A HYBRID FRLS/NLMS STEREO ACOUSTIC ECHO CANCELER

Peter Eneroth\textsuperscript{1}, Steven Gay\textsuperscript{2}, Tomas G"ansler\textsuperscript{1} and Jacob Benesty\textsuperscript{2}

\textsuperscript{1}Department of Applied Electronics
Lund University, Sweden
\{pe, tg\}@tde.lth.se

\textsuperscript{2}Bell Laboratories, Lucent Technology,
Murray Hill, NJ, USA
\{slg, jbenesty\}@bell-labs.com

ABSTRACT
Teleconferencing systems employ acoustic echo cancelers to reduce echoes that result from the coupling between loudspeaker and microphone. To enhance the sound realism, two-channel audio is necessary. However, stereophonic acoustic echo cancellation is more difficult to solve because of the necessity to uniquely identify two acoustic paths, which becomes problematic since the two excitation signals are highly correlated. In this paper a wideband stereophonic acoustic echo canceler is presented. The fundamental difficulty of stereophonic acoustic echo cancellation (SAEC) is described and an echo canceler based on a fast recursive least squares algorithm in a subband structure is proposed. The structure has been used in a real-time implementation, with which experiments have been performed. In the paper, simulation results of this implementation on real life recordings, with 8 kHz bandwidth, are studied. The results clearly verify that the theoretic fundamental problem of SAEC also applies in real-life situations. They also show that more sophisticated adaptive algorithms are needed in the lower frequency regions than in the higher regions.

1. INTRODUCTION
In conferencing systems, such as teleconferencing and computer conferencing, acoustic echo cancelers (AEC’s) are needed to reduce the echo that results from the acoustic coupling between the loudspeaker and the microphone. The AEC identifies the echo path and simultaneously reduces the echo by means of adaptive filtering. If the conferencing system has dual audio channels in each direction, the classical monophonic AEC’s will not provide sufficient echo suppression, and special stereophonic acoustic echo cancelers (SAEC’s) are needed. The fundamental problem of SAEC [1] is described, and we propose a structure that has proven to perform well in a real-time implementation in this paper.

In stereophonic conferencing system, spatial audio information is also transmitted. Not only will the listener get a more realistic sound, but she will also be able to aurally localize the speaker at the other end. Studies have shown that this improves perception, especially when speech from several speakers overlap [2]. However, there are now four acoustic echo paths to identify, two to each microphone, Fig. 1. This will not only cause increased calculation complexity, but also new fundamental problems of the solution, as we will see.

Four mono AEC’s, straightforwardly implemented in the stereo case, not only would have to track changing echo paths in the receiving room but also in the transmission room! For example, the canceler has to reconverge if one talker stops talking and another starts talking at a different location in the transmission room. There is no adaptive algorithm that can track such a change sufficiently fast and this scheme therefore results in poor echo suppression. Thus, a generalization of the mono AEC in the stereo case does not result in satisfactory performance.

The theory explaining the problem of SAEC was first described in an early paper, [3] and later on in [4, 1]. The fundamental problem is that the two channels usually carry linearly related signals which in turn make the normal equations to be solved by the adaptive algorithm singular. This implies that there is no unique solution to the equations but an infinite number of solutions and it can be shown that all (but the physically true) solutions depend on the transmission room. In [1] it is also shown that the only solution to the non-uniqueness problem is to reduce the correlation between the stereo signals from the transmission room and an efficient low complexity method for this purpose was also given.

Lately, attention has been focused on the investigation of other methods that decrease the cross-correlation between the channels in order to get well behaved estimates of the echo paths, [5, 6, 7]. The main problem is how to reduce the correlation sufficiently without affecting the stereo perception and the sound quality. Even though the above methods may improves the SAEC’s ability to find the true solution, the normal equations to be solved are still ill conditioned. The standard Normalized Least Mean Square (NLMS) adaptive algorithm is known to converge slowly in these situations. More sophisticated algorithms such as the Recursive Least Squares (RLS), that are less affected by a high condition number, are preferred in SAEC’s. The combination of four adaptive filters per AEC and sophisticated adaptive algorithms results in high calculation complexity, imposing a subband structure to be used.

