclo,moonen}@esat.kuleuven.ac.be

Erik De Clippel

Philips ITCL
Interleuvenlaan 74-76, 3001 Leuven, Belgium
erik.de.clippel@philips.com

ABSTRACT

This paper describes two schemes for combining acoustic echo and noise reduction using a GSVD-based optimal filtering technique. The GSVD-based filtering technique is a signal enhancement technique which has recently been proposed for noise reduction in multi-microphone speech signals. In many speech communication applications however also a far-end echo source is present. Therefore a combined echo and noise reduction scheme is needed.

The first scheme combines a standard multi-channel adaptive echo canceller with the GSVD-based noise reduction technique. The second scheme incorporates the far-end echo reference directly into the GSVD-based signal enhancement technique without cancelling the echo in every microphone signal. The two different schemes are compared with regard to performance and computational complexity.

1. INTRODUCTION

In many speech communication applications, like audio-conferencing and hands-free mobile telephony, the recorded multi-microphone speech signals are corrupted by acoustic background noise and echo signals (see figure 1). This causes a signal degradation which can lead to total unintelligibility of the speech and which degrades the performance of speech recognition and speech coding devices.

For cancelling far-end echo some well-known solutions (RLS, NLMS, APA, ...) exist. Also for reducing background noise multi-microphone signal enhancement techniques are available (*e.g.* fixed and adaptive beamforming techniques). Recently a GSVD-based optimal filtering technique for noise reduction in multi-microphone speech signals has been proposed, which has a better performance than classical beamforming techniques and which is more robust to deviations from the nominal situation [1][2]. This technique is briefly discussed in section 2. The used simulation environment is outlined in section 3.

Section 4 discusses the effect of adding a far-end echo source. Compared to the situation without echo source, the performance of the GSVD-based noise reduction technique drops considerably. However since the GSVD-based technique considers the echo source as an additional noise source, the echo signal itself will be reduced to some extent.

In order to improve the global performance a combined scheme for echo and noise reduction is needed, which has to be able to cancel the echo signal as well as obtain a noise reduction performance which is equally good as in the situation without echo source. Two different schemes are discussed in section 5. One scheme uses a multi-channel adaptive echo canceller, the other scheme incorporates the echo reference directly into the GSVD-based optimal filtering technique without cancelling the echo in every microphone signal. Simulations indicate that the scheme using a multi-channel adaptive echo canceller has a better performance and a lower computational complexity than the scheme incorporating the echo reference directly into the GSVD-based optimal filtering technique.

2. GSVD-BASED OPTIMAL FILTERING

The GSVD-based optimal filtering technique [1] considers problems where the observed signal vector $\mathbf{u}_k \in \mathbb{R}^N$ contains a signal-of-interest $\mathbf{s}_k \in \mathbb{R}^N$ (*e.g.* a speech signal) and an additive noise term $\mathbf{n}_k \in \mathbb{R}^N$, such that $\mathbf{u}_k = \mathbf{s}_k + \mathbf{n}_k$.

If we consider speech applications and use a robust speech-noise detection algorithm [3], noise-only observations can be made during speech pauses. Our goal is to reconstruct the signal-of-interest $\mathbf{s}_k$ from $\mathbf{u}_k$ by means of a linear filter $\mathbf{W} \in \mathbb{R}^{N\times N}$ using $\hat{\mathbf{s}}_k = \mathbf{u}_k^T\mathbf{W}$. It can be shown that using a MMSE-criterion the optimal filter $\mathbf{W}$ is equal to

$$\mathbf{W}_{WF} = \mathcal{E}\left\{\mathbf{u}_k \cdot \mathbf{u}_k^T\right\}^{-1}\left(\mathcal{E}\left\{\mathbf{u}_k \cdot \mathbf{u}_k^T\right\} - \mathcal{E}\left\{\mathbf{n}_k \cdot \mathbf{n}_k^T\right\}\right). \quad (1)$$

