Stereophonic Acoustic Echo Cancellation: Theory and Implementation

Eneroth, Peter

2001

Stereophonic Acoustic Echo Cancellation: Theory and Implementation

Peter Eneroth

Lund 2001

Department of Electroscience
Lund University
Box 118, S-221 00 LUND
SWEDEN

No. 22
ISSN 1402-8662
ISBN 91-7874-110-6

Printed in Sweden by KFS AB, Lund.
January 2001

Abstract

The thesis treats theory and implementation aspects for stereophonic acoustic echo cancellation.

In Paper I a complete implementation of a stereophonic acoustic echo canceler based on the two-channel fast recursive least-squares algorithm is presented. An analysis of the system calculation complexity is also given, in addition to simulation results on recorded real-life data.

Paper II presents a comparison between adaptive filters for usage in stereophonic acoustic echo cancellation. The comparison includes, in addition to the standard normalized least mean square algorithm, the two-channel fast recursive least-squares algorithm and a two-channel frequency-domain adaptive algorithm. In the paper, the convergence rate, the calculation complexity, the signal transmission delay, and the memory usage for the evaluated systems are shown.

Adaptive filters applied in subband structures may need to model a few non-causal taps even if the fullband impulse response is causal. This phenomenon is analyzed in Paper III. Formulas to calculate the number of non-causal taps needed are also presented in the paper.

The fundamental problem in stereophonic acoustic echo cancellation is the signal correlation between the two channels. In Paper IV it is shown how a perceptual audio coder may decrease this correlation and thereby increase the performance of the stereophonic acoustic echo canceler.

Paper V investigates the possibilities of using joint subband filterbanks or time to frequency-domain transforms for echo cancellation and perceptual audio coding. The usage of joint filterbanks/transforms not only decreases the computational complexity of the system, it also reduces the total signal transmission delay of the system if properly designed.

Contents

Acknowledgment

Without all the help and encouragement I have received during the last five years, the contents of this thesis would be significantly different, and I am truly grateful to everyone who has made this work possible. Not only have I been given a lot of help, I have also been given the opportunity to become friends with so many generous, interesting and intelligent people.

I am without competition most grateful towards my supervisor Tomas Gänsler. I have always felt totally confident with his knowledge and his support to me. He has guided me towards very interesting technical problems, and in times when I have been too committed in solving these problems, he has forced me to write. After leaving the department, Tomas continued being fully committed as my supervisor.

I would also like to thank Göran Salomonsson, now retired from his position at the department. It was Göran who accepted me as a graduate student and who guided me in my early work.

Inspiration is a fundamental corner-stone behind research, and much of my inspiration has come from Bell Laboratories. Jacob Benesty taught me plenty about multichannel adaptive filtering, and has generously shared everything from knowledge to deep friendship. During the fall of 1998, Steven Gay and I worked on the project presented in Paper I of this thesis. We spent many evenings at the labs, together fighting against time and software bugs. Also Steve generously shared his knowledge and friendship. Dennis Morgan was always happy to discuss everything from mathematical and technical details to international politics. Being a perfectionist, he also taught me plenty about technical writing. Finally, I would like to mention the head of the group, Gary Elko, who enthusiastically pushed the work in the right direction, and everybody who helped creating such a friendly atmosphere at Bell Laboratories, Lucent Technologies.

When Tomas Gänsler and I started our work on stereophonic acoustic echo cancellation, it was a joint project together with Telia Research AB in Haninge, now situated in Farsta. Our cooperation was extremely fruitful, and I will always be grateful to Ove Till, Niklas Johansson and Birgitte Wikman.

Most of the time was of course spent at the Department of Applied Electronics, now the Department of Electroscience. Academic and educational inspiration was given by Leif Sörnmo and Per Ola Börjesson, and administrative issues were easily

resolved with help from Birgitta Holmgren. Clas Angvall has been supportive in diverse matters and Erik Johnsson has provided me a bullet proof computer system. I'm also grateful to all colleagues at the department, making almost every day at the department joyful.

Finally, I would like to thank my wife, Elisabeth Anderberg, my family and my friends for encouragement, and for giving me quality of life.

General Introduction

Echo: *"the repetition of a sound caused by reflection of sound waves."*

Merriam-Webster's Dictionary

Historically, echo was probably an interesting and amusing phenomenon that could be experienced in rare places, e.g. between mountain walls in the Alps. However, to most people today, echo is also something annoying which can occur, e.g., during a long distance telephone call. The definition of echo is still valid though, it is all about (delayed) reflections of sound waves.

In this thesis, methods for removing echo in communication systems are proposed. While most of the theories presented here can be applied to reduce echoes in telephone calls, the target application is hands-free communications systems, with dual audio channels. Examples include stereophonic video conferencing systems and desktop conferencing on computers with dual loud-speakers. The thesis will both consider the theoretical problems and implementation aspects.

This part includes general background information needed in order to get the other parts of the thesis in a proper context. Here the underlying application is discussed, together with notes about the historical background. The theoretical problem in echo cancellation is shown and basic echo cancellation methods are discussed. Finally, Section 5 summarizes the five parts of the thesis.

1 Communication System

For the purpose of this thesis, a communication system is a system that allows two or more persons, or perhaps one person and one machine, to interact with speech over an artificial channel, for example a telephone line. This system may also deliver other components, such as pictures, video, documents, or any other type of media, but in this thesis we will only discuss the audio component. The most commonly used system

in this category is of course the standard telephone system, also called the public switched telephone network (PSTN), but in recent years many additional systems have emerged, e.g., cellular telephones, video conferencing, and desktop conferencing.

Natural speech signals are analogue signals, and they can be modeled as continuous functions, with infinitely high resolution. Such functions are very ill-suited for today's digital communication and computer systems. Therefore, in the systems under consideration, the speech signals are sampled and quantized. This process involves measuring the amplitude of the signal at given time intervals (sampling) and partitioning the signal amplitude span in sub-intervals, and representing each sub-interval with a number (quantization). The digitized signal can advantageously be processed with a digital system, e.g., a computer, and be transmitted via a digital communication system, such as a modern PSTN.

The quality of the digital signal can be made arbitrarily good, by decreasing the time between successive samples and by increasing the number of quantization intervals. Since the bandwidth of the original speech signal is limited, as is the sub-group of audio signals that are audible to humans, it is not controversial to state that high quality sampled audio signals are extremely redundant. That is, it is possible to reduce the amount of data used to represent the signal without, or with very small, loss of quality. The sampling and quantization of the analog signal can be viewed as one form of data compression, and later in the thesis we will see other compression methods, where more mathematical methods are used to reduce the redundancy further.

Almost everyone who has made an intercontinental telephone call, or a call to a cellular telephone in a poor network, has experienced echoes. This thesis presents methods for how to reduce echoes in communication systems, with the emphasis on echoes due to an acoustic coupling between the loud-speaker and the microphone. Three papers in the thesis focus on systems with two audio channels, stereophonic audio, describing why this situation is more complicated and proposes possible solutions for stereophonic acoustic echo cancellation.

2 Echo Cancellation

Echoes in the PSTN have been a problem since the early days of telephone technology, when the first transcontinental telephone networks were built. The long distance between the two parties introduces a time delay to the echo signal. This delay in combination with low echo attenuation reduces the perceptual quality of the system. Fortunately, the development of high speed transmission systems in the 1920's reduced the problem [1].

In the 1960's, satellite communication was introduced, and again the round-trip delay became long enough to cause noticeable echo problems. A device called echo

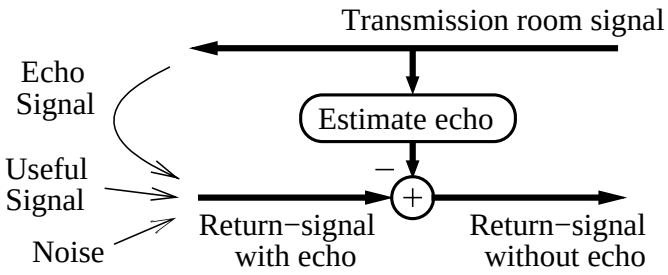

Figure 1: The echo canceler estimates the echo signal, and subtracts it from the return signal.

suppressor which had been used for some time became the widely used solution. In principle, the device identifies which one of the two parties who is talking, and accordingly attenuates the transmitted signal in the opposite direction [2]. The disadvantage with this solution is that major speech detection errors, results in attenuation of the actual speech, not only the echo.

In the late 1960's, the foundation of today's echo eliminating technology, the echo canceler, was invented by J. L. Kelly and B. F. Logan, and presented in a paper by M. M. Sondhi [3]. These gentlemen also filed two patents on the same day [4], [5]. This device adaptively estimates the echo-path transfer function, and subtracts an estimated echo from the returning signal, Fig. 1. The device requires no à priori information of the echo-path transfer function, instead it recursively estimates the echo-path transfer function, using an *adaptive filter*. Even though an analog implementation was presented in the same paper [3], useful implementations had to wait for improvements in digital signal processing technology. In today's public switched telephone network, this type of echo canceler is standard in most transmission lines with a long enough round-trip delay [6].

Echo only arise in situations where the forward and backward transmission lines are not completely separated. In the PSTN, different nodes in the network are interconnected with a 4 wire-line, i.e., the signals from subscriber A to B are transmitted in channels separate from the signals from subscriber B to A. At each end, beyond the switching center in the national network, the 4-wire transmission is converted to 2-wire transmission via a hybrid coil circuit before the connection to the telephone set, see Fig. 2. This conversion circuit is the source of the echo, since a fraction of the signal traveling on the A-to-B path returns on the B-to-A path. This is denoted *line echo*. The leakage is due to impedance mismatch in the hybrid coil [1]. Because of impedance differences between different 2-wire local telephone lines and impedance differences between different telephone sets, perfect impedance match is not possible to achieve. Fortunately, these echo path transfer functions are usually well modeled by a rather short linear FIR-filter [7], and the echo can in most situations relatively

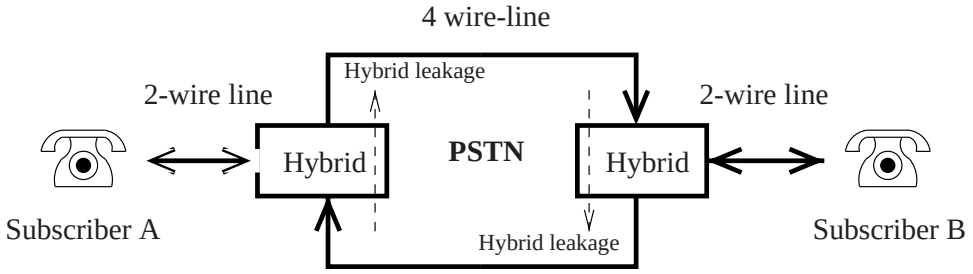

Figure 2: Telephone network model. Telephone subscribers are connected to the closest node in the telephone system via a 2-wire cable. A hybrid coil connects the 2-wire line with the 4-wire line used in the national and international telephone systems. The cause of the echo is leakage in the hybrid.

easily be removed from the return channel with an echo canceler. However, as mentioned above, it is only in the situation with significant signal transmission delay that the echo disturbs the conversation [1]. This signal transmission delay can be caused by the physical distance between the two subscribers, but also by advanced speech coders and radio channel coding schemes used in e.g. the cellular telephone network.

