A real-time implementation of a stereophonic acoustic echo canceler

P. Eneroth

Steven L. Grant
Missouri University of Science and Technology, sgrant@mst.edu

T. Gansler

J. Benesty

Follow this and additional works at: http://scholarsmine.mst.edu/faculty_work

Part of the Electrical and Computer Engineering Commons

Recommended Citation
http://scholarsmine.mst.edu/faculty_work/268

This Article is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Faculty Research & Creative Works by an authorized administrator of Scholars' Mine. For more information, please contact weaverjr@mst.edu.
A Real-Time Implementation of a Stereophonic Acoustic Echo C canceler

Peter Eneroth, Steven L. Gay, Tomas Gänsler, Member, IEEE, and Jacob Benesty, Member, IEEE

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 (SAEC) 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 is described and an echo canceler based on a fast recursive least squares (RLS) algorithm in a subband structure, with equidistant frequency bands, is proposed. The structure has been used in a real-time implementation, with which experiments have been performed. In this 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.

Index Terms—Real-time implementation, recursive least squares (RLS), stereophonic acoustic echo cancellation, subband filtering.

I. INTRODUCTION

In conferencing systems, such as teleconferencing and desktop conferencing, acoustic echo cancelers (AECs) 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 AECs will not provide sufficient echo suppression, and more sophisticated stereophonic acoustic echo cancelers (SAECRs) are needed. In this paper, we will show the fundamental problem of stereophonic acoustic echo cancellation (SAEC), possible solutions, and propose a structure that has proven to perform well in a real-time implementation.

In a stereophonic conferencing system, spatial audio information is also transmitted. Not only will the listener get a more realistic sound, but the listener 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 [1]. 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 a new fundamental problem of the solution, as we will see.

Four mono AECs, 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 [2] and later in [3] and [4]. 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 [4], it is also shown that the only solution to the nonuniqueness 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]–[8]. The main problem is how to reduce the correlation sufficiently without affecting the stereo perception and the sound quality.

Even though these methods may improve the SAECRs’ 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 affine projection algorithm (APA) or the recursive least squares (RLS) that are less affected by a high condition number, are preferred in SAECRs. The combination of four adaptive filters per AEC and sophisticated adaptive algorithms results in high calculation complexity, imposing the need for a subband structure.

In the following, we will explain the fundamental problem with SAECRs and possible solutions. We will propose a high-performance SAECR structure, that has been verified in a real-time implementation. Finally we will show and discuss results from real-life recordings using the proposed structure.

II. PROBLEMS AND SOLUTIONS OF STEREO PHONIC ACOUSTIC ECHO CANCELLATION

In stereophonic acoustic echo cancellation, there are four independent transmission paths between the two microphones and the two loudspeakers (Fig. 1). The impulse responses of all four
echo paths need to be estimated by the echo canceler. Usually the two transmission room signals, $x_1(n)$ and $x_2(n)$, originate from the same source, and are therefore highly correlated. Because of this, it is difficult to estimate the impulse responses, $h_{1,N}(n)$ and $h_{2,N}(n)$. The circumstances under which convergence to the true echo paths of a SAECR is achieved has been thoroughly analyzed in [4]. A problem formulation and a summary of methods to reduce the correlation between the channels are given in the following.

A. Problem Formulation

Assume that the transmission room microphone signals are given by (Fig. 1)

$$
x_i(n) = g_i(n) * s(n), \quad i = 1, 2
$$

where $s(n)$ is the source signal in the transmission room and $g_i(n), i = 1, 2$, are the transmission room echo paths of length $M$. The symbol $*$ denotes convolution. For simplicity, we will only study one return path from the receiving room to the transmission room, but similar remarks will be valid for the other path. The residual echo for this channel, $e(n)$, after the EC is

$$
e(n) = y(n) - \hat{h}_{1,L}(n)x_{1,L}(n) - \hat{h}_{2,L}(n)x_{2,L}(n)
$$

$$
y(n) = h_{1,N}x_{1,N}(n) + h_{2,N}x_{2,N}(n)
$$

$$
h_{i,N} = [h_{i,0} \cdots h_{i,N-1}]^T
$$

$$
x_{i,N}(n) = [x_i(n) \cdots x_i(n-N+1)]^T.
$$

Here, $h_{i,N}, i = 1, 2$ are the true responses of length $N$ of the receiving room and $\hat{h}_{i,L}(n), i = 1, 2$ are the estimated responses of length $L$. The symbol $H$ denotes the Hermitian transposition operator. Minimization of the weighted least squares criterion

$$
J(n) = \sum_{l=1}^{n} \lambda^{n-l}|e(l)|^2, \quad 0 < \lambda < 1
$$

results in solving the system of linear equations [9]

$$
R_{xx}(n)\begin{bmatrix}
\hat{h}_{1,L}(n) \\
\hat{h}_{2,L}(n)
\end{bmatrix} = r_{yy}(n)
$$

where $r_{xx}(n)$ is the estimated cross-correlation vector and $R_{xx}(n)$ is the estimated correlation matrix

$$
R_{xx}(n) = \sum_{l=1}^{n} \lambda^{n-l}\begin{bmatrix}
x_{1,L}(l)x_{1,L}^H(l) \\
x_{2,L}(l)x_{2,L}^H(l)
\end{bmatrix},
$$

The challenging problem with stereophonic acoustic echo cancellation lies in the condition number of this matrix. If we define the misalignment as $\varepsilon(n) = ||h - \hat{h}(n)||^2/||h||^2$ and $\hat{h}(n) = [\hat{h}_{1,L}(n) \hat{h}_{2,L}(n)]^T$, $h = [h_{1,N} \ h_{2,N}]^T$, it can be shown that

$$
L \geq M \Rightarrow R_{xx}(n) \text{ is singular } \forall n
$$

$$
L < M \Rightarrow R_{xx}(n) \text{ is ill-conditioned}
$$

$$
L \geq N \Rightarrow \varepsilon(n) = 0, \quad n \geq N
$$

$$
L < N \Rightarrow \varepsilon(n) \neq 0, \quad \forall n
$$

where the two latter statements in (9) require $R_{xx}(n)$ not to be singular. Equation (9) is valid in the situation where no noise is added to the microphone signal $y(n)$ (see Fig. 1).