In practice this filter is computed by means of a generalized singular value decomposition (GSVD) [4] of a speech data matrix $\mathbf{U}_k \in \mathbb{R}^{p\times N}$ and a noise data matrix $\mathbf{N}_k \in \mathbb{R}^{q\times N}$,

$$\mathbf{U}_k = \begin{bmatrix} \mathbf{u}_k^T \\ \mathbf{u}_{k+1}^T \\ \vdots \\ \mathbf{u}_{k+p-1}^T \end{bmatrix} \qquad \mathbf{N}_k = \begin{bmatrix} \mathbf{n}_k^T \\ \mathbf{n}_{k+1}^T \\ \vdots \\ \mathbf{n}_{k+q-1}^T \end{bmatrix}. \quad (2)$$

Simon Doclo is a Research Assistant supported by the I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). Marc Moonen is a Research Associate with the F.W.O.-Vlaanderen (Fund for Scientific Research-Flanders). Erik De Clippel is a Project Engineer DSP working for Philips ITCL. This research work was carried out in the frame of the Concerted Research Action *Model-based Information Processing Systems* of the Flemish Government, the F.W.O. Research Project *Design and implementation of adaptive digital signal processing algorithms for broadband applications* and the IT-project *Multi-microphone Signal Enhancement Techniques for handsfree telephony and voice controlled systems (MUSETTE)* of the I.W.T., and was partially sponsored by Philips ITCL.

The GSVD of the matrices $\mathbf{U}_k$ and $\mathbf{N}_k$ is defined as

$$\left\{ \begin{array}{rcl} \mathbf{U}_k & = & U \cdot \mathrm{diag}\{\sigma_i\} \cdot X^T \\ \mathbf{N}_k & = & V \cdot \mathrm{diag}\{\eta_i\} \cdot X^T, \end{array} \right. \tag{3}$$

with U and V orthogonal matrices, X an invertible (but not necessarily orthogonal) matrix and $\frac{\sigma_i}{\eta_i}$ the generalized singular values. Substituting these formulas into (1) gives

$$\mathbf{W}_{WF} = X^{-T} \cdot \mathrm{diag}\{\frac{\sigma_i^2 - \eta_i^2}{\sigma_i^2}\} \cdot X^T. \tag{4}$$

We are also interested in the diagonal elements of the error covariance matrix $\mathcal{E}\left\{\mathbf{e}_k \cdot \mathbf{e}_k^T\right\}_{ii}$, with $\mathbf{e}_k = \mathbf{s}_k - \mathbf{u}_k^T \mathbf{W}$, since these elements indicate how well the i^{th} component of $\mathbf{s}_k$ is estimated. The smallest element on the diagonal corresponds to the best estimator. The best estimator, which is the corresponding column of $\mathbf{W}_{WF}$, will be denoted as $\mathbf{w}_{WF}^{min}$.

When considering M microphones where each microphone signal $m_j(k)$, $j = 1 \dots M$, consists of a filtered version of the speech signal and an additive noise term, the vector $\mathbf{u}_k \in \mathbb{R}^{MN}$ takes the form

$$\mathbf{u}_k = \left[\begin{array}{cccc} \mathbf{m}_{1k} & \mathbf{m}_{2k} & \dots & \mathbf{m}_{Mk} \end{array} \right]^T \tag{5}$$

$$\mathbf{m}_{jk} = \left[\begin{array}{cccc} m_j(k) & m_j(k-1) & \dots & m_j(k-N+1) \end{array} \right]. \tag{6}$$

The enhanced speech signal $\hat{s}(k)$ is then computed as

$$\hat{\mathbf{s}}(k) = \left[\begin{array}{cccc} \hat{s}(k) & \hat{s}(k+1) & \dots & \hat{s}(k+p-1) \end{array} \right]^T = \mathbf{U}_k \cdot \mathbf{w}_{WF}^{min}.$$

This can be considered a multi-channel filtering operation (see figure 5), where each of the M channels is filtered with an N-taps filter A_j, with $\mathbf{w}_{WF}^{min} = \left[\begin{array}{cccc} A_1^T & A_2^T & \dots & A_M^T \end{array} \right]^T$.