In many situations it is an advantage if a hands-free communication system can be used instead of a handset. Hands-free systems include video conference systems, and also hands-free mobile telephones which are becoming a legal requirement in some countries for using a telephone while driving a car. In these systems, there is another cause of the echo, namely the acoustic coupling between the loud-speaker and the microphone in the receiving room, see Fig. 3. Depending on the acoustic properties of the room, the echo path can have a rather long impulse response (several seconds) and the transfer function may be non-linear. The non-linearity is mainly due to non-linear elements in the loud-speaker and the use of high-quality loud-speakers usually reduces it to the extent where it is sufficient to use a linear model in the echo canceler. Another significant difference from the line echo situation, is the constant change of the echo path transfer function. The smallest change in the room, such as any movements or even a temperature change can significantly change the echo path transfer function. Therefore, the adaptive algorithm used to estimate the transfer function is required to have good tracking properties.

Throughout this thesis we will use the following room definitions for a video conference situation: *The transmission room*, is where the excitation signal is created, e.g., by a speaker and *the receiving room* where the acoustic coupling between the loud-speaker and the microphone is the cause of the echo. That is, the echo canceler needs to estimate the echo path in the receiving room.

In order to be able to derive an algorithm that can estimate the transfer function, we need a mathematical model of the echo signal. We start by naming the excitation

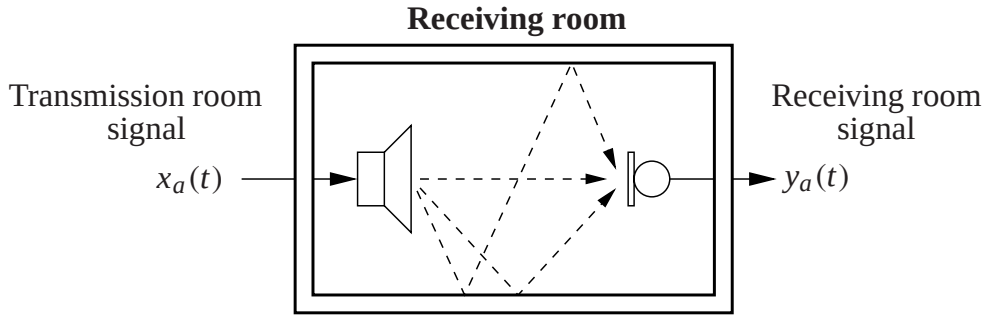

Figure 3: Acoustic echo: An infinite number of reflections of the loud-speaker signal is mixed in the microphone signal.

signal $x_a(t)$, a continuous function where t denotes the time, see Fig. 3. Similarly, the microphone signal is denoted $y_a(t)$. In Fig. 3, it can also be seen that the physical distances for the reflections vary, and they will therefore arrive at the microphone at different time points. Each reflection is an individually attenuated and delayed version of the excitation signal $x_a(t)$. If we denote the attenuation of the sum of all reflections with the delay τ with $h_a(\tau)$, we can write the received signal as an integral over all possible delay values,

$$y_a(t) = \int_0^\infty h_a(\tau)x_a(t-\tau)d\tau + v_a(t), \tag{1}$$

The function $h_a(t)$ is called the impulse response of the system, and $v_a(t)$ represents the signals generated in the receiving room, called "useful signal" in Fig. 1. The impulse response is said to be causal, since $h_a(\tau) = 0$ for $\tau < 0$. That is, the output signal of the causal system is only dependent on present and past input signals.

In this thesis we will consider sampled systems, i.e. we will only have knowledge of the signal amplitude at specific time instances. For example, if we sample the signal $x_a(t)$ with a time interval of Δ_t, the time discrete version of the excitation signal $x(n)$ can be written as $x(n) = x_a(n\Delta_t)$, $n \in \mathbf{Z}$. The discrete version of (1) can be written as,

$$y(n) = \sum_{l=0}^{\infty} h(l)x(n-l) + v(n). \tag{2}$$

In (2), the signal $x(n)$ is filtered with the filter $h(n)$. This filter-operation is called convolution, and will be denoted with the symbol "$*$". That is, (2) may also be written as, $y(n) = h(n) * x(n) + v(n)$.

The echo canceler should estimate $h(n)$, and this estimate is denoted $\hat{h}(n)$. Even if $h(n)$ has an infinite length, i.e. infinite number of taps, the taps will decay with

time index n. Therefore, the estimate $\hat{h}(n)$ will in a practical situation have a limited number of taps, and in the following it will have L taps. The echo canceler will have knowledge of the excitation signal $x(n)$, see Fig. 1. With an impulse response estimate $\hat{h}(n)$, the echo canceler estimates the echo signal as,

$$\hat{y}(n) = \sum_{l=0}^{L-1} \hat{h}(l)x(n-l). \tag{3}$$

By subtracting this estimate from the return signal, the echo canceler is able to cancel the echo, and the canceled signal, sometimes called the residual echo signal, is simply,

$$e(n) = y(n) - \hat{y}(n) \approx v(n). \tag{4}$$

The approximation is due to the fact that $\hat{h}(n)$ only is an approximation of $h(n)$.

In this thesis the performance of echo cancelers is discussed. In order to do so, we use the following performance index,

$$\varepsilon = \frac{\|\mathbf{h} - \hat{\mathbf{h}}\|}{\|\mathbf{h}\|}, \tag{5}$$

where $\|\cdot\|$ denotes the l_2-norm of a vector and $\mathbf{h} = \begin{bmatrix} h(0) & \cdots & h(L-1) \end{bmatrix}$. This performance index is called the misalignment and measures the mismatch between the true and estimated impulse response of the receiving room. The misalignment is a very accurate measurement of the performance of an echo canceler, however, it requires knowledge of the true impulse response, $h(n)$. This knowledge is only available in simulations, where the receiving room signal, $y(n)$, is artificially calculated from a known transmission room signal, $x(n)$, and known impulse response $h(n)$. In this thesis, most simulations are conducted on real-life data, where both the signals $x(n)$ and $y(n)$ are recorded in an authentic room. In such simulations, we do not know the true impulse response, $h(n)$. As this function also may change over time, it is actually impossible to derive an exact estimate of $h(n)$. In these situations, an alternative performance index is needed.

Another performance index is the mean square error (MSE) energy of the residual echo signal. The MSE is given by,

$$\text{MSE} = \frac{\text{LPF}[e(n) - v(n)]^2}{\text{LPF}[y(n) - v(n)]^2}, \tag{6}$$

where LPF denotes a lowpass filter. We do not need any knowledge of the impulse response $h(n)$ in order to calculate the MSE performance of a system. In the recorded real-life data situation described above, we have no knowledge of the background

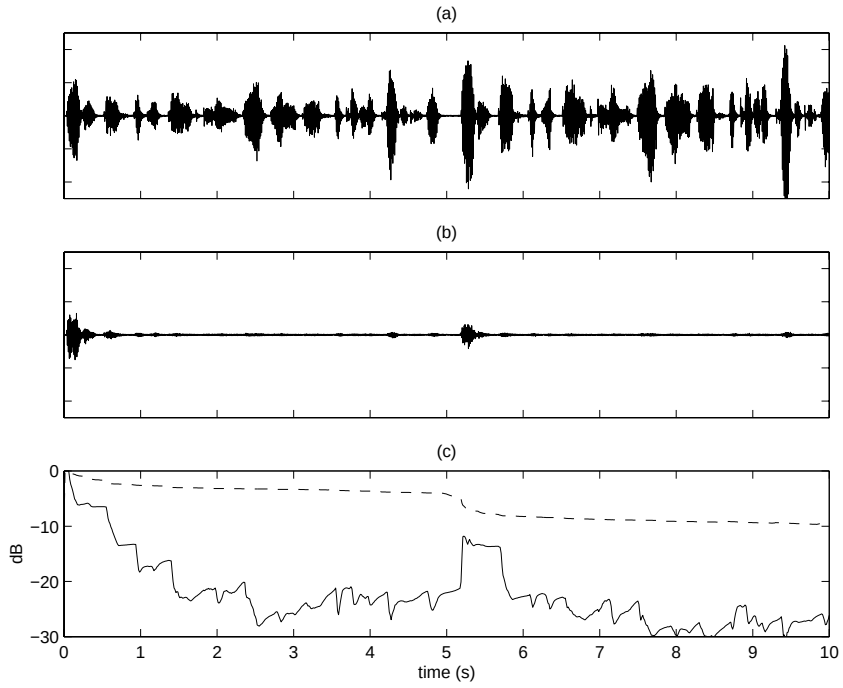

Figure 4: Echo canceler performance example. (a) Echo signal, $y(n)$. (b) Residual echo signal $e(n)$. (c) MSE (solid line) misalignment (dashed line).

noise or other signals generated in the receiving room, denoted $v(n)$ in (6). However, if $v(n)$ is small, it can be neglected in the calculation of the MSE.

A third popular performance index is the echo return loss enhancement (ERLE). In principal, the ERLE can be regarded as the inverse of the (normalized) MSE eq. (6), and since the MSE has been commonly used in earlier papers investigating stereophonic acoustic echo cancellation, the MSE was chosen as the preferred performance index in this thesis.

In Fig. 4, the result of a stereophonic acoustic echo canceler in a video conferencing situation is illustrated. In Fig. 4(a) the echo signal situation before the echo canceler is plotted and 4(b) is the situation afterwards. In Fig. 4(c) the MSE and the misalignment performance indexes are shown. In this situation, we have fair suppression of echo, however, the misalignment shows that the impulse response estimate $\hat{h}(n)$ is far from being equal to the true impulse response $h(n)$. How this can be possible will be discussed in the next section.

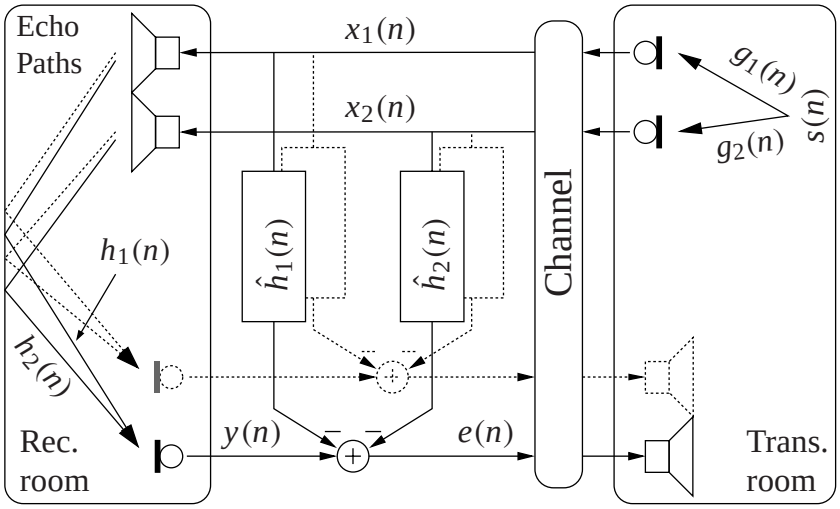

Figure 5: Schematic diagram of stereophonic acoustic echo cancellation.