2. SYSTEM DESCRIPTION
As described in the previous section, a stereophonic acoustic echo canceler needs four adaptive filters, each adapted with a sophisticated algorithm. This will of course lead to a system with high computational complexity, but by applying the echo canceler in a subband scheme, it is possible to reduce this complexity. The subband structure, decomposes the signals into $M$ downsampled subband signals, where each subband corresponds to a given frequency interval. Four analysis filter banks are needed to decompose the transmission room signals, $x(t)$ and receiving room
As described in the introduction, the fundamental problem in SAEC lies in the correlation between the two channels. One method to determine how correlated the two channels are, is to calculate the coherence function [1].

\[
\gamma(f) = \frac{S_{xx}(f)}{S_{xx}(f)S_{xx}(f)}
\]

where \(S_{xx}(f)\) is the power spectral estimate, and \(\gamma(f) = 1\) shows that the two channels are completely linearly related to each other. In the worst case, and also usually the practical case, the two signals \(x_1(n)\) and \(x_2(n)\) originate from the same source, Fig. 1. Due to background noise in the transmission room, the signals are not completely linearly related. Considering that the source signal is speech, the SNR in an normal office environment would be higher in the lower frequency region than in the higher. This due to the low speech energy in the higher frequency region. Examples showing this is given in the next section.

In [1], a half-wave rectifier, described by

\[
x'_1(n) = x_1(n) + \alpha x_1(n) + |x_1(n)|
\]

\[
x'_2(n) = x_2(n) + \alpha x_2(n) - |x_2(n)|
\]

was presented as a simple and effective non-linear method to decrease the correlation between the channels. Informal listening tests have verified that \(\alpha \leq 0.5\) insignificantly affects the audio quality. As described above, the channels are usually most correlated in the lower frequencies. Therefore a simple enhancement to (2) would be to divide the signals into a couple of subbands and apply the half-wave rectifier in these subbands. This way, more distortion can be introduced in the lower frequency regions by increasing \(\alpha\) for these bands.

### 4. REAL-LIFE SIMULATIONS

The purpose of this experiment is to show the phenomenon of reconvergence of the echo canceler when one talker stops and another starts talking at a different location in the transmission room. This is caused by the inherent non-uniqueness problem of SAEC. Perceptually transparent nonlinear processing, [1], has theoretically been proven to solve this problem. We show how one such nonlinear processor, the half-wave rectifier (2), decreases the correlation between the two channels, and improves the performance of the SAEC. Comparison between the FRLS, the NLMS algorithm and a of these mixed structure is also performed.

The recordings were made in actual rooms at the Bell labs facility (HuMaNet I, [13]). The receiving room SNR is fairly high, \(\approx 40\) dB, and the sampling rate is 16 kHz. The echo canceler uses a 64 band filter bank, and the downsampling rate is 48. The residual echo signals in the lower 33 subbands are estimated by the adaptive filters, whereas the upper subbands are reconstructed as described confirming this is shown in Section 4.

A real-time implementation of the system was made on a parallel floating point DSP board. Devices for residual echo suppression were implemented and used during the real-time experiments, but were deactivated in the simulations, presented in the last section.

### 3. CHANNEL DECORRELATION

The recordings were made in actual rooms at the Bell labs facility (HuMaNet I, [13]). The receiving room SNR is fairly high, \(\approx 40\) dB, and the sampling rate is 16 kHz. The echo canceler uses a 64 band filter bank, and the downsampling rate is 48. The residual echo signals in the lower 33 subbands are estimated by the adaptive filters, whereas the upper subbands are reconstructed as described confirming this is shown in Section 4.

A real-time implementation of the system was made on a parallel floating point DSP board. Devices for residual echo suppression were implemented and used during the real-time experiments, but were deactivated in the simulations, presented in the last section.
in Section 2. In each subband, the length of the adaptive filter is 66. This corresponds to 3168 fullband taps by which it is possible to cancel an echo path of length 198 ms.