3. SIMULATION ENVIRONMENT

The used simulation environment is depicted in figure 1. Three different sources are present: the desired speech source $s(k)$, the background noise source $n(k)$ and the far-end echo source $f(k)$. Each microphone signal $m_j(k)$, $j = 1 \dots M$, consists of the sum of filtered versions of the speech, noise and echo signal,

$$m_j(k) = g_j^s(k) \otimes s(k) + g_j^n(k) \otimes n(k) + g_j^f(k) \otimes f(k), \tag{7}$$

with $g_j^s(k)$ the room impulse response between the speech source and the j^{th} microphone and $g_j^n(k)$ and $g_j^f(k)$ similarly defined for the noise and the echo source.

In our simulations we use $M = 5$ microphones. The room impulse responses are calculated using the image method [5], with a filterlength of 800 taps ($f_s = 8kHz$) and reflection coefficient $\alpha = 0.6$. The noise source is a white noise signal, the echo source is a music signal.

Since we are using simulations, the signal-to-noise ratio (SNR) and signal-to-echo ratio (SER) at each stage of the algorithms (see section 5) can be computed as

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum \tilde{s}^2(k)_{sp}}{\sum \tilde{n}^2(k)_{sp}}, \quad \mathrm{SER} = 10 \log_{10} \frac{\sum \tilde{s}^2(k)_{sp}}{\sum \tilde{f}^2(k)_{sp}}, \tag{8}$$

where $\tilde{s}(k)$, $\tilde{n}(k)$ and $\tilde{f}(k)$ correspond to the speech-, noise- and echo-related part of the considered signal during speech activity. The SNR of the corrupted first microphone signal $m_1(k)$ is 10.2 dB, the SER is 6.8 dB.

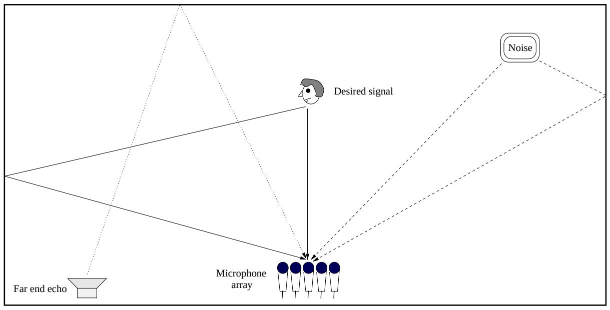

Figure 1: Simulation environment

4. EFFECT OF FAR-END ECHO SOURCE

In this section the effect of a far-end echo source on the GSVD-based noise reduction technique is discussed. First consider the situation with no echo source present (*i.e.* $f(k) = 0$). The SNR of the signal enhanced by the GSVD-based optimal filtering technique is depicted by the solid line in figure 3 in function of the filterlength N.

When a far-end echo source is present and we use the same GSVD-based technique, the SNR of the enhanced signal drops considerably. This is depicted in figure 3 by the dashed line ($L_m = 0$). Although less noise is reduced compared to the situation without echo, some of the echo itself has been reduced by the GSVD-based technique. This can be seen in figure 4 ($L_m = 0$), where the SER of the enhanced signal is plotted for different values of the filterlength N.

This can be explained by the fact that the GSVD-based optimal filtering technique just considers the echo source as an additional noise source. Because this procedure now has to cancel both noise and echo source instead of the one noise source, it is obvious that the noise reduction performance drops. On the other hand this shows that the noise reduction procedure also works for highly non-stationary signals (*e.g.* music), since it relies on the more stationary characteristics of the impulse responses between source and microphone array and not on the (highly non-stationary) signal characteristics.

In the next section different combined echo and noise reduction schemes are discussed which try to cancel the echo signal as well as obtain a noise reduction performance which is equally good as in the situation without echo source.