2.1 Stereophonic Acoustic Echo Cancellation

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 [8]. However, there are now four acoustic echo paths to identify – two to each microphone, see Fig. 5. This will not only cause increased calculation complexity, but also a new fundamental problem of the solution, as we shall see.

The fundamental problem is that the two audio channels may carry linearly related signals which in turn may cause 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 we will show that all (but the physically true) solutions depend on the transmission room.

Four monophonic echo cancelers straightforwardly implemented in the stereo case would not only 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 person stops talking and another starts talking from a different location in the transmission room. There is currently 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 monophonic AEC in the stereo case does not give satisfactory performance.

If we assume that the transmission room microphone signals, Fig. 5, are given by,

$$x_i(n) = g_i(n) * s(n),\ i = 1, 2, \tag{7}$$

where $s(n)$ is the source signal in the transmission room and $g_i(n),\ i = 1, 2$, are the transmission room echo paths. 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 echo canceler is

$$e(n) = y(n) - \hat{y}(n), \tag{8}$$

$$y(n) = h_1(n) * g_1(n) * s(n) + h_2(n) * g_2(n) * s(n), \tag{9}$$

$$\hat{y}(n) = \hat{h}_1(n) * g_1(n) * s(n) + \hat{h}_2(n) * g_2(n) * s(n). \tag{10}$$

That is, the residual echo signal, to be minimized by the echo canceler, can be written as,

$$e(n) = \left\{\left[h_1(n) - \hat{h}_1(n)\right] * g_1(n) + \left[h_2(n) - \hat{h}_2(n)\right] * g_2(n)\right\} * s(n). \tag{11}$$

The adaptive algorithm in the echo canceler recursively tries to minimize the sum of the squared residual echo signal, $e^2(n)$, by finding suitable values for the filters $\hat{h}_1(n)$ and $\hat{h}_2(n)$. As expected, one possible solution is $\hat{h}_1(n) = h_1(n)$ and $\hat{h}_2(n) = h_2(n)$, but in contrast to monophonic echo cancellation, this in not the only solution. This can be shown with a simple example. Let $g_1(n) = a$, $g_2(n) = b$, $h_1(n) = c$, and $h_2(n) = d$. Then the squared residual echo signal $e^2(n) = 0$ for the true solution $\hat{h}_1(n) = h_1(n)$ and $\hat{h}_2(n) = h_2(n)$, and for an infinite number of solutions, where the only requirement is the following relation between $\hat{h}_1(n)$ and $\hat{h}_2(n)$,

$$\hat{h}_1(n) = c + \frac{b}{c}(d - \hat{h}_2(n)). \tag{12}$$

The relation in (12) is, in contrast to the true solution, dependent on $g_1(n)$ and $g_2(n)$. Therefore, the echo canceler has to track changes in both the transmission room and receiving room, if it is unable to find the true solution.

Let us return to Fig. 4. After just a few seconds the echo canceler has found a solution that suppresses the echo with more than 20 dB. Still this is far from the true solution, which can be seen by studying the misalignment. After 5 seconds there is actually a change of $g_1(n)$ and $g_2(n)$. Since the echo canceler has not found the true solution, the residual echo signal increases. However, the misalignment will actually decrease. In this simulation there is a small amount of independent background noise added to the excitations signals $x_1(n)$ and $x_2(n)$. This noise is pushing the system slowly to a solution closer to the true solution, and the misalignment will therefore gradually decrease.

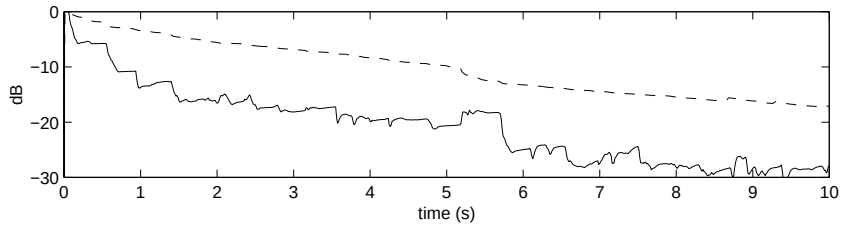

Figure 6: The simulation in Fig. 4 repeated, this time the excitation signals are processed with the non-linear function in (13), $\alpha = 1$.

In Part I, Section 2.1, an extensive description of the fundamental problem of stereophonic acoustic echo cancellation is given. The problem was originally described in [9], and then further analyzed and described in [10], [11], [12] and [13].

In [12] it is also concluded, that the practical solution to the fundamental problem of stereophonic echo cancellation is to reduce the correlation between channel $x_1(n)$ and $x_2(n)$. This reduction must of course be performed without decreasing the signal quality. In [12], the authors suggest that the transmission room signals can be processed with the following non-linear functions,

$$\begin{aligned} x_1'(n) &= x_1(n) + \alpha \frac{x_1(n) + |x_1(n)|}{2}, \\ x_2'(n) &= x_2(n) + \alpha \frac{x_2(n) - |x_2(n)|}{2}, \end{aligned} \tag{13}$$

where α determines the amount of added distortion. Since the functions are non-linear, the correlation between the channels will be reduced. For small α, the distortion is hardly audible in an office environment [14], and at the same time the non-linearity rather efficiently reduces the channel correlation. In Fig. 6, the simulation in Fig. 4 is repeated. This time the excitation signals have been processed with (13), $\alpha = 1$. Here, the adaptive filters are forced to converge to a solution that is closer to the true solution. Therefore, the MSE is decreasing slower and the misalignment is decreasing faster. The problem that occurs in Fig. 4 when $g_1(n)$ and $g_2(n)$ are altered after 5 s is now nearly removed. It should be noted that $\alpha = 1$ will in most situations add unacceptable amount of distortion to the signal. It is used here, however, in order to visualize the fundamental problem in stereophonic acoustic echo cancellation.

Other decorrelation methods have been presented. In [15], it is suggested that shaped, independent noise is added to the two signals $x_1(n)$ and $x_2(n)$. Adding white noise is not efficient enough, since it will either have to little energy to decorrelate the signals or it will be unacceptable for the listener [10]. Paper IV, originally published as [16] at the same time as [15], presents a method where an audio coder, such as an MPEG 1 layer 3 coder, is used to add shaped independent noise to the channels.

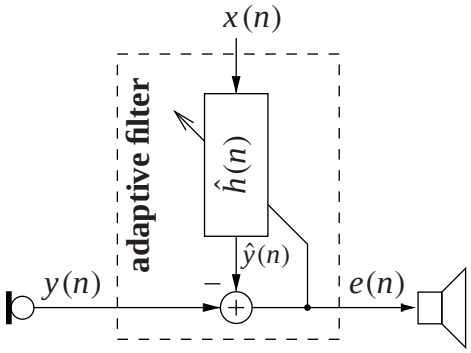

Figure 7: One adaptive filter extracted from Fig. 5.

This method is shown to efficiently reduce the correlation with only a small increase of calculation complexity in systems where an audio coder is needed. Other decorrelation methods have been studied in [17], [18], [19].

Stereophonic acoustic echo canceler implementations will have higher calculation complexity requirements than monophonic implementations since four adaptive filters with high performance algorithms are needed. In an early implementation [20], it was chosen to use a stereophonic echo canceler for frequencies below 1 kHz and monophonic echo canceler for the higher frequencies. A fullband stereophonic acoustic echo canceler is presented in [21] and Paper I. Implementations have also been conducted utilizing the affine projection algorithm [22] and the two-channel frequency-domain algorithm [23].

3 Adaptive Filtering

The echo path impulse response is estimated with an adaptive filter in the echo canceler, as we have seen in Fig. 5. This is done by recursively minimizing the sum of the squared residual echo signal $e^2(n)$. The block diagram of an adaptive filter in an echo canceler situation is illustrated in Fig. 7. As is depicted in this figure, the adaptive filter will only have knowledge of the transmission room signal, $x(n)$, and the receiving room signal, $y(n)$. Internally, the adaptive filter then estimates the receiving room impulse response, $h(n)$, and generates the residual echo signal, $e(n)$.

In an adaptive filter there is a cost function to be minimized. Usually this cost function is equal to the sum of the squared residual echo signal,

$$J(n) = \sum_{l=0}^{n} e^2(l), \tag{14}$$

and the minimization parameters are the L filter coefficients,

$$\hat{\mathbf{h}}(n) = \begin{bmatrix} \hat{h}(0,n) & \cdots & \hat{h}(L-1,n) \end{bmatrix}^T, \tag{15}$$

where $\hat{h}(l,n)$ is filter tap l after n recursive updates, and ^{T} denotes the matrix transpose operator. The adaptive filter requires no à priori information of the true impulse response $h(n)$, instead it begins with an initial value, say the zero vector $\hat{\mathbf{h}}(0) = \mathbf{0}$, and recursively updates the values of $\hat{\mathbf{h}}$. The gradient vector,

$$\nabla J(n) = \begin{bmatrix} \frac{\partial J(n)}{\partial \hat{h}(0,n)} & \frac{\partial J(n)}{\partial \hat{h}(1,n)} & \cdots & \frac{\partial J(n)}{\partial \hat{h}(L-1,n)} \end{bmatrix}^T, \tag{16}$$

describes how the cost function $J(n)$ is affected by small changes of the estimate $\hat{\mathbf{h}}$. A straightforward algorithm for estimating $\hat{\mathbf{h}}$ is to use the steepest-descent method [24],

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu \nabla J(n), \tag{17}$$

where μ is the update step-size parameter. In order to derive the gradient vector $\nabla J(n)$ we need to know the first and second order statistic information of the signals $x(n)$ and $y(n)$.

In the least mean square (LMS) algorithm the gradient vector $\nabla J(n)$ is estimated as,

$$\nabla J(n) = -\mathbf{x}(n)e(n), \tag{18}$$

where

$$\mathbf{x}(n) = \begin{bmatrix} x(n) & \cdots & x(n-L+1) \end{bmatrix}^T, \tag{19}$$

and $x(n)$ is the transmission room signal.

Since no explicit averaging is done in (18), each recursive update of the LMS algorithm will suffer from a gradient noise. That is, the filter coefficients in $\hat{\mathbf{h}}(n)$ will only on average be updated correctly, making the convergence of the LMS algorithm slower than the steepest descent algorithm.