As shown in [4] and (9), the tails of the impulse responses both in the transmission and receiving rooms play a key role. Thanks to the impulse response tails in the transmission room, we can obtain a unique solution to the normal equation. However, because of the impulse response tails in the receiving room, we have potentially large misalignment. We assume of course that $L \geq M \geq N$, since this is the realistic case to be dealt with. Theoretically, $M$ and $N$ are infinitely long, but the normal reverberation time in an office room is approximately 0.3 s.

There are two ways to decrease the misalignment. The first way is to use longer adaptive filters, but commensurately, the adaptive algorithm becomes very slow in terms of convergence speed and is more expensive to implement in terms of memory, arithmetic complexity, etc. Moreover, the solution is not robust in the sense that it is ill-conditioned and still sensitive to transmission room changes. A second way, the practical approach, is to decorrelate partially (or in totality) the two input signals. The difficulty is to decorrelate the signals without decreasing the signal quality.

B. Decorrelation Methods

The most straight-forward method to reduce the correlation between two channels is probably to add independent random noise to each channel, $x_i(n), i = 1, 2$. This was described in [3], but it is also pointed out that in order to sufficiently reduce the correlation, the noise-level had to be greater than the level of the maximum nonperceptible noise.

To reduce the perceived distortion it would be preferable if the decorrelating signal is similar to the original signal. But the core problem in SAEC is that the two channels are linearly related, i.e., adding a signal that is linearly related to the original signal will not reduce the correlation between the two channels. In [4], it is suggested that a nonlinearly processed source signal should be added to the source signal itself. It was found that adding a half-wave rectified signal to the original signal performed well in addition to having a simple and low-complexity structure.
This can be expressed as
\begin{align}
\hat{x}_1(n) &= x_1(n) + \alpha \frac{x_1(n) + \chi x_1(n)}{2} \quad (10) \\
\hat{x}_2(n) &= x_2(n) + \alpha \frac{x_2(n) - \chi x_2(n)}{2} \quad (11)
\end{align}
where \( \alpha \) determines the amount of added distortion. It has been found that \( \chi \) between 0.3 and 0.5 decreases the channel correlation significantly and that the distortion is hardly audible in an office environment [10]. The stereo perception is not affected.

In systems where the transmission path between the transmission and the receiving rooms includes an audio codec, certain coders will decorrelate the channels. In [5], the influence that a perceptual audio coder, MPEG layer III [11], has on a SAECR is analyzed. It is shown that the coder can decorrelate the signal because nonperceivable quantization noise is added to the source signal. How efficiently the coder decorrelates the signal depends on what features are used for compression. For example, advanced stereo coders usually operate in a joint stereo mode, where two correlated channels are coded jointly. This can actually increase the correlation between the channels, and should not be used if the coder also has to serve as decorrelator.

Traditional perceptual audio coders achieve better audio quality for a given compression ratio than speech coders when the source signal contains music, whereas speech coders perform better on pure speech signals. In an attempt to combine the best of the two worlds, coders, that softly switch mode depending on the source signal, have been proposed. The decorrelating properties of one such coder, the MTPC coder [12], is also shown in the examples.

III. PROPOSED STRUCTURE FOR THE STEREOPHONIC ACOUSTIC ECHO CANCELER

In this section, all vital parts of the real-time implemented SAECR are presented. First of all, a decorrelator is needed to reduce the correlation between the two transmission room signals, \( x_1(n) \) and \( x_2(n) \). In the system we have chosen to use the half-wave rectifier, presented in the previous section.

Even after the decorrelator, finding the correct echo paths is still not a well-conditioned problem. Therefore, the two-channel RLS algorithm, which has shown great promise in SAEC applications [13], was chosen for the adaptive filters. This algorithm has very fast convergence rate, even for signals with a large eigenvalue spread of the correlation matrix. The two main disadvantages with the RLS algorithm are the high calculation complexity and stability problems with nonstationary signals, such as speech. The stability is improved by monitoring the state of the RLS algorithm, and to reinitialize parameters when they become unstable. A two-path structure [14] is used to further reduce the effects of an unstable algorithm. This structure also serves as a double-talk detector. Due to the high calculation complexity, the adaptive filters are performed in a subband structure, illustrated in Fig. 2.

Fig. 2. Stereophonic subband acoustic echo canceler. In total four analysis filterbanks decompose the transmission room and receiving room signals. In each subband, the two-path two-channel FRLS algorithm cancels the major part of the echo. The residual echo is reduced even further by the suppressor before reconstructing the fullband error signal, \( e(n) \).

A. Fast Recursive Least-Squares Adaptive Algorithm

A complete analysis of the FRLS algorithm is beyond the scope of this article. However, a general analysis of the RLS algorithm can be found in [9] and stabilized two-channel versions are described in [13] and [15]. Nevertheless, the definition of the specific version of the two-channel FRLS used in the implementation is given below.