5. COMBINED ECHO AND NOISE REDUCTION

5.1. Multi-channel echo cancelling

Because the noise reduction performance drops if an echo source is present, one solution consists in first cancelling the echo in every microphone signal before applying the GSVD-based optimal filtering technique. This is depicted in figure 2, where the first stage represents a multi-channel NLMS-based adaptive echo canceller and the second stage represents the GSVD-based optimal filtering technique for noise reduction and additional echo suppression.

In figures 3 and 4 the SNR and SER of the enhanced signal are depicted for different filterlengths L_m (of the multi-channel echo canceller) and N (of the GSVD-based filtering

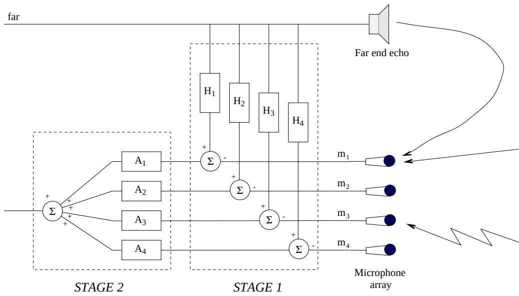

Figure 2: GSVD-based optimal filtering technique (stage 2) with multi-channel echo cancelling (stage 1)

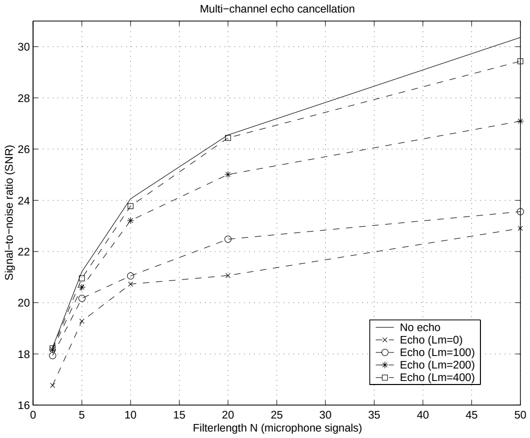

Figure 3: Effect of multi-channel echo cancelling on SNR

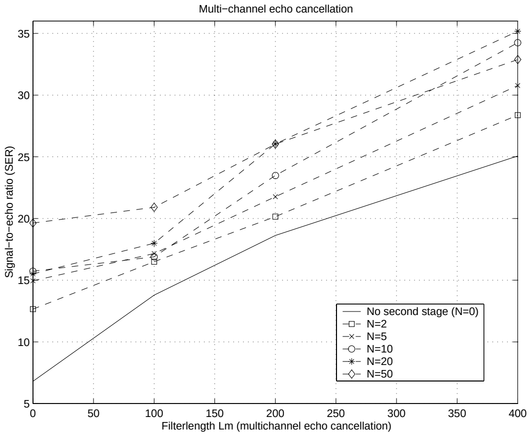

Figure 4: Effect of multi-channel echo cancelling on SER

technique). Figure 3 shows that the larger L_m, the better the noise reduction performance of the situation without echo source is approached. This figure also shows that even when using short filters a satisfactory SNR can be obtained. Figure 4 shows that the second stage achieves an additional echo suppression, depending on the filterlength N.
The next schemes incorporate the echo reference directly into the GSVD-based optimal filtering technique, hereby avoiding the multi-channel echo cancellation problem.

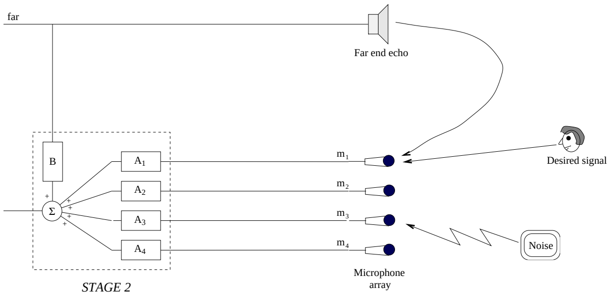

Figure 5: GSVD-based optimal filtering technique with incorporation of far-end echo signal (scheme 1)