In the LMS algorithm, the gradient vector estimate, $\nabla J(n)$, is directly proportional to the transmission room signal $x(n)$. Therefore the convergence rate is also proportional to the transmission room signal. This can be corrected by normalizing the gradient estimate vector with a factor equal to the energy in $\mathbf{x}(n)$, namely the scalar $\mathbf{x}^T(n)\mathbf{x}(n)$. In order to avoid the risk of division by zero, or even a value close to zeros, we regularize the algorithm with a small scalar, ϵ_{reg}. The filter update in the normalized least mean square (NLMS) algorithm can then be written as,

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + \frac{\mu}{\mathbf{x}^T(n)\mathbf{x}(n) + \epsilon_{\text{reg}}}\mathbf{x}(n)e(n). \tag{20}$$

The NLMS algorithm exhibits slow convergence in frequency regions (modes) with little excitation signal energy [25]. For example, let us study the situation where the excitation signal $x(n)$ has lower energy level in the higher frequencies than in the lower, like e.g. a speech signal. Then the gradient vector estimate $\mathbf{x}(n)e(n)$ will be small for those modes that correspond to higher frequencies. The normalization factor, the scalar $\mathbf{x}^T(n)\mathbf{x}(n)$, will be large due to the large energy in the lower frequencies of $x(n)$. Therefore the recursive updates of the modes of $\hat{\mathbf{h}}(n)$, that correspond to small excitation signal energies, will be slow. Other algorithms, such as the recursive least-squares algorithm (RLS) [24], [26], and affine projection algorithms [27], incorporates knowledge of the spectrum of the excitation signal $x(n)$ into the filter update. These algorithms will therefore converge faster for non-white autocorrelated excitation signals.

Adaptive filters may become unstable, i.e., instead of converging to the correct solution they diverge. Usually the filter-tap estimates will then grow very rapidly. When this happens in an echo canceler, it will not only stop removing the annoying echo, but will also create a very strong distortion. It is therefore of greatest importance that the adaptive filter does not become unstable. It has been mathematically proven that the NLMS algorithm is stable for step-size values in the range $0 < \mu < 2$ [24]. The RLS algorithm, on the other hand, is notorious for being unstable (in finite precision), and without extra stabilization features, it is useless as an echo canceler. However, in Paper I and II, stabilization methods for the RLS algorithm are discussed.

In a stereophonic acoustic echo canceler, the adaptive filter is the main contributer to the system's calculation complexity. This is mainly due to the long adaptive filters needed. As an example, the NLMS and the fast RLS algorithms used in Paper I need $8L$ and $32L$ multiplications respectively per input sample, where L is the length of the adaptive filter. And in a typical acoustic echo canceler, L may be in the range of one thousand to several thousands. Methods to reduce the calculation complexity are therefore needed.

3.1 Adaptive Filtering in Subbands

One way to decrease the calculation complexity of the adaptive filters, is to introduce a filterbank structure. Here, the fullband signals $x(n)$ and $y(n)$ are decomposed by an analysis filterbank into several signals with lower band-width, $x_m(n)$ and $y_m(n)$ in Fig. 8. Then one adaptive filter in each subband estimates the subband residual echo signal, $e_m(n)$, before a synthesis filterbank reconstructs the fullband residual echo signal, $e(n)$ [28], [29]. Filterbanks are used in several signal processing applications and general descriptions can be found in numerous books and articles, e.g., [30], [31], [32], [33], [34].

If the filterbanks are not properly designed, aliasing can be a major source of

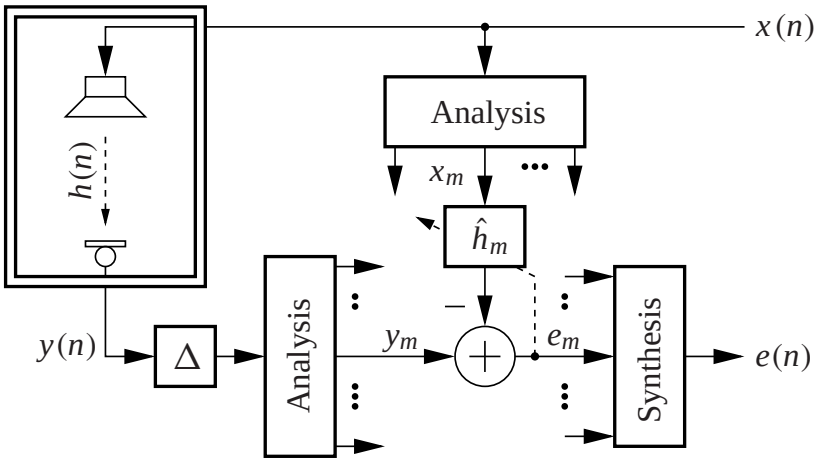

Figure 8: Standard subband echo canceler.

distortion. In one class of filterbanks, alias distortion is allowed in the subbands signal, and this distortion is then canceled by the synthesis filterbank when the fullband signal is reconstructed. While this is acceptable in many applications, e.g. in audio coding, alias distortion in the subbands will significantly decrease the performance of the adaptive filters [35]. In order to avoid aliasing, the filterbanks used in the echo canceler are usually non-critically downsampled. That is, the downsampling factor is less than the number of subbands. In Paper I, the real-time system has $M = 64$ subbands and the downsampling factor $r = 48$.

The calculation complexity reduction is a result of downsampling. Even if we need one adaptive filter for each subband, or at least half of them, each adaptive filter can be r times shorter, and they only need to be updated for each r fullband input sample. Efficient structures for filterbanks are available, and the calculation complexity overhead created by the filterbank is far less than the complexity gain in the adaptive filters. That is, from a calculation complexity viewpoint, we wish to have as high a downsampling factor as possible.

In the filterbank, bandpass filters are needed to suppress signals outside the frequency interval of a given subband. These filters will not only add calculation complexity to the system, but worse, they will also delay the signal. A higher downsampling factor requires longer filters in order to suppress aliasing. Since the signal delay is increased with filter length, the downsampling factor is partly a compromise between calculation complexity and signal transmission delay. The design of the filters are of course crucial, and in Paper I, a design method based on quadratic programming is proposed. An older design method was described in [34].

An alternative subband structure, which does not introduce extra delay to the signal path has been introduced in [36]. Like the previously described structure, the

adaptive filters estimate the impulse responses in the subbands. Instead of performing the compensation in the subbands, the fullband impulse response estimate is reconstructed from the estimated subband impulse responses. Using this reconstructed fullband impulse response, echo cancellation can be performed in the time domain without introducing any extra signal path delay to the receiving room signal. The price payed for zero signal path delay is increased calculation complexity, though it is still lower than in the fullband LMS situation [36]. In [37], a delay-less subband structure which is a mixture between this structure and the structure in Fig. 8, is presented.

One peculiar effect with adaptive filtering in subband structures is that a causal fullband impulse response may be best modeled with non-causal subband impulse responses. Remember, a casual impulse response means that the output signal from a system is only dependent on present and past input signals. The output signal of a non-causal system is also dependent on *future* input signals. Non-causality in subband structures was described in [28], and is also the topic of Paper III. Non-causality in delayless subband structures was studied in [38].

3.2 Frequency-Domain Adaptive Filtering

Another way to reduce the calculation complexity of adaptive filters is to perform the filtering in the frequency domain, Fig. 9. Here, the discrete Fourier transform is used to transform blocks of the input signal, $x(n)$, to the frequency domain. Then the room transfer function, $\hat{H}(k)$, and the estimated echo signal, $\hat{Y}(k)$, are derived in the frequency domain, before the estimated time-domain echo signal $\hat{y}(n)$ is calculated with an inverse discrete Fourier transform [39].

The calculation complexity reduction comes from the fact that convolution in the time-domain is equal to multiplication in the frequency domain, including the availability of efficient discrete Fourier transform structures, usually denoted the fast Fourier transform (FFT) [40], [41]. An FFT of size L can be computed with $\frac{L}{2}\log_2 L$ multiplications [42]. If we are to calculate L estimated echo samples using the time-domain convolution in (3), L^2 multiplications are needed. In the frequency domain, the convolution is replaced with multiplication, i.e. $\hat{Y}_m(k) = \hat{H}_m(k)X_m(k)$. Since multiplication in the discrete Fourier domain actually corresponds to circular convolution in the time domain [42], L zeros must be appended in each FFT in order to avoid time-domain aliasing. Therefore, in total we need $2L$ multiplications for the actual filtering, plus $2L\log_2 2L$ multiplications for the FFT and the inverse FFT. That is, the larger the block-size L, the bigger calculation complexity reduction compared to time-domain convolution. Since the algorithm is block-based, it will impose a delay to the transmission signal, and as for the subband echo canceler, this is the biggest disadvantage with a frequency-domain-based echo canceler.

It was earlier described that some fullband adaptive filters have slower conver-

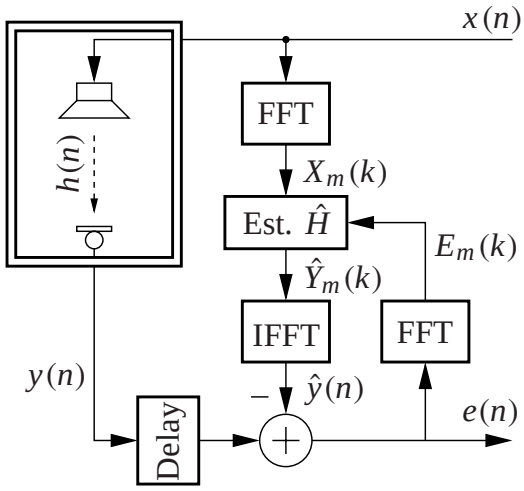

Figure 9: Frequency-domain-based echo canceler.

gence rate for signals with larger energy in some frequency regions than others. For the NLMS algorithm, it is due to the fact that the same normalization factor was used for all modes. In a frequency-domain echo canceler, we may have an individual normalization factor in each frequency region [43], and therefore, the convergence rate in frequency regions where the excitation signal has less energy can be significantly improved over a fullband NLMS algorithm.

In [44], [45] and [46] a delayless frequency-domain adaptive filter algorithm is presented. Like for the delayless subband based echo canceler, the price payed for zero signal path delay is increased calculation complexity.

3.3 Two-channel Adaptive Filtering

A direct generalization of the single channel NLMS (20) to a two channel version that is suitable for the structure depicted in Fig. 5 is presented below. In a situation where the two excitation signals $\mathbf{x}_1$ and $\mathbf{x}_2$ are statistically independent, this version is actually equivalent to a single channel NLMS filter with the double adaptive filter length.

$$e(n) = y(n) - \mathbf{x}_1^T \hat{\mathbf{h}}_1(n) - \mathbf{x}_2^T \hat{\mathbf{h}}_2(n), \tag{21}$$

$$\hat{\mathbf{h}}_i(n+1) = \hat{\mathbf{h}}_i(n) + \frac{\mu}{\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_2^T \mathbf{x}_2} \mathbf{x}_i e(n), \quad i = 1, 2. \tag{22}$$

In Paper I a two channel fast RLS algorithm is used. This algorithm was introduced in [47] and, in contrast to the two channel NLMS algorithm, it compensates for both

correlation within each channel and cross-correlation between the two channels.

In Paper II, a comparison is conducted between the two algorithms above, and a two-channel frequency-domain algorithm. The latter was introduced in [48] and further described in [49]. This algorithm also compensates both for correlation within each channel and cross-correlation between the two channels.

In [50], a multi-channel affine projection algorithm is introduced and in [51] a stereophonic acoustic echo canceler based on the affine projections algorithm is evaluated.