From Fig. 1 the following quantities must be defined (see (12)–(14) at the bottom of the page) where
\[
L \quad \text{length of the adaptive filters}; \\
h_{1,1}(n) \quad \text{coefficient number 1 of channel 1}; \\
h_{2,1}(n) \quad \text{coefficient number 1 of channel 2},
\]
Note that the channels of the filter and state vector \( \tilde{x}(n) \) are interleaved in this algorithm. The complete two-channel FRLS adaptive filter is given in Table I, where the following quantities are used:
\[
A(n), B(n) = \text{Forward and backward prediction coefficient matrices}; \\
E_A(n), E_B(n) = \text{Forward and backward prediction error energy matrices}; \\
e_A(n), e_B(n) = \text{Forward and backward prediction error vectors}; \\
G(n) = \text{Kalman gain vector}; \\
\varphi(n) = \text{Inverse conversion factor}; \\
k \in [1,5,25], \quad \text{Stabilization parameter}; \\
\lambda \in (0,1], \quad \text{Forgetting factor}.
\]

\begin{align}
\chi(n) &= [x_1(n) \quad x_2(n)]^T, \quad (2 \times 1) \\
\tilde{x}(n) &= [\chi^T(n) \quad \chi^T(n-1) \quad \cdots \quad \chi^T(n-L+1)]^T, \quad (2L \times 1) \\
\hat{h}(n) &= [\hat{h}_{1,1}(n) \quad \hat{h}_{2,1}(n) \quad \cdots \quad \hat{h}_{1,L}(n) \quad \hat{h}_{2,L}(n)]^T, \quad (2L \times 1)
\end{align}
TABLE I
STEREOPHONIC FRLS ALGORITHM FOR COMPLEX ARITHMETIC. THE CONJUGATE AND THE HERMITIAN TRANSPOSITION OPERATOR ARE DENOTED * AND H, RESPECTIVELY. VARIABLES ARE DEFINED IN SECTION III.A. THE PREDICTION PART IS COMPLETE, WHEREAS THE FILTERING PART ONLY DESCRIBES ONE RETURN PATH

<table>
<thead>
<tr>
<th>Input signals</th>
<th>Matrix sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x(n) = [x_1(n) \ldots x_{2L}(n)]^T$</td>
<td>$(2 \times 1)$</td>
</tr>
<tr>
<td>$y(n) = [y^T(n) \ldots y^T(n-L+1)]^T$</td>
<td>$(2L \times 1)$</td>
</tr>
</tbody>
</table>

**Prediction**

$e_A(n) = x(n) - A^H(n-1)x(n-1)$

$\varphi_1(n) = \psi(n-1) + E_A^{-1}(n-1)e_A(n)$

$M(n) = \begin{bmatrix} 0 & (n - 1) \\ -A(n-1) \\ \end{bmatrix} E_A^{-1}(n-1)e_A(n)$

$e_{B2}(n) = x(n-L) - B^H(n-1)x(n)$

$\psi(n) = \varphi(n) - E_B(n)\psi(n)$

$A(n) = A(n-1) + G(n-1)e_A^H(n)/\varphi(n-1)$

$E_A(n) = \lambda(E_A(n-1) + e_A(n)e_A^H(n)/\varphi(n-1))$

$G(n) = M(n) + B(n-1)m(n)$

$e_B(n) = e_B(n-1)m(n)$

$e(n) = k_e e_B(n) + (1-k_e)e_B(n)$

$B(n) = B(n-1) + G(n)e_B^H(n)/\varphi(n)$

$E_B(n) = \lambda(E_B(n-1) + e_B(n)e_B^H(n)/\varphi(n))$

**Filtering**

$e(n) = y(n) - h^H(n-1)x(n)$

$\hat{h}(n) = \hat{h}(n-1) + G(n)e(n)/\varphi(n)$

**Definition**

$\hat{h}(n) = [\hat{h}_{1,0}(n) \hat{h}_{2,0}(n) \ldots \hat{h}_{1,L-1}(n) \hat{h}_{2,L-1}(n)]^T$

In this version, stability is improved with the stability parameter $k$. But for operation on nonstationary signals, like speech-signals, further enhancements are needed. First of all, by monitoring $\varphi$, it is possible to detect if the algorithm is about to become unstable. If this is the case, the parameters in the prediction part are reset to their start values, while the adaptive filter estimate, $\hat{h}$, can be left unchanged. A suitable initial value for $\Delta(n)$, $B(n)$ and $G(n)$ is 0 whereas the energy estimates, $E_A(n)$ and $E_B(n)$ could be initialized with a recursive estimate of the speech energy. During the time between restart and until the algorithm has reconverged, echo cancellation may be poor. A two-path structure, presented next, improves performance in these situations.

**B. Two-Path Adaptive Filter**

In situations of large disturbances, for example double-talk, or if the adaptive filter becomes unstable, the filter may diverge from a good estimate. If the estimate diverges, it would be better to use an earlier filter estimate until the adaptive filter has reconverged. This is the purpose of the two-path adaptive filter structure [14], illustrated in Fig. 3. In this structure, the adaptive filter is used only to estimate the impulse response $\hat{h}_{RLS}$. Then, it has to be determined if this new estimate is better than a previous estimate, denoted $\hat{h}$. If the new estimate is better, $\hat{h}$ is updated. The output signal from the echo canceler, $e(n)$, is calculated in the two-path structure using $\hat{h}$. It should be noted that in the implementation, the two-path structure is adapted in the subbands, and that the decision made in one subband is independent of the states in all other subbands.

The most crucial condition to be met for a filter update is that the short-time residual echo energy in the adaptive filter, $E_{RLS,i}^*$, is less than the residual echo energy in the two-path structure, $E_{\varphi,i}$, multiplied with a fixed value $C < 1$

$$E_{RLS,i} < CE_{\varphi,i}$$  \hspace{1cm} (15)

where $i \in \{1,2\}$ denotes the channel number.