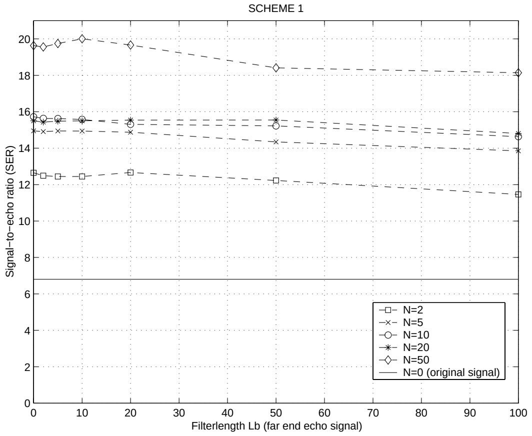

Figure 6: Effect of incorporating far-end echo signal on SER

5.2. Incorporating echo reference: scheme 1

Figure 5 depicts the scheme which incorporates the far-end echo signal $f(k)$ into the GSVD-based filtering technique by considering the echo reference as an additional microphone input and filtering it by means of a filter B with filterlength L_b. This means that the signal vector $\mathbf{u}_k$ now has the form

$$\mathbf{u}_k = \left[\begin{array}{ccccc} \mathbf{m}_{1k} & \mathbf{m}_{2k} & \dots & \mathbf{m}_{Mk} & \mathbf{f} \end{array} \right]^T \quad (9)$$

$$\mathbf{m}_{jk} = \left[\begin{array}{cccc} m_j(k) & m_j(k-1) & \dots & m_j(k-N+1) \end{array} \right] \quad (10)$$

$$\mathbf{f} = \left[\begin{array}{cccc} f(k) & f(k-1) & \dots & f(k-L_b+1) \end{array} \right] \quad (11)$$

Figure 6 shows the SER of the enhanced signal for different filterlengths L_b and N. As can be seen the filter B has no effect - even a slightly negative effect - on the SER. The reason is that the filter $b(k)$ has to model

$$b(k) \rightarrow -\sum_{j=1}^{M} g_j^f(k) \otimes a_j(k), \quad (12)$$

having a filterlength much larger than L_b. The filterlength L_b has to be limited because of the computational complexity of the GSVD-based optimal filtering technique (see section 5.4).

5.3. Incorporating echo reference: scheme 2

Instead of incorporating the far-end echo reference $f(k)$ into the GSVD-based filtering technique, one can incorporate a

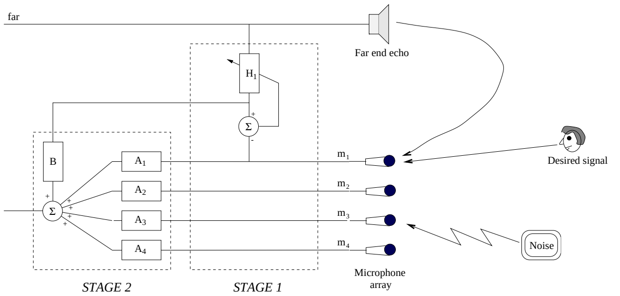

Figure 7: GSVD-based optimal filtering technique with incorporation of filtered far-end echo signal (scheme 2)

filtered version $f_1(k)$ of the echo reference,

$$f_1(k) = h_1(k) \otimes f(k), \tag{13}$$

with $h_1(k)$ an adaptive filter modelling the impulse response $g_1^f(k)$ between the echo source and the first microphone. This is depicted in figure 7. By doing this, the (short) filter $b(k)$ now has to model

$$h_1(k) \otimes b(k) \rightarrow -\sum_{j=1}^{M} g_j^f(k) \otimes a_j(k) \tag{14}$$

with $h_1(k) \simeq g_1^f(k)$. Figure 8 shows the effect of the filter B on the SNR and SER of the enhanced signal for two different lengths of the adaptive filter H_1 and for $N = 20$. The filter B only has a small effect on the SNR, but has a considerable effect on the SER.
However, compared to the scheme using multi-channel adaptive echo cancelling (section 5.1) the effect on the SNR is very small. Also the echo reduction achieved by scheme 2 is easily obtained by using a multi-channel adaptive filter (*e.g.* with a filterlength $L_m = 200$, see figure 4).