4 Audio Coding

In the digital world, the originally continuous audio signals are represented with quantized samples. As the information in these samples is very redundant, audio compression algorithms are commonly used to reduce the redundancy. One class of audio compression algorithms is denoted perceptual audio coders, and several different perceptual audio coders have been presented, and published [52], [53], [54], [55], [56], [57] and [58]. In this thesis we are interested in two possible combination gains when both an echo canceler and a perceptual audio coder are used. First, the performance of the stereophonic echo canceler can be increased if the audio coder reduces the correlation between the two excitation signals, which can actually be the effect of using a perceptual audio coder. Secondly, both the audio coder and the echo canceler may use filterbanks or transforms, and by jointly designing these, we may be able to reduce both calculation complexity and signal transmission delay.

In an audio coder, a model of the human ear is used for determining what is audible and what is not. Strong audio components will namely mask weaker components that are close in time and frequency [59]. Therefore, the audio coder starts by estimating the global masking threshold, a function of time and frequency which describes the minimum energy needed for a specific tone to be audible, see Fig. 10. Having this knowledge, we need only to store/transmit tones that are stronger than this masking threshold. The compression algorithms in audio coders are usually not loss-less, i.e. it is not possible to reconstruct the original signal exactly, only a signal that hopefully sounds like the original signal.

To be a little bit more specific. In order to reduce the amount of data in an audio signal we need to increase the quantization interval. When a signal is quantized, it can be modeled as if quantization noise is added to the signal. In an audio coder, short segments of the signal are transferred to the frequency domain, and the frequency bins are then quantized with the restriction that the level of the quantization noise should be below the global masking threshold.

In this thesis two possible gains which result from combining an echo canceler and

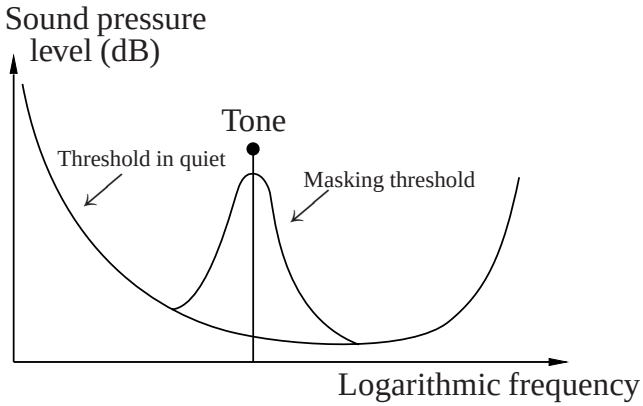

Figure 10: Global masking threshold. Acoustic events below the two curves will not be audible.

an audio coder are discussed. The first is about the fundamental problem of stereophonic acoustic echo cancellation. In Section 2.1 it was shown that the fundamental problem is due to the fact that the two excitation signals $x_1(n)$ and $x_2(n)$ can be strongly correlated. If a perceptual audio coder is used, it may add independent quantization noise to the two channels, and thereby decreasing the correlation between the two channels. This is analyzed in Paper IV for both an MPEG 1 layer III [55] and an MTPC [60] audio coder.

The second possible gain is to use a joint filterbank or transforms for the audio coder and the echo canceler. As is described above, the audio coder needs to decompose the signal in order to increase the frequency resolution. This is usually done either with a filterbank [32], [55], [61], or with a modified discrete cosine transform (MDCT) [62], [63]. In Paper V, we analyze what gains are possible by combining a subband based perceptual audio coder with a subband echo canceler and a MDCT based coder with a frequency-domain echo canceler.

5 Outline of the Thesis

The thesis consists of five papers, out of which two focus on stereophonic acoustic echo cancellation, the third analyses impulse responses in a subband based echo canceler, the fourth describes how an audio coder may enhance the performance of a stereophonic acoustic echo canceler and the fifth finally shows how to combine an echo canceler and an audio coder. Although the papers are written independently of each other and are self-contained, we have tried to be as consistent as is possible with the notation.

Paper I - A Real-time Implementation of a Stereophonic Acoustic Echo Canceler

The aim of this paper is to propose a structure for a stereophonic acoustic echo canceler suitable for a real-time implementation. The paper starts by describing the fundamental problems of stereophonic acoustic echo cancellation, and discusses some possible methods for solving these problem.

In earlier work [47], the two-channel fast recursive least-squares (FRLS) algorithm has been derived, and this algorithm has shown fast convergence properties for the ill-conditioned stereophonic echo cancellation problem. Therefore, this algorithm was chosen for the implementation, and the algorithm is also presented in the paper. The two-channel FRLS algorithm has two major disadvantages, instability problems and high calculation complexity. In the paper, a two-path adaptive filter is proposed to be used in order to mitigate instability. With this filter, we will not always use the latest FRLS adaptive filter impulse response estimate, but rather the *best* impulse response estimate. That is, in a situation where the FRLS adaptive filter is becoming unstable, an earlier impulse response estimate is used until the FRLS adaptive filter has reconverged.

The high calculation complexity of the FRLS adaptive filter is reduced by applying the filters in a subband structure. Here the signals are decomposed, with a filterbank, into several subband signals, each one approximately corresponding to a given frequency region. Then one adaptive filter is applied in each subband. In the paper, an efficient structure for filterbanks with non-critical downsampling is described. Also described is a filter design method for the prototype filter needed in the filterbank. The design method is written as a minimization problem, and it is shown how quadratic programming can be used to solve the minimization problem. Prototype filters designed with this method generally have fewer coefficients than filters designed with other methods, e.g. [34]. An intuitive table, exemplifying the calculation complexity reduction achieved with the filterbank structure, is also given.

In a simulation section, we once again return to the fundamental problem of stereophonic acoustic echo cancellation. Simulations performed on real-life recorded data show how the channel correlation can be reduced with the earlier discussed methods. It is also shown that the fundamental problem is not only a theoretical problem, but also a severe problem in e.g. a quiet office environment. Other examples study the convergence rate of the adaptive filters in individual subbands, and it is discovered that the two-channel FRLS is significantly better than the NLMS algorithm, mainly in the frequency region 0-4 kHz. Therefore, it is proposed that the NLMS algorithm is used in the upper frequency region, thereby reducing the instability problems and the high calculation complexity inherent with FRLS algorithm.

A real-time implementation of the system has been constructed, and the implementation has been used for perceptual evaluations. This implementation is also discussed

in the paper.

The paper has been co-authored with Steven Gay, Tomas Gänsler and Jacob Benesty, and it has been accepted for publication [64]. The paper in the thesis has been extended with Appendix D and E, where parts of Appendix E has been presented in [65]. Parts of the results presented in the paper is also published in [21], and presented at [66], [67].

Paper II - Comparison of Different Adaptive Algorithms for Stereophonic Acoustic Echo Cancellation

This paper compares different two-channel adaptive algorithms for stereophonic acoustic echo cancellation. The comparison includes convergence rate, tracking, calculation complexity, transmission signal delay and memory usage.

Since the first echo canceler was presented in the 1967 [3], the NLMS algorithm has, without competition, been the most commonly used adaptive algorithm. The performance of the NLMS algorithm is sufficient in most monophonic echo cancelers. However, it has been shown that NLMS has too slow convergence rate in stereophonic echo cancelers. Nevertheless, the NLMS algorithm is often used as the base algorithm, to which other algorithms are compared. This is also done in this paper.

In recent years, several two-channel and multi-channel versions of adaptive algorithms have been derived [47], [50], [48]. In addition to the two-channel version of the standard NLMS, the paper evaluates two other algorithms, namely the two-channel fast recursive least-squares (FRLS) and a two-channel frequency-domain algorithm. The two-channel FRLS algorithms is used in Paper I. We know from this work that the FRLS algorithm has a very fast convergence rate, but severe disadvantages as instability problems and high calculation complexity. In this paper, stabilization methods are discussed, including a method that in most cases detects in advance if the FRLS algorithm is about to become instable. The paper also includes a section on how filterbanks can be used in order to reduce the calculation complexity.

The two-channel frequency-domain algorithm is a block-based algorithm, i.e. blocks of the input-signals are transformed by a discrete Fourier transform from the time domain to the frequency domain. The adaptive filtering is then performed in the frequency domain. Therefore, we can have individual adaptive step-size parameters in each frequency bin, thereby increasing the convergence rate for frequency regions with low signal energy in comparison with the standard NLMS algorithm. Moreover, in the two-channel frequency-domain algorithm, the cross-correlation between the channels is also taken into account. In the paper, it is also discussed how the convergence rate can be increased by overlapping input data.

To have an exhaustive example, including all possible values on all parameters is of course outside the scope of this paper. Instead, examples on a practical usable

parameter set is given. For this parameter set, the calculation complexity, the signal transmission delay and the memory usage is given for both fullband and subband versions of the NLMS and the two-channel FRLS algorithm, and for a two-channel frequency-domain algorithm. Examples of convergence rate and tracking on real-life recorded speech signals are given for all systems except the fullband two-channel FRLS.

The paper has been co-authored with Jacob Benesty, Tomas Gänsler and Steven Gay, and has been published in [68].

Paper III - Analysis of Subband Impulse Responses in Subband Echo Cancelers

In a subband echo canceler, the signals are decomposed into several subband signals, with one adaptive filter in each subband. It has previously been shown that a causal fullband impulse response is best modeled by non-causal subband impulse responses [28], [29]. In this paper, we analyze the rationale of this non-causality.

For adaptive filters in subband schemes we usually make a distinction between an open-loop and a closed loop structure. In the open-loop structure the residual signals in each subband are minimized individually, whereas in the closed-loop structure, the fullband residual signal is minimized. The open-loop structure is the most commonly used in subband echo cancelers, and it is also the impulse responses of this structure that is analyzed in this paper. Since each subband is minimized individually, without interaction from signals in other subbands, it is also possible to analyze them separately. The paper begins with describing the model in order to analyze one subband and to model non-causal filter taps with a causal adaptive filter, we introduce a delay Δ to the transmission room signal.

In the paper, several different situations are analyzed. It starts out with a completely ideal situation, where it is assumed that both the filters used in the subband filterbank and the adaptive filter are allowed to be non-causal and of infinite length. The filterbank is also ideal in the sense that it has an perfect frequency response, with the amplification of either 1 or 0. It is shown that in this situation the non-causal taps can be viewed as a result of interpolation.

Next an expression for the transfer function of the subband impulse response when the filterbank is non-ideal, i.e. a realizable filterbank is used. The adaptive filter can still be non-causal and of infinite length. This is denoted the Wiener solution. Even if the uncertainty is too large for a general solution of the transfer function, a couple of different situations are studied. It is concluded that the transfer function can have values that significantly differ from the expected value close to the band edges. These transfer functions are best modeled with non-causal filters.

Finally, the minimum mean square error solution is derived for a fully realizable system. It is shown that the transfer function can be significantly different from the Wiener solution in frequency regions that are close to the band edges of the subbands. Actually, since the signal energy is close to zero, the transfer function may have a near arbitrary shape close to the band edges. With formulas given in the paper, it is also possible to calculate the optimal number of non-causal taps, Δ, for a given filterbank and a given fullband impulse response.

It is a well known fact that the NLMS adaptive algorithm has a slow convergence rate for signals with a large eigenvalue spread [24]. In a subband system, this will be the case, since the signal energy close to the band edges are close to zero. What implications this has on the number of non-causal taps is also analyzed.