**C. Subband Filterbank Design**

The main reason for using a subband scheme is the reduction of calculation complexity, but other positive effects include increased stability of the adaptive filter, because fewer taps are adapted in each subband, and a structure that allows for efficient implementations on parallel systems. The condition number of the correlation matrix in each subband is also reduced, resulting in increased convergence rate for the LMS class of adaptive filters. The two biggest disadvantages are the transmission path delay that is introduced (exemplified in Table II) and possible aliasing due to downsampling. In [16] it has been shown that if critical downsampling is used, i.e., if we have the same down-sampling ratio as number of subbands $M$, aliasing will significantly decrease the performance of the adaptive filters. Therefore noncritical downsampling, i.e., $\tau < M$, in conjunction with filters that have good stopband attenuation were chosen. General discussions of filterbanks can be found in numerous books and articles [17]–[20], but since the emphasis has been on critical downsampled filterbanks, an efficient structure with noncritical downsampling is shown in the Appendix. Methods on how to design prototype filters are also discussed.

**D. Computational Complexity**

In this section, we calculate the number of real valued multiplications and additions the adaptive filter and the filterbanks need per fullband sample period. The Fourier transform in the filterbank and most parts of the adaptive filter (FRLS) are performed with complex arithmetic. In this analysis, multiplication between two complex numbers is counted as four real multiplications and two real additions.
The number of multiplications needed for the real-valued two channel FRLS [13] is $32L$, and the number of additions is $32L$, where $L$ is the length of the adaptive filter. This includes calculation of the residual signals for both channels. The subband signals are complex valued, and the complex version of the two channel FRLS is given in Table I. This algorithm needs $128L_{\text{sub}}$ real-valued multiplications and $128L_{\text{sub}}$ real additions, where $L_{\text{sub}}$ denotes the length of the adaptive filter in the subbands. Echo cancellation in the subband structure described in previous sections, needs $M/2 + 1$ adaptive filters, Appendix A, but due to downsampling they are updated $1/r$ times the fullband rate. The length of the adaptive filters are $L_{\text{sub}} = L/r + C_{\text{nc}}$, where $C_{\text{nc}}$ compensates for the noncausal taps [21]. The total number of real multiplications per fullband sample is $128(M/2 + 1)L_{\text{sub}}/r$ and the number of additions $128(M/2 + 1)L_{\text{sub}}/r$.

The analysis, (26), and synthesis, (35), filterbanks include two parts, polyphase filtering and fast Fourier transformation. The polyphase filtering uses $K$ multiplications and $K$ additions per filterbank, where $K$ is the length of the prototype filter. Since the input signal of the Fourier transform in the analysis filterbank is real-valued, and the output signal from the Fourier transform in the synthesis filterbank is also real-valued, only two Fourier transforms are needed for the four analysis filterbanks and one Fourier transform for the two synthesis filterbanks. In order to separate the one complex valued Fourier transform into two real-valued transforms, $2M - 4$ extra additions are needed [22]. If the Fourier transforms are implemented with a Radix 2 structure, each need $2M \log_2 M - 7M + 12$ real multiplications and $3M \log_2 M - 3M + 4$ additions [22]. Also the filterbanks need to be updated at $1/r$ times the fullband rate. The total number of multiplications of the four analysis filterbanks is $(1/r)(4K + 4M \log_2 M - 14M + 24)$ and the total number of additions is $(1/r)(4K + 6M \log_2 M - 2M)$. In the synthesis filterbank, $K$ extra additions are needed to update the state vector in each filterbank, (35). The total number of multiplications for the two synthesis filterbanks is $(1/r)(2K + 2M \log_2 M - 7M + 12)$ and the number of additions $(1/r)(4K + 3M \log_2 M - M)$.

In Table II, complexity examples with different number of subbands are given. Three system configurations are considered: Two-channel FRLS in all subbands, NLMS in all subbands, and finally, FRLS in half of the subbands, and NLMS in the other half.

### IV. Simulations

In order to validate the system in a controlled manner and verify the necessity to decorrelate the stereo signals, simulations have been performed on data recorded in HumaNet room B [23]. First, a recording representing the transmission room was performed. In this recording, the excitation signal was a high-quality speech signal recorded in an anechoic chamber. Two loudspeakers were used, representing two speakers at different positions, in order to create signals with spatial changes in the transmission room. Also, signals representing the receiving room were recorded, using the transmission room signal above as excitation signal. The signals were recorded at a sampling rate of 16 kHz and the average SNR (echo-to-noise ratio) was approximately 40 dB, i.e., a very low background noise level. In the simulations in this section, the echo canceller had the following settings:

- number of subbands: $M = 64$;
- downsampling rate: $r = 48$;
- number of estimated impulse response taps in each subband: $L_{\text{sub}} = 66$;
- number of corresponding fullband impulse response taps: $L = rL_{\text{sub}} = 3168$;
- number of possible fullband noncausal taps [21]: 300.

As shown in Section II-A, the echo canceler problem has an infinite number of solutions when the two input signals, $x_1$ and $x_2$, in Fig. 1, are linearly related. The magnitude coherence function

$$
\gamma(f) = \frac{|S_{x_1 x_2}(f)|}{\sqrt{S_{x_1 x_1}(f)S_{x_2 x_2}(f)}}
$$

(16)

is a measure of how correlated the two signals are [4], where $\gamma(f) = 1$ shows that the two signals are completely linearly related to each other. That is, the SAECR will have difficulties to converge to the correct solution in those frequency regions where the magnitude coherence function is close to one. In Fig. 4, the magnitude coherence function is shown for speech signals recorded as described above. The power spectral estimates, $S_{x_1 x_j}$, were calculated using the Welch method [24] with a Hanning window of 8196 samples on the signals recorded at 16 kHz. In the estimations, 30 s long signals were used. Fig. 4(a) shows the magnitude coherence between the left and right channels for an unprocessed transmission room speech recording. The correlation between the channels can be reduced by pre-processing the signals. In Fig. 4(b), the magnitude coherence for signals preprocessed with half-wave rectifiers, Section II-B, with $\alpha = 0.5$ is shown. The signals can also be decorrelated by the use of a coder and a decoder [5]. In Fig. 4(c), the magnitude coherence function is shown for a signal that has been coded/decoded by an MPEG Layer III coder [11] and finally in Fig. 4(d), the result using an MTPC coder [12] is shown. Both