5.4. Computational complexity

Even using recursive GSVD-updating algorithms [6] the computational complexity of the GSVD-based optimal filtering technique is still rather high. If we count both an addition and a multiplication as 1 flop, the computational complexity of the GSVD-based filtering technique amounts to $27.5(MN + L_b)^2$. The computational complexity of a standard time-domain NLMS-based multi-channel adaptive echo canceller is $4ML_m$. For normal values of these parameters, one can easily verify that the scheme using the multi-channel adaptive echo canceller has a lower computational complexity than the scheme incorporating the far-end echo reference into the GSVD-based optimal filtering technique.

6. CONCLUSION

In this paper we have described two schemes for combining acoustic echo and noise reduction. One scheme uses a multi-channel adaptive echo canceller, the other scheme incorporates the echo reference directly into the GSVD-based optimal filtering technique. Simulations indicate that the scheme using a multi-channel adaptive echo canceller outperforms the scheme incorporating the echo reference directly into the GSVD-based optimal filtering technique.

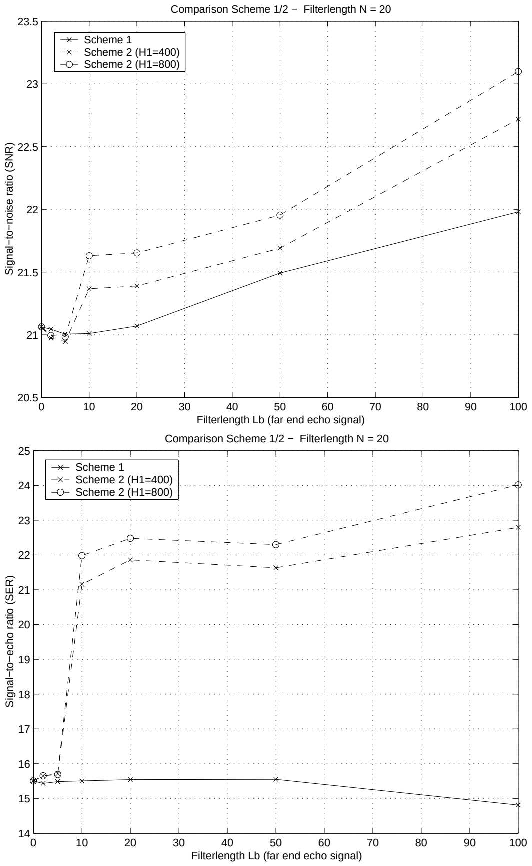

Figure 8: Effect of incorporating filtered far-end echo signal on SNR and SER

REFERENCES

[1] S. Doclo and M. Moonen, “SVD-based optimal filtering with applications to noise reduction in speech signals,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’99)*, New Paltz, New York, USA, Oct. 1999, pp. 143–146.

[2] S. Doclo and M. Moonen, “Robustness of SVD-based Optimal Filtering for Noise Reduction in Multi-Microphone Speech Signals,” in *Proc. of the 1999 IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC’99)*, Pocono Manor, Pennsylvania, USA, Sept. 1999, pp. 80–83.

[3] F. Xie and S. Van Gerven, “Comparative study of 3 speech detection methods,” Tech. Rep. MI2-SPCH-95-8, ESAT, K.U.Leuven, Belgium, Oct. 1995.

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

[5] J. Allen and D. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

[6] M. Moonen, P. Van Dooren, and J. Vandewalle, “A systolic algorithm for QSVD updating,” *Signal Processing*, vol. 25, pp. 203–213, 1991.