In the paper, simulation examples are given for the ideal situation, and also minimum mean square error solution, showing the need for non-causal filter taps. Other simulations show the behavior of systems with an NLMS or an RLS adaptive filter. The number of non-causal taps needed as a function of number of subbands for a typical system configuration is also plotted in a figure.

The paper has been co-authored with Tomas Gänsler, and has been submitted for publication [69]. Parts of the results are also published in [70].

Paper IV - Influence of Audio Coding on Stereophonic Acoustic Echo Cancellation

This paper presents a method to reduce the correlation between the right and the left channel in a stereophonic hands-free communication system. Correlation reduction is needed since the non-uniqueness problem in stereophonic acoustic echo cancellation is exactly due to the channel correlation. The method in this paper uses a perceptual audio coder in order to decrease this correlation.

The paper starts with a description of the fundamental problem in stereophonic acoustic echo cancellation and why channel correlation reduction is the only solution to this problem. It is also discussed how the magnitude coherence function is a good measurement of the channel correlation and thereby of how efficiently a decorrelator reduces channel correlation. This is also verified with simulations.

Even if the results in the paper apply to perceptual audio coders in general, the layer III audio coder in MPEG-1 and 2 is used both in the context of describing how a perceptual audio coder can be used to reduce the correlation and in the simulations. It is shown that one of the fundamental principals of the audio coder actually is to add none-perceivable quantization noise to the signals. This since the coder estimates the global masking threshold, i.e., a function of time and frequency that describes the energy needed for a tone to be audible for humans when it is masked by other tones. The source signal is then quantized as much under the restriction that the quantization

noise should have less energy than the global masking threshold. By adding quantization noise to the signals, we will also decorrelate the signals.

It is also found that the quantization level is not optimal, in the sense that a coder never adds a maximal amount of none-perceivable quantization noise. Because, if each component were to be individually maximally quantized, the overhead information needed to describe how each component is quantized would be too large. Therefore, several components will use the same quantization scheme. For some components, this will result in quantization noise far below the masking threshold. By adding noise to these components, it is possible to reduce the channel correlation even further.

It is also shown how an MPEG audio coder, and a modified MPEG audio coder reduces the magnitude coherence function between two channels of recorded real-life speech signals. The magnitude coherence function of the output signal of the standard and the modified MPEG audio coder are compared with an unprocessed signal, and a signal that has been processed with the non-linear function presented in [12] (also used in Paper I). The performance of a stereophonic acoustic echo canceler, given as the misalignment error of the impulse response estimate, is also shown for a system without a decorrelator, and for systems with the three decorrelation methods discussed above.

The paper has been co-authored with Tomas Gänsler and has been published in [16]. This work has been supported be Telia Research AB, and the method has been patented by Telia AB [71].

Paper V - Joint Filterbanks for Echo Cancellation and Audio Coding

This paper investigates the possibilities of combining the filterbanks or the transforms of an echo canceler and a perceptual audio coder. Possible gains are a reduction of the total signal transmission delay and reduced calculation complexity.

One class of perceptual audio coders uses a filterbank to decompose the signal into narrow-band subband signals. In a subband based echo canceler, a filterbank is also used in order to decompose the signal, and it would be advantageous to use one filterbank for the two systems. However, the two filterbanks used are fundamentally different. In the audio coder the filterbank usually use critical downsampling, i.e., the same number of subbands as downsampling factor. This results in downsampling aliasing. Aliasing is not a problem in the coder, and the aliasing is canceled when the fullband signal is reconstructed by decoder. On the other hand, the performance of the adaptive filter in an echo canceler will decrease on signals with significant amount of aliasing [35]. Therefore, typically a non-critically downsampled filterbank with high suppression of aliasing components is used in filterbank based echo cancelers. In the paper, two possible solutions are proposed. To start with, a modified audio coder

filterbank, with non-critical downsampling, that can be used for the echo canceler is presented. The modification is such that the output of a standard audio coder filterbank can be reconstructed with a trivial operation. Secondly, a method to convert subband signals directly between a critically downsampled audio coder filterbank and a non-critically downsampled echo canceler filterbank.

In some modern perceptual audio coders, such as the AAC coder [52], the frequency resolution is increased by interchanging the filterbank for a modified discrete cosine transform (MDCT). In the paper, it is shown that this transform can be written as a function of a discrete Fourier transform. Therefore, it can be of interest to combine a MDCT based audio coder with a frequency-domain adaptive filter (FDAF) based echo canceler. Methods how to combine these two systems are evaluated in the paper. Situations where joint transforms are unwanted are also discussed, and it is concluded that in these situations, the system designer should at least use a block based adaptive filter for the echo canceler and common buffers for the audio coder and the echo canceler, in order to reduce the signal transmission delay.

Also in the paper, the calculation complexity, the signal transmission delay and the adaptive filter convergence rate examples are given for the discussed system combination.

This paper is authored by Peter Eneroth and has been submitted to [72].

References

[1] J. W. Emling and D. Michell, "The effects of time delay and echo on telephone conversations," *The Bell Syst. Tech. J.*, vol. XLII, no. 6, pp. 2869–2891, Nov. 1963.

[2] P. T. Brady and G. K. Helder, "Echo suppressor design in telephone communications," *The Bell Syst. Tech. J.*, vol. XLII, no. 6, pp. 2893–2917, Nov. 1963.

[3] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. XLVI, no. 3, pp. 497–510, Mar. 1967.

[4] M. M. Sondhi, "Echo canceller," U.S. Patent 3,499,999, Mar. 10, 1970 (filed Oct. 31, 1966).

[5] J. L. Kelly and B. F. Logan, "Self-adjust echo suppressor," U.S. Patent 3,500,000, Mar. 10, 1970 (filed Oct. 31, 1966).

[6] A. Eriksson, G. Eriksson, J. Karlsen, A. Roxström, and T. V. Hulth, "Ericsson echo cancellers - a key to improve speech quality," *Ericsson Review*, no. 1, pp. 25–33, 1996.

[7] C. W. K. Gritton and D. W. Lin, "Echo cancellation algorithms," *IEEE ASSP Mag.*, pp. 30–37, Apr. 1984.

[8] J. Benesty, D. R. Morgan, J. Hall, and M. M. Sondhi, "Synthesized stereo combined with acoustic echo cancellation for desktop conferencing," *Bell Labs Tech. J.*, vol. 3, no. 3, pp. 148–158, July-Sept. 1998.

[9] M. M. Sondhi and D. R. Morgan, "Acoustic echo cancellation for stereophonic teleconferencing," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio Acoustics*, 1991.

[10] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation — An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[11] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1997, pp. 303–306.

[12] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[13] J. Benesty, T. Gänsler, and P. Eneroth, "Multi-channel sound, acoustic echo cancellation, and multi-channel time-domain adaptive filtering," in *Acoustic Signal Processing For Telecommunication*, S. L. Gay and J. Benesty, Eds., vol. 551, chapter 6, pp. 101–120. Kluwer Academic Publishers, 2000, ISBN 0-7923-7814-8.

[14] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. on Speech Audio Processing*, submitted.

[15] A. Gilloire and V. Turbin, "Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers," in *Proc. IEEE ICASSP*, 1998, pp. 3681–3684.

[16] T. Gänsler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1998, pp. 3649–3652.

[17] S. Shimauchi, Y. Haneda, S. Makino, and Y Kaneda, "New configuration for a stereo echo canceller with nonlinear pre-processing," in *Proc. IEEE ICASSP*, 1998, pp. 3685–3688.

[18] M. Ali, "Stereophonic echo cancellation system using time-varying all-pass filtering for signal decorrelation," in *Proc. IEEE ICASSP*, 1998, pp. 3689–3692.

[19] Y. Joncour and A. Sugiyama, "A stereo echo canceler with pre-processing for correct echo path identification," in *Proc. IEEE ICASSP*, 1998, pp. 3677–3680.

[20] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 468–475, Sept. 1998.

[21] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A real-time stereophonic acoustic echo canceler," in *Acoustic Signal Processing For Telecommunication*, S. L. Gay and J. Benesty, Eds., vol. 551, chapter 8, pp. 135–152. Kluwer Academic Publishers, 2000, ISBN 0-7923-7814-8.

[22] S. Shimauchi, S. Makino, Y. Haneda, A. Nakagawa, and S. Sakauchi, "A stereo echo canceller implemented using a stereo shaker and a duo-filter control system," in *Proc. IEEE ICASSP*, 1999, pp. 857–860.

[23] V. Fischer, T. Gänsler, E. J. Diethorn, and J. Benesty, "A software stereo acoustic echo canceler under microsoft windows," *Private Communication*, 2000.

[24] S. Haykin, *Adaptive Filter Theory*, Prentice Hall International, 1996.

[25] D. R. Morgan, "Slow asymtotic convergence of LMS acoustic echo cancelers," *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 2, pp. 126–136, Mar. 1995.

[26] M. G. Bellanger, *Adaptive Digital Filters and Signal Analysis*, Marcel Dekker, 1987.

[27] S. L. Gay, *Fast Projection Algorithms with Application to Voice Echo Cancellation*, Ph.D. thesis, State University of New Jersey, Oct. 1994.

[28] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. of ICASSP*, 1988, pp. 2570–2573.

[29] W. Kellermann, *Zur Nachbildung physikalischer Systeme durch parallelisierte digitale Erzatzsysteme im Hinblick auf die Kompensation akustischer Echos*, Ph.D. thesis, Darmstadt, 1989.

[30] N. J. Fliege, *Multirate Digital Signal Processing*, John Wiley & Sons, 1994.

[31] G. Strang and T. Nguyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1996.

[32] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall PTR, 1993.

[33] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall PTR, 1995.

[34] G. Wackersreuther, "On the design of filters for ideal QMF and polyphase filter banks," *AEÜ*, vol. 39, no. 2, pp. 123–130, 1985.

[35] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[36] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Trans. on Signal Processing*, vol. 43, no. 8, pp. 1819–1830, Aug. 1995.

[37] M. Dörbecker and P. Vary, "Reducing the delay of an acoustic echo canceller with subband adaptation," in *The Int. Workshop on Acoust. Echo and Noise Control*, 1995, pp. 103–106.

[38] P. Eneroth and T. Gänsler, "A modified open-loop delayless subband adaptive echo canceler," in *The Int. Workshop on Acoust. Echo and Noise Control*, 1997, pp. 156–159.

[39] D. Mansour and A. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 30, no. 5, pp. 726–734, Oct. 1982.

[40] P. Duhamel, "Implemetation of "Split-Radix" FFT algorithms for complex, real, and real-symmetric data," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 2, pp. 285–295, Apr. 1986.

[41] H. Sorensen, D. Jones, M. Heideman, and S. Burrus, "Real-values fast Fourier transform algorithms," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, pp. 849–863, June 1987.

[42] J. Proakis and D. Manolakis, *Digital Signal Processing*, Prentice Hall Inc., 1996.