<table>
<thead>
<tr>
<th>System configuration</th>
<th>Number of subbands, $M$</th>
<th>1</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype filter length, $K$</td>
<td>–</td>
<td>207</td>
<td>415</td>
<td>831</td>
<td>1663</td>
<td></td>
</tr>
<tr>
<td>Signal delay (ms)</td>
<td>0</td>
<td>13</td>
<td>26</td>
<td>52</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td><strong>System components complexity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filterbank</td>
<td>–</td>
<td>1.8</td>
<td>1.9</td>
<td>2.0</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Two-channel FRLS</td>
<td>1.6k</td>
<td>410</td>
<td>200</td>
<td>100</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>NLMS</td>
<td>410</td>
<td>100</td>
<td>50</td>
<td>25</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Two-path</td>
<td>200</td>
<td>52</td>
<td>25</td>
<td>12</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td><strong>System complexity (adaptive filter, filterbank, two-path)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRLS</td>
<td>1.6k</td>
<td>470</td>
<td>230</td>
<td>110</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>Hybrid NLMS/FRLS</td>
<td>200</td>
<td>150</td>
<td>76</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLMS</td>
<td>610</td>
<td>160</td>
<td>76</td>
<td>39</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Table II: Calculation complexity comparison given as number of $10^6$ real valued multiplications per second at 16 kHz sampling period. Corresponding fullband impulse response length $L = 3168$, downsampling rate $r = (3/4)M$, number of non-causal taps per subband $C_{\text{nc}} = 5$ [21]. In the hybrid NLMS/FRLS system, FRLS is used in the lower half of the subbands, and NLMS in the upper half.
The magnitude coherence function between right and left transmission room signal in the echo canceler. Average signal to noise ratio is approximately 40 dB. (a) Unprocessed transmission room signals, (b) signals preprocessed with the half-wave rectifier, $\alpha = 0.5$, (c) signal coded/decoded with an MPEG layer III audio coder [11], coded at 32 kbit/s per channel, and (d) MTPC coder [12], same conditions as Fig. 4(c).

One way to show the effectiveness of the decorrelation is to study how the performance of the echo canceler decreases after a position change of the transmission room speaker. As a performance index the normalized mean square error (MSE) energy of the residual is used. The MSE is given by

$$\text{MSE} = \frac{P_{\text{res}}}{P_{\text{in}}} \quad (17)$$

$$P_{\text{res}} = \text{LPF}[e(n) - w(n)]^2 \quad (18)$$

where $w$ denotes the receiving room background noise signal and LPF denotes a lowpass filter; in this case it has a single real pole at 0.999. $P_{\text{res}}$ is analogously calculated. In all examples, except the double-talk example, the background noise signal $w$ is unknown, and cannot be subtracted according to (18). This will somewhat increase the MSE. The excitation signals to be used in the examples are shown in Fig. 5. The left channel is shown above the right channel. In this signal, the speaker moves from a position close to the left microphone, to a position closer to the right microphone at 5.1 s. Shown in Fig. 6(a) is the MSE resulting from echo cancellation of a signal that has not been processed with the half-wave rectifier. Especially notice the severe increase of MSE after the instantaneous transmission room speaker position change at 5.1 s. In Fig. 6(b), the mean square error for the same excitation signal but processed with the half-wave rectifier, $\alpha = 0.5$, is shown. Under these conditions, the SAECR converges toward the true solution, and is therefore less sensitive to echo path changes in the transmission room. Because of this, the MSE is almost unaffected by the transmission room speaker position change at 5.1 s. Simulations have shown similar behavior for the coded/decoded signals used in Fig. 4(c) and (d) as for signals processed with the half-wave rectifier $\alpha = 0.5$. This suggests that the magnitude coherence function is an effective measure of how the condition of the correlation matrix $R_{xx}(n)$ (8) affects the performance of the SAECR.

In Fig. 7, the MSE as a function of subband number is shown for two time instances, before (solid line) and directly after (dashed line) the transmission room speaker changes positions. In Fig. 7(a), where an unprocessed signal was used as input for the echo canceler, it is shown that the increase in MSE is severest in the lower subbands. This corresponds to the region where the channels are highly correlated, compare Fig. 4(a). In Fig. 7(b), a signal processed with the half-wave rectifier, $\alpha = 0.5$, was used. Since the channels are less correlated in this case, there are only minor differences in the MSE before and after the transmission.

---

1Since we normalize with the power of the echo. We can regard this as the inverse of the echo return loss enhancement (ERLE).
Fig. 8. Subband convergence comparison between the two-channel FRLS and NLMS algorithms. The figures show the MSE performance for one typical lower and one typical higher subband. Solid line FRLS, dashed line NLMS. (a) Subband 7 (center frequency at 1.75 kHz) and (b) subband 18 (center frequency at 4.5 kHz).

Fig. 9. Fullband convergence comparison between the two-channel FRLS and NLMS algorithms. The figure shows the MSE performance for different SAECR setups. Solid line: FRLS is used in all subbands. Dashed line: NLMS is used in all subbands. Dotted line: FRLS in the lower 16 subband, NLMS in the upper 17 subbands.

room speaker changes positions. Fig. 7 also indicates that the channel decorrelation is more important in the lower frequency regions than in the higher ones, in practical situations.

Other simulations have shown that when a background noise source, in our case fan noise from a personal computer, was emitted in the transmission room, the correlation between the channels was reduced. Convergence of the adaptive filters was improved, especially in the higher frequency regions. Though channel decorrelation is still needed in normal office environments, especially in the lower frequency regions.