In Fig. 2, the coherence function, (1), is shown for the recorded speech signals described above. The power spectral estimates, 

\[ S_{x_1 x_2} \text{ and } S_{y_1 y_2} \]

were calculated using the Welch method [14], with a 8196 taps long Hanning window. Figure 2a shows the coherence between the left and right channels for unprocessed transmission room speech, and Fig. 2b, the coherence for the same signal, but processed with the half-wave rectifiers, \( \alpha = 0.5 \).

One way to show the effectiveness of the decorrelation is to study how the performance of the echo canceler decreases after a speaker position change in the transmission room. If the physically true solution is found by the adaptive filter, the performance is independent of impulse response changes in the transmission room, \( g_1 \) and \( g_2 \) in Fig. 1. As performance index, the mean square error (MSE) energy of the residual is used. The MSE is given by,

\[ \text{MSE} = \frac{P_e}{P_y}, \quad P_e = \text{LPF}\{e^2(n)\}, \]  

(3)

where LPF denotes a low-pass filter. \( P_y \) is calculated analogously.

Figure 3b shows the MSE for unprocessed transmission room speech and speech processed with the half-wave rectifier (\( \alpha = 0 \), and 0.5 respectively, where \( \alpha \) is the degree of nonlinearity). The active talker is changed after 5.1 s (dashed vertical line). A significant stability improvement, \( \approx 20 \text{ dB} \) better echo attenuation, is achieved with the higher \( \alpha \) in this real life experiment.

As have been discussed in Section 3 and is shown in Fig. 2, the correlation between the channels is higher in the lower frequency regions than in the higher. Therefore, the stability problems, due to changes of the transmission room signal path (\( g_1, g_2 \)), should be severer in the lower subbands. In Fig. 4a, the MSE before the transmission room speaker position change for an unprocessed signal (solid line) versus the MSE directly after the position change (dashed line) is plotted as a function of the subband number. As can be seen, the echo canceler performance decrease is largest in the lower subbands. In other speech signals, where the transmission room background noise is higher (SNR \( \approx 30 \text{ dB} \)), the stability problem is only apparent in the lower half of the subbands. That is, stronger decorrelation is needed in the lower frequency regions, also discussed in Section 3. In Fig. 4b, it is shown that the half-wave rectifier effectively eliminates the stability problems in all subbands.

In the previous examples, the two-channel FRLS algorithm is used in all subbands. In order to reduce the calculation complexity, it is possible to switch to an NLMS algorithm in the upper subbands without significantly reducing the performance of the echo canceler. In Fig. 5 the MSE performance difference between the FRLS and the NLMS algorithm is shown for one typical lower and one typical higher subband. The impressive performance gain of the FRLS algorithm only applies for the lower subbands. In Fig. 6, the performance for SAEC with NLMS, FRLS and with FRLS in the lower subbands and NLMS in the higher subbands is shown.

5. CONCLUSIONS

With the real-time stereophonic acoustic echo canceler presented in this paper, we have been able to confirm that the use of two channels significantly enhances the ability to separate speech signals in video conferencing systems. Therefore a listener has an improved ability to distinguish one speaker in the other room, when another speaker, situated at another position, is talking at the same time.

It has also been confirmed that decorrelation is crucial for the stability of the system, both in real-time experiments and in off-line simulations on the real-life recorded signals, presented in the previous section. The studies confirm also that without a decorrelator, it is unlikely that the echo canceler converges to the correct solution. This is especially the case in the lower subbands, and in situations when the far-end background noise is low. Finally it is shown that the two-channel FRLS adaptive algorithm is superior to the NLMS in the lower subbands, but the performance gain is less obvious in the upper subbands.
Figure 4: Subband mean square error at two time instances, solid line at 5.1 s and dotted line at 5.4 s, where the latter is directly after a speaker position change in the transmission room. (a) Unprocessed signal, shown in Fig. 3b. (b) Signal processed with the half-wave rectifier, $\alpha = 0.5$, shown in Fig. 3c.

6. REFERENCES