[43] P. Sommen, P. Van Gerwen, H. Kotmans, and A. Janssen, "Convergence analysis of a frequency-domain adaptive filter with exponetial power averaging and generalized window function," *IEEE Trans. on Circuits and Systems*, vol. 34, no. 7, pp. 788–798, July 1987.

[44] T. Gänsler, P. Eneroth, and G. Salomonsson, "A frequency domain adaptive echo canceller with post-processing residual echo supression by decorrelation," in *Proc. of ICSPAT*, 1996, pp. 394–398.

[45] P. Eneroth and T. Gänsler, "A frequency domain adaptive echo canceller with post-processing residual echo supression by decorrelation," Signal Processing Report SPR–40, Dept. of Applied Electronics, LTH, 1997.

[46] T. Gänsler and G. Salomonsson, "Nonintrusive measurements of the telephone channel," *IEEE Trans. on Commun.*, vol. 47, no. 1, pp. 158–167, Jan 1999.

[47] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.

[48] J. Benesty, A. Gilloire, and Y. Grenier, "A frequency domain stereophonic acoustic echo canceler exploiting the coherence between the channels," *J. Acoust. Soc. Am.*, vol. 106, pp. L30–L35, Sept. 1999.

[49] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 789–792.

[50] J. Benesty, P. Duhamel, and Y. Grenier, "A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation," *IEEE Signal Processing Lett.*, vol. 3, no. 2, pp. 35–37, Febr. 1996.

[51] S. Makino et. al., "Subband stereo echo canceller using the projection algorithm with fast convergence to the true echo path," in *Proc. of ICASSP*, 1997, pp. 299–302.

[52] M. Bosi et. al., "ISO/ICE MPEG-2 advanced audio coding," *J. Audio Eng. Soc. (AES)*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[53] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, vol. 14, no. 5, pp. 59–81, Sept. 1997.

[54] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to* MPEG-*2*, chapter 4, pp. 55–79, Digital Multimedia Standards Series. Chapman & Hall, 1997.

[55] ISO/IEC 11172–3, "Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio," ISO/IEC JTC 1/SC 29, Case Postale 56, CH1211 Genève 20, Switzerland, 1993.

[56] Steven Vernon, "Design and implementation of AC–3 coders," *IEEE Trans. on Consumer Electronics*, vol. 41, no. 3, Aug. 1995.

[57] Fraunhofer IIS, "MPEG-1 LAYER III shareware audio coder," 1995, Am Weichselgarten 3 D-91058 Erlangen Germany, encoder and decoder code: http://www.iis.fhg.de/amm/techinf/layer3/index.html, Public domain decoder source code (ANSI c): ftp://ftp.fhg.de/pub/iis/layer3/public_c/.

[58] M. Iwadare et. al., "A 128 kb/s Hi-Fi audio CODEC based on adaptive transform coding with adaptive block size MDCT," *IEEE J. on Selected areas in Communications*, vol. 10, no. 1, pp. 138–144, Jan. 1992.

[59] P. Noll, "Wideband speech and audio coding," *IEEE Communication Mag.*, pp. 34–44, Nov. 1993.

[60] S. A. Ramprashad, "A multimode transform predictive coder (MTPC) for speech and audio," in *IEEE Speech Coding Workshop*, June 1999.

[61] T. Saramäki and J. Yli-Kaakinen, "Design of digital filters and filter banks by optimization: Applications," in *Proc. of EUSIPCO*, Sept. 2000.

[62] J. Princen and A. Bradley, "Analysis/synthesis filterbank designed based on time domain aliasing cancellation," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 34, no. 5, pp. 1153–1161, Oct. 1986.

[63] J. Princen, A. Johnson, and A. Bradley, "Suband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE ICASSP*, Apr. 1987, pp. 50.1.1–50.1.4.

[64] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A real-time implementation of a stereophonic acoustic echo canceler," *IEEE Trans. Speech Audio Processing*, accepted for publication.

[65] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "An implementation of a stereophonic acoustic echo canceler on a general purpose DSP," in *Proc. ICSPAT*, 1999.

[66] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A hybrid FRLS/NLMS stereo acoustic echo canceler," in *The Int. Workshop on Acoust. Echo and Noise Control*, 1999, pp. 20–23.

[67] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "Studies of a wideband stereophonic acoustic echo canceler," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio Acoustics*, 1999, pp. 207–210.

[68] P. Eneroth, J. Benesty, T. Gänsler, and S. L. Gay, "Comparison of different adaptive algorithms for stereophonic acoustic echo cancellation," in *Proc. of EUSIPCO*, 2000.

[69] P. Eneroth and T. Gänsler, "Analysis of subband impulse responses in subband echo cancelers," *IEEE Trans. Signal Processing*, submitted.

[70] P. Eneroth and T. Gänsler, "Analysis of subband impulse responses in subband echo cancelers," Signal Processing Report SPR–46, Dept. of Applied Electronics, LTH, 1999.

[71] P. Eneroth, T. Gänsler, and O. Till, "Metod och anordning vid stereoakustisk ekosläckning," Swedish Patent 512,903, June 5, 2000 (filed Oct. 29, 1997).

[72] P. Eneroth, "Joint filterbanks for echo cancellation and audio coding," *IEEE Trans. Speech Audio Processing*, submitted.

Paper I

Paper I

A Real-time Implementation of a Stereophonic Acoustic Echo Canceler

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 (FRLS) 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 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.

Based on: P. Eneroth and S. Gay and T. Gänsler and J. Benesty "A Real-time Implementation of a Stereophonic Acoustic Echo Canceler," *IEEE Trans. Speech and Audio Processing*, accepted for publication.

This work was partly done at Bell Labs, Lucent Technologies.

1 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 on in [3], [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 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], [8]. 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 improve the SAECR's ability to find the true solution, the normal equations to be solved are still ill conditioned. The standard

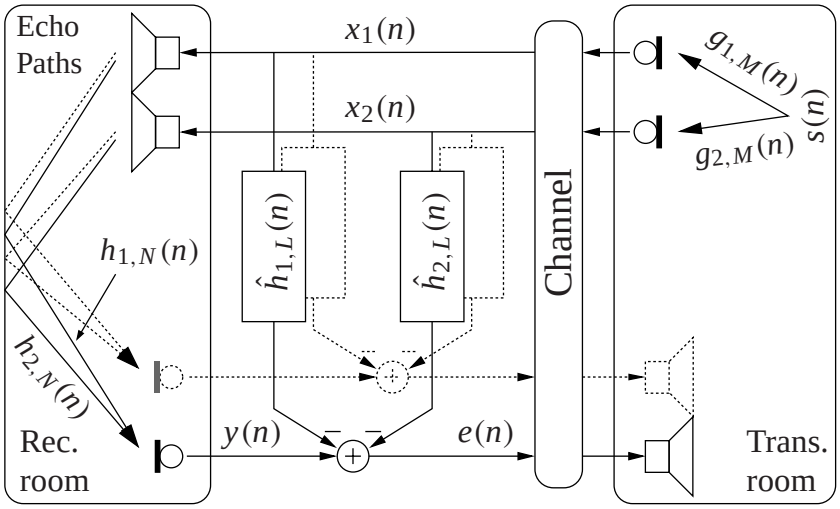

Figure 1: Schematic diagram of stereophonic acoustic echo cancellation.

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.

2 Problems and Solutions of Stereophonic 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 below.

2.1 Problem Formulation

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

$$x_i(n) = g_i(n) * s(n),\ i = 1, 2, \tag{1}$$

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

$$e(n) = y(n) - \hat{\mathbf{h}}_{1,L}^H(n)\mathbf{x}_{1,L}(n) - \hat{\mathbf{h}}_{2,L}^H(n)\mathbf{x}_{2,L}(n), \tag{2}$$

$$y(n) = \mathbf{h}_{1,N}^H\mathbf{x}_{1,N}(n) + \mathbf{h}_{2,N}^H\mathbf{x}_{2,N}(n), \tag{3}$$

$$\mathbf{h}_{i,N} = \begin{bmatrix} h_{i,0} & \dots & h_{i,N-1} \end{bmatrix}^T, \tag{4}$$

$$\mathbf{x}_{i,N}(n) = \begin{bmatrix} x_i(n) & \dots & x_i(n-N+1) \end{bmatrix}^T. \tag{5}$$

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

$$J(n) = \sum_{l=1}^{n} \lambda^{n-l}|e(l)|^2, \qquad 0 < \lambda \leq 1, \tag{6}$$

results in solving the system of linear equations [9]

$$\mathbf{R}_{xx}(n)\begin{bmatrix} \hat{\mathbf{h}}_{1,L}(n) \\ \hat{\mathbf{h}}_{2,L}(n) \end{bmatrix} = \mathbf{r}_{yx}(n), \tag{7}$$

where $\mathbf{r}_{yx}(n)$ is the estimated cross-correlation vector and $\mathbf{R}_{xx}(n)$ is the estimated correlation matrix,

$$\mathbf{R}_{xx}(n) = \sum_{l=1}^{n} \lambda^{n-l}\begin{bmatrix} \mathbf{x}_{1,L}(l)\mathbf{x}_{1,L}^H(l) & \mathbf{x}_{1,L}(l)\mathbf{x}_{2,L}^H(l) \\ \mathbf{x}_{2,L}(l)\mathbf{x}_{1,L}^H(l) & \mathbf{x}_{2,L}(l)\mathbf{x}_{2,L}^H(l) \end{bmatrix}. \tag{8}$$

The challenging problem with stereophonic acoustic echo cancellation lies in the condition number of this matrix. If we define the misalignment as

$$\varepsilon(n) = ||\mathbf{h} - \hat{\mathbf{h}}(n)||^2/||\mathbf{h}||^2, \tag{9}$$

where

$$\hat{\mathbf{h}}(n) = [\hat{\mathbf{h}}_{1,L}^T(n)\ \hat{\mathbf{h}}_{2,L}^T(n)]^T, \tag{10}$$

$$\mathbf{h} = [\mathbf{h}_{1,N}^T\ \mathbf{h}_{2,N}^T]^T, \tag{11}$$

it can be shown that [4],

$$\begin{aligned} L \geq M &\Rightarrow \mathbf{R}_{xx}(n) \text{ is singular } \forall n, \\ L < M &\Rightarrow \mathbf{R}_{xx}(n) \text{ is ill-conditioned}, \\ L \geq N &\Rightarrow \varepsilon(n) = 0,\ n \geq N, \\ L < N &\Rightarrow \varepsilon(n) \neq 0,\ \forall n, \end{aligned} \tag{12}$$

where the two latter statements in (12) require $\mathbf{R}_{xx}(n)$ not to be singular. Equation (12) is valid in the situation where no noise is added to the microphone signal $y(n)$, see Fig. 1.

As shown in [4] and stated in (12), 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 $L < M$ and $L < N$, 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.

2.2 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 non-perceptible 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 non-linearly 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,

$$x_1'(n) = x_1(n) + \alpha \frac{x_1(n) + |x_1(n)|}{2}, \tag{13}$$

$$x_2'(n) = x_2(n) + \alpha \frac{x_2(n) - |x_2(n)|}{2}, \tag{14}$$

where α determines the amount of added distortion. It has been found that α 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 non-perceivable 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.