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. 8, the MSE performance of the FRLS and the NLMS algorithm are shown for one typical lower and one typical higher subband. The impressive performance gain of the FRLS algorithm only applies to the lower subbands. The performance of a system with FRLS in the lower and NLMS in the higher subbands is shown in Fig. 9.

Finally, a double-talk situation is shown. In contrast to the previous figures, the receiving room is simulated, using 4096 taps long room impulse responses. This in order to be able to remove the double-talk signal before calculating the MSE, (17). The signal was processed with the half-wave rectifier, \( \alpha = 0.5 \), and the two-channel FRLS algorithm was used in all subbands. The result is shown in Fig. 10.

The real-time system also includes a device to suppress the residual echo after the adaptive filter (see Fig. 2). This device was however not used in the simulations. The suppressor consists of three parts. The first is a short-time transmission room energy based suppressor, which increases suppression as the speech energy in the transmission room increases. The second suppressor, an echo path gain based suppressor [25], can be viewed as a mild form of center clipping in that if the residual echo is very strong, it is left unaffected, but when it is below a threshold, it is attenuated by an amount roughly proportional to the residual echo signal. Finally, comfort noise is added to the residual echo signal. Without comfort noise, the listener may be annoyed by the rapid changes of suppression by the two suppressors.

The subband structure enhances the system in several ways, including reducing the calculation complexity as has been shown in this paper. Another important advantage is the ability to run the adaptive algorithms on parallel processing units. In the real-time system, the analysis and synthesis filterbanks are processed on one DSP, whereas the adaptive filters are distributed over several DSPs. To be more specific, the adaptive filter, \( h_{\text{FRLS}} \) in Fig. 3, is distributed, whereas the filtering, \( \hat{h} \) in Fig. 3, is performed on the DSP which also executes the filterbanks [26]. This way, no extra signal path delay is introduced by the parallel structure. As the inherent transmission signal delay is a clear disadvantage of subband structures, this is of importance. Examples of the delay introduced by the filterbank are given in Table II.

V. SUMMARY

With the real-time implementation of the 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 aurally separate the speakers in video conferencing systems. Therefore a listener in the receiving room has an improved ability to distinguish one speaker in the transmission room, when other speakers, also situated in the transmission room but at other positions, are talking at the same time.
It has also been confirmed that decorrelation is crucial for stability of the system, both in real-time experiments and in off-line simulations on real-life recorded signals presented in the previous section. The studies also confirm 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 transmission room background noise is low. Finally it is shown that the two-channel FRLS adaptive algorithm is superior to the NLMS algorithm in the lower subbands, but the performance gain is less in the upper subbands.

The RLS algorithm is notorious for its stability problems. However, with the stabilization enhancements presented in this paper, and a proper initialization, it has been possible to use the algorithm in a controlled manner, and it has shown a very fast convergence rate. Even if tracking/reconverging is somewhat slower than initial convergence (see Fig. 10). Double-talk situations can severely decrease the performance of the adaptive filter. In the system, the two-path structure handles these situations. It also takes care of problems when the FRLS is restarted. Restarts are necessary for stabilization of the FRLS, e.g., in a simulation where 24 h of real-life data was processed, each subband was restarted on average every 18 s.

Finally, the authors would like to comment simulation data used in this paper. Simulations have been conducted under several different situations. It was chosen to use data with fairly high SNR (40 dB) in the paper. This since background noise will decorrelate the channels, and thereby reduce the "stereophonic" problem. That is, with lower SNR the MSE for a converged system will increase, but the increase of MSE due to transmission room speaker change [Fig. 6(a)] will be less obvious.

**APPENDIX I
ANALYSIS FILTERBANK**

The signal for subband \( m \), \( x_m(k) \), is calculated by bandpass filtering and downsampling of the fullband input signal \( x(n) \)

\[
x_m(k) = \sum_{l=0}^{K-1} f_m(l)x(rk - l)
\]

where the subband filter is denoted \( f_m(n) \) with length \( K \) and the downsampling factor is \( r \). The subband filters, \( f_m(n) \), are modulated versions of the low-pass prototype filter \( f(n) \). If \( M \) denotes the number of subbands in the filterbank, the individual subband filters can be expressed with the modulation, \( f^T_m = w^T_m F \), where

\[
f_m = [f_m(0) f_m(1) \cdots f_m(K-1)]^T
\]
\[
w_m = [1 e^{j2\pi m/1} \cdots e^{j2\pi m(M-1)}]^T
\]
\[
F = \text{diag}(\hat{f}_0) \cdots \text{diag}(\hat{f}_M) \cdots \text{diag}(\hat{f}_{K-M}).
\]

The prototype filter matrix \( F \) is of size \( M \times K \), and \( \text{diag}(\hat{f}_i) \) denotes a diagonal matrix with elements from the prototype filter as

\[
\text{diag}(\hat{f}_i) = \begin{bmatrix} f(i) & 0 & \cdots & 0 \\ 0 & f(i+M-1) & \cdots & 0 \end{bmatrix}.
\]

Due to the downsampling, \( r \) new input samples are needed for each new subband sample, denoted a frame. Now the output sample in subband \( m \), frame \( k \) can be expressed as \( x_m(k) = w^T_m F x(k) \), where the fullband input vector \( x(k) \) is defined as

\[
x(k) = [x(rk) x(rk-1) \cdots x(rk - K + 1)]^T.
\]

By exchanging the modulation vector \( w_m \) for the modulation matrix

\[
W = [w_0 \ w_1 \ \cdots \ w_{M-1}]^T
\]

a vector containing all subband samples at frame \( k \) may be calculated as

\[
[x_0(k) \ x_1(k) \ \cdots \ x_{M-1}(k)]^T = WFx(k).
\]

The efficient DFT polyphase filterbank structure can be seen in (26). Since \( F \) is a real sparse matrix, \( WFx(k) \) can be calculated with \( K \) real multiplications. Secondly, the modulation matrix divided by the number of subbands, \( W/M \), is also known as the inverse DFT matrix. Therefore, the calculation complexity of (26) can be reduced by using efficient inverse fast Fourier transforms (IFFT). Let us return to (25) once again. Due to the symmetry in the modulation vector (21), the following relation between rows can be seen in (25)

\[
x_i(k) = w^T_{M-i} x(k) \quad 1 \leq i \leq \frac{M}{2} - 1
\]

where \( * \) denotes the conjugate operator. Consequently, \( x_{M-i}(k) = x^*_i(k) \) as long as the input signal \( x(n) \) and the elements of the prototype filter matrix \( F \) are real-valued. Therefore, it is only necessary to calculate the first \( M/2 + 1 \) subbands, and also only necessary to apply the adaptive filters, the echo canceler, in the \( M/2 + 1 \) lowest subbands. It should be noted that complex valued adaptive filters are needed.

**APPENDIX II
SYNTHESIS FILTERBANK**

The synthesis filterbank reconstructs the fullband signal, \( c(n) \), by a summation of the interpolated and filtered subband signals \( c_m(k) \). For reasons that will be seen later in the section, it is advantageous to use a state description of the filterbank filters. The input signal to the filterbank, \( c_m(k) \), is upsampled \( r \) times. Then, \( K \) copies of the interpolated signal are each multiplied with the corresponding synthesis filterbank filter-tap weight, \( g_m(l) \). Each sample factor is added to a state variable \( s_m(n) \), in the state vector \( s_m(n) \). For every fullband sample interval, the state vector is shifted one position. The fullband
signal can then be calculated as the sum of the upper state variables, \( e(n) = \sum_{m=0}^{M-1} s_{m0}(n) \). Let us now define the state vector as one upper and one lower vector

\[
\begin{align*}
\mathbf{s}_m(k) &= \begin{bmatrix} s^u_m(k) \\ s^l_m(k) \end{bmatrix} \\
\mathbf{s}_{m0}(k) &= \begin{bmatrix} s_{m0}(k) \\ \vdots \\ s_{m0}(M-1) \end{bmatrix} \\
\mathbf{s}^l_m(k) &= \begin{bmatrix} s_{m1}(k) \cdots s_{mM-1}(k) \end{bmatrix}^T
\end{align*}
\]

Due to the interpolation, every subband sample \( e_m(k) \) is followed by \( r-1 \) zeros. For these zeros, the state vectors are only updated with shifts. The output signal \( e(n) \) can therefore be calculated for \( r \) samples, i.e., one frame at a time

\[
\begin{align*}
\begin{bmatrix} e(kr - r + 1) & e(kr - r + 2) & \cdots & e(kr) \end{bmatrix}^T &= \sum_{m=0}^{M} \mathbf{s}^u_m(k).
\end{align*}
\]

Similar to the analysis filters, the synthesis filterbank filters \( g_m(n) \) are modulated versions of the low-pass prototype filter \( g(n) \). Therefore, the filter vector

\[
\mathbf{g}_m = \begin{bmatrix} g_m(0) & g_m(1) & \cdots & g_m(K-1) \end{bmatrix}^T
\]

is defined as \( \mathbf{g}_m = \mathbf{Gw}_m \). Here, \( \mathbf{G} \) is a sparse prototype matrix of size \( M \times K \)

\[
\mathbf{G} = \begin{bmatrix} \text{diag}(\mathbf{g}_0) & \text{diag}(\mathbf{g}_M) & \cdots & \text{diag}(\mathbf{g}_{K-M}) \end{bmatrix}^T
\]

where \( \text{diag}(\mathbf{g}_0) \) is defined as in (23) and \( \mathbf{w}_m \) is defined in (21).

As previously mentioned, the state vectors can be updated on a frame basis. The state vectors are shifted \( r \) positions and the input subband sample, multiplied with \( g_m \), is added

\[
\mathbf{s}_m(k) = \begin{bmatrix} \mathbf{s}^u_m(k-1) \\ 0_{r \times 1} \end{bmatrix} + \mathbf{Gw}_m e_m(k).
\]

When reconstructing the output signal \( e(n) \), it is enough to know the sum of the state vectors, as is shown in (31). Therefore, the state vectors do not need to be updated individually, instead it is sufficient to update the sum of the state vectors, as

\[
\begin{align*}
\mathbf{s}(k) &= \begin{bmatrix} \mathbf{s}^u(k-1) \\ 0_{r \times 1} \end{bmatrix} + \mathbf{GWe}^\text{sub}(k)
\end{align*}
\]

where \( \mathbf{W} \) is defined in (25) and \( \mathbf{e}^\text{sub}(k) = \begin{bmatrix} e_0(k) & e_1(k) & \cdots & e_{M-1}(k) \end{bmatrix}^T \). Finally, (35) can be calculated with a computationally efficient IFFT algorithm, and the output for frame \( k \) is then the upper \( r \) elements of the state vector.

**Appendix III**

**Prototype Filter for Noncritical Downsampling**

In this Appendix, we will formulate a filterbank design method as a minimization problem.

Since aliasing in the filterbank will drastically decrease the performance of the SAECR, aliasing suppression is an important property of the prototype filter. Hence, alias suppressing subbands in combination with noncritical down-sampling, allow us to neglect aliasing cancellation in the filterbank. The only requirement for the filterbank, neglecting small amounts of aliasing, is to ensure that the analysis–synthesis system is a pure delay,

\[
\frac{1}{r} \sum_{m=0}^{M-1} F \left( e^{i(\omega - \frac{2\pi m}{r})} \right) G \left( e^{i(\omega - \frac{2\pi m}{r})} \right) = e^{-i\omega K}
\]

where \( K \) is the prototype filter length, \( M \) is the number of subbands, \( r \) is the downsampling factor, \( F(z) = \sum_{k=0}^{K-1} f(k)z^{-k} \) is the analysis filter, and \( G(z) = \sum_{k=0}^{K-1} g(k)z^{-k} \) is the synthesis filter. In order to guarantee linear phase, we let \( G(z) = z^{-K}F(z^{-1}) \). If we define the noncausal filter \( R(z) = F(z)F(z^{-1}) \), which only differ from the analysis-synthesis system, \( F(z)G(z) \), by a delay, an equivalent requirement to (36) would be

\[
\frac{1}{r} \sum_{m=0}^{M-1} R \left( e^{i(\omega - \frac{2\pi m}{r})} \right) = 1.
\]

Now, we can formulate the filter design problem as a minimization problem. First we need to make sure that the passband is sufficiently flat, by minimizing

\[
\Phi_1 = \int_{0}^{\frac{\pi r}{2}} [R(e^{i\omega}) - r]^2 d\omega.
\]

The region where two adjacent bands overlap also needs to be flat, i.e., we minimize

\[
\Phi_2 = \int_{\frac{\pi r}{2}}^{\pi} \left[ R(e^{i\omega}) + R \left( e^{i(\omega - \frac{2\pi}{r})} \right) - r \right]^2 d\omega.
\]

Finally, we need to minimize aliasing by enforcing good stop-band attenuation

\[
\Phi_3 = \int_{\frac{\pi}{2}}^{\pi} [R(e^{i\omega})]^2 d\omega.
\]

The total minimization problem can now be expressed as

\[
\min_{R(z)} \alpha_1\Phi_1 + \alpha_2\Phi_2 + \alpha_3\Phi_3
\]

subject to the constraints

\[
\begin{align*}
\alpha_i &\geq 0 & i = 1, 2, 3 \quad \text{(42)} \\
r(0) &\geq r(n) & n \neq 0 \quad \text{(43)} \quad \forall n
\end{align*}
\]

where \( \alpha_i \geq 0 \) are tradeoff parameters and \( R(z) = \sum_{k=0}^{K-1} r(k)z^{-k} \). Quadratic programming [27] has been found to be a fast and stable method to solve this minimization problem. The filters used in the simulations are designed with this method, and are shown in Fig. 11. In [28], a method for calculating \( F(z) \) from a given \( R(z) \) is presented. An alternative method to design filters that satisfies (36) can be found in [29].
This method has more stringent requirements, forcing $\Phi_1$ and $\Phi_2$ to be equal to zero, but also needs a larger filter length $K$ in order to achieve good stopband attenuation.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. W. Elko without whose enthusiastic engagement this project would not have been nearly as successful. The authors also highly value the dialogue with Dr. D. Morgan during the project.

REFERENCES


Peter Eneroth was born in Kristianstad, Sweden, in 1969. He received the M.S. degree in electrical engineering and the Ph.D. degree in signal processing from Lund University, Lund, Sweden, in 1995 and 2001, respectively.

During the Fall of 1998 and 1999, he was with Bell Labs, Lucent Technologies, Murray Hill, NJ, as a Consultant. Since March 2001, he has been with Telia Research, Sweden. His research interests include adaptive filtering, subband signal processing, echo cancellation, and real-time implementations.

Steven L. Gay was born in Kansas City, MO, in 1958. He received the B.S. degree in electrical engineering from the University of Missouri, Rolla, in 1979, the M.S. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1981 under the Bell Labs one-year-on-campus (OYOC) program, and the Ph.D. degree in electrical engineering from Rutgers, The State University of New Jersey, New Brunswick, in 1994.

He joined Bell Labs, Whippany, NJ, in 1980. After working on speech coding at JTT-Defense Communications Division from 1982 to 1984, he returned to Bell Labs to work in speech coding, echo cancellation, and other digital signal processing research and applications in various development departments. In 1991, he joined the Acoustics and Audio Research Department, Bell Labs, Murray Hill, NJ. His main research topics include network and acoustic echo control, efficient adaptive filtering algorithms, and real-time implementations. He co-edited the book Acoustic Signal Processing for Telecommunication.

Dr. Gay co-chaired the 1999 International Workshop on Acoustic Echo and Noise Control and continues to serve on its standing committee.
Tomas Gänsler (S’91–M’96) was born in Sweden in 1966. He received the M.S. degree in electrical engineering and the Ph.D. degree in signal processing from Lund University, Lund, Sweden, in 1990 and 1996, respectively.

From 1997 to September 1999, he was an Assistant Professor at Lund University. During 1998 he was with Bell Labs, Lucent Technologies, Murray Hill, NJ, as a Consultant; in October 1999, he became Member of Technical Staff. His research interests include robust estimation, adaptive filtering, mono/multichannel echo cancellation and subband signal processing.

Jacob Benesty (M’98) was born in Marrakesh, Morocco, in 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Orsay, France, in April 1991. During his Ph.D. studies, he worked on adaptive filters and fast algorithms at the Centre National d’Etudes des Telecommunications (CNET), Paris, France.

From January 1994 to July 1995, he was with Telecom Paris working on multichannel adaptive filters and acoustic echo cancellation. He joined Bell Labs, Lucent Technologies (formerly AT&T), Murray Hill, NJ, in October 1995, first as a Consultant and then as Member of Technical Staff. Since then, he has been working on stereophonic acoustic echo cancellation, adaptive filters, source localization, robust network echo cancellation, and blind deconvolution. He is co-editor/co-author of the book Acoustic Signal Processing for Telecommunication.

Dr. Benesty was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control.