3 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

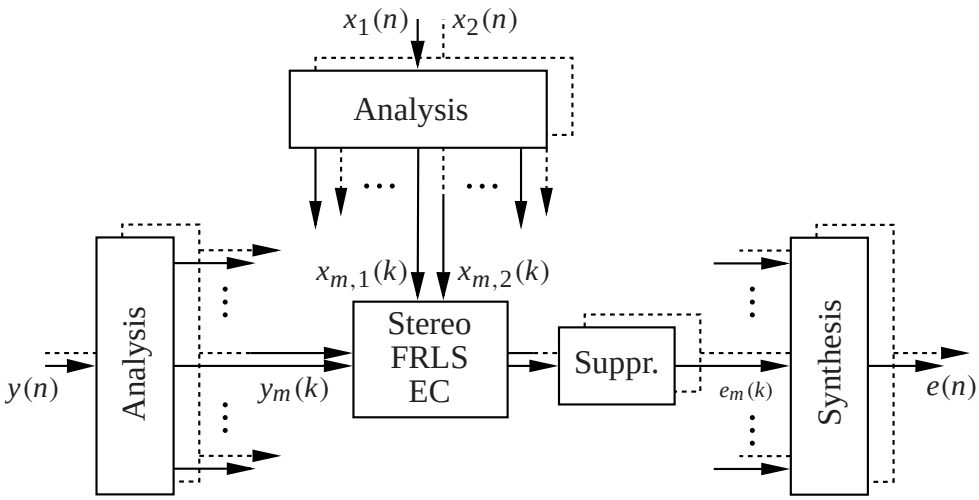

Figure 2: Stereophonic subband acoustic echo canceler. In total four analysis filterbanks decompose the transmission room and receiving room signals. In each subband, the two-path two-channel FRLS algorithm cancels the major part of the echo. The residual echo is reduced even further by the suppressor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

high calculation complexity and stability problems with non-stationary 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.

3.1 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], [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 have to be defined:

$$\boldsymbol{\chi}(n) = \begin{bmatrix} x_1(n) & x_2(n) \end{bmatrix}^T, \qquad (2 \times 1) \qquad (15)$$

$$\underline{\mathbf{x}}(n) = \begin{bmatrix} \boldsymbol{\chi}^T(n) & \boldsymbol{\chi}^T(n-1) & \dots & \boldsymbol{\chi}^T(n-L+1) \end{bmatrix}^T, \qquad (2L \times 1) \qquad (16)$$

$$\underline{\hat{\mathbf{h}}}(n) = \begin{bmatrix} \hat{h}_{1,1}(n) & \hat{h}_{2,1}(n) & \dots & \hat{h}_{1,L}(n) & \hat{h}_{2,L}(n) \end{bmatrix}^T, \qquad (2L \times 1) \qquad (17)$$

where L denotes the length of the adaptive filters and $h_{1,1}(n)$ denotes coefficient number 1 of channel 1 and $h_{2,1}(n)$ coefficient number 1 of channel 2 etc. Note that the channels of the filter and state vector $\underline{\mathbf{x}}(n)$ are interleaved in this algorithm. The complete two-channel FRLS adaptive filter is given in Table 1, where the following quantities are used:

$$\begin{aligned}
\mathbf{A}(n),\ \mathbf{B}(n) &= \text{Forward and backward prediction coefficient matrices,}\\
\mathbf{E}_{\mathrm{A}}(n),\ \mathbf{E}_{\mathrm{B}}(n) &= \text{Forward and backward prediction error energy matrices,}\\
\mathbf{e}_{\mathrm{A}}(n),\ \mathbf{e}_{\mathrm{B}}(n) &= \text{Forward and backward prediction error vectors,}\\
\mathbf{G}(n) &= \text{Kalman gain vector,}\\
\varphi(n) &= \text{Inverse conversion factor,}\\
k \in [1.5, 2.5],&\ \text{Stabilization parameter,}\\
\lambda \in (0, 1],&\ \text{Forgetting factor.}
\end{aligned}$$

In this version, stability is improved with the stability parameter k. But for operation on non-stationary signals, like speech-signals, further enhancements are needed. First of all, by monitoring φ, 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, $\underline{\hat{\mathbf{h}}}$, can be left unchanged. A suitable initial value for $\mathbf{A}(n)$, $\mathbf{B}(n)$ and $\mathbf{G}(n)$ is $\mathbf{0}$ whereas the energy estimates, $\mathbf{E}_{\mathrm{A}}(n)$ and $\mathbf{E}_{\mathrm{B}}(n)$ 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.

3.2 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 $\hat{h}_{\mathrm{RLS}}$. 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, $E_{e_{RLS,i}}$, is less than the residual echo energy in

Table 1: The stereophonic FRLS algorithm for complex arithmetic. The conjugate and the Hermitian transposition operator are denoted * and ^{H}, respectively. Variables are defined in Sec. 3.1. The prediction part is complete, whereas the filtering part only describes one return path.

Prediction	**Matrix sizes**
$\mathbf{e}_{\mathrm{A}}(n) = \boldsymbol{\chi}(n) - \mathbf{A}^H(n-1)\underline{\mathbf{x}}(n-1),$	(2×1)
$\varphi_1(n) = \varphi(n-1) + \mathbf{e}_{\mathrm{A}}^H(n)\mathbf{E}_{\mathrm{A}}^{-1}(n-1)\mathbf{e}_{\mathrm{A}}(n),$	(1×1)
$\begin{bmatrix}\mathbf{M}(n)\\ \mathbf{m}(n)\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{G}(n-1)\end{bmatrix} + \begin{bmatrix}\mathbf{I}_2\\ -\mathbf{A}(n-1)\end{bmatrix}\mathbf{E}_{\mathrm{A}}^{-1}(n-1)\mathbf{e}_{\mathrm{A}}(n),$	$((2L+2) \times 1)$
$\mathbf{e}_{\mathrm{B}_2}(n) = \boldsymbol{\chi}(n-L) - \mathbf{B}^H(n-1)\underline{\mathbf{x}}(n),$	(2×1)
$\varphi(n) = \varphi_1(n) - \mathbf{e}_{\mathrm{B}_2}^H(n)\mathbf{m}(n),$	(1×1)
$\mathbf{A}(n) = \mathbf{A}(n-1) + \mathbf{G}(n-1)\mathbf{e}_{\mathrm{A}}^H(n)/\varphi(n-1),$	$(2L \times 2)$
$\mathbf{E}_{\mathrm{A}}(n) = \lambda[\mathbf{E}_{\mathrm{A}}(n-1) + \mathbf{e}_{\mathrm{A}}(n)\mathbf{e}_{\mathrm{A}}^H(n)/\varphi(n-1)],$	(2×2)
$\mathbf{G}(n) = \mathbf{M}(n) + \mathbf{B}(n-1)\mathbf{m}(n),$	$(2L \times 1)$
$\mathbf{e}_{\mathrm{B}_1}(n) = \mathbf{E}_{\mathrm{B}}(n-1)\mathbf{m}(n),$	(2×1)
$\mathbf{e}_{\mathrm{B}}(n) = k\mathbf{e}_{\mathrm{B}_2}(n) + (1-k)\mathbf{e}_{\mathrm{B}_1}(n),$	(2×1)
$\mathbf{B}(n) = \mathbf{B}(n-1) + \mathbf{G}(n)\mathbf{e}_{\mathrm{B}}^H(n)/\varphi(n),$	$(2L \times 2)$
$\mathbf{E}_{\mathrm{B}}(n) = \lambda[\mathbf{E}_{\mathrm{B}}(n-1) + \mathbf{e}_{\mathrm{B}_2}(n)\mathbf{e}_{\mathrm{B}_2}^H(n)/\varphi(n)],$	(2×2)
Filtering	
$e(n) = y(n) - \hat{\underline{\mathbf{h}}}^H(n-1)\underline{\mathbf{x}}(n),$	(1×1)
$\hat{\underline{\mathbf{h}}}(n) = \hat{\underline{\mathbf{h}}}(n-1) + \mathbf{G}(n)e^*(n)/\varphi(n).$	$(2L \times 1)$

the two-path structure, E_{e_i}, multiplied with a fixed value $C < 1$,

$$E_{e_{RLS,i}} < CE_{e_i}, \tag{18}$$

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

3.3 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 2, 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 downsampling ratio r as number of subbands M, aliasing will significantly decrease the performance of the adaptive filters. Therefore non-critical downsampling, i.e. $r < M$, in conjunction with filters that have good stopband attenuation were chosen. General discussions of filterbanks can be found in numerous books and articles [17], [18], [19], [20], but since the emphasis has been on critical downsampled filterbanks, an efficient structure with non-critical downsampling is shown in the Appendix. Methods on how to design prototype filters are also discussed.

3.4 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 1. This algorithm needs $128L_{\text{sub}}$ real-valued multiplications and $128L_{\text{sub}}$ real additions, where L_{sub} denotes the length of the adaptive filter in the subbands. Echo

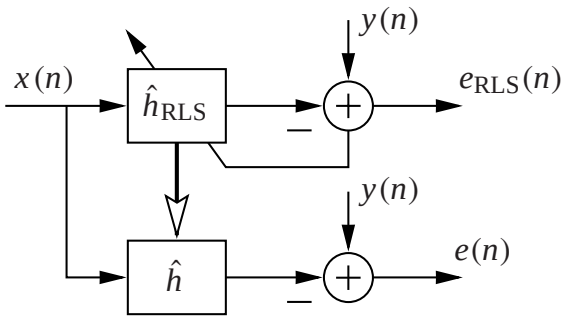

Figure 3: Two-path adaptive filtering.

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_{nc} compensates for the non-causal taps [21]. The total number of real multiplications per fullband sample is $128\frac{M/2+1}{r}L_{\text{sub}}$ and the number of additions $128\frac{M/2+1}{r}L_{\text{sub}}$.

The analysis, (29), and synthesis, (38), 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, they 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 $\frac{1}{r}(4K + 4M\log_2 M - 14M + 24)$ and the total number of additions is $\frac{1}{r}(4K + 6M\log_2 M - 2M)$. In the synthesis filterbank, K extra additions are needed to update the state vector in each filterbank, (38). The total number of multiplications for the two synthesis filterbanks is $\frac{1}{r}(2K + 2M\log_2 M - 7M + 12)$ and the number of additions $\frac{1}{r}(4K + 3M\log_2 M - M)$. In Table 2, 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.

4 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 canceler had the following settings.

Table 2: Calculation complexity comparison given as number of 10^6 real-valued multiplications per second at 16 kHz sample period. Corresponding fullband impulse response length $L = 3168$, downsampling rate $r = \frac{3}{4}M$, 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.

System configuration					
Number of subbands, M	1	16	32	64	128
Prototype filter length, K	–	207	415	831	1663
Signal delay (ms)	0	13	26	52	104
System components complexity					
Filterbank	–	1.8	1.9	2.0	2.1
Two-channel FRLS	1.6k	410	200	100	53
NLMS	410	100	50	25	13
Two-path	200	52	25	12	6.6
System complexity (adaptive filter, filterbank, two-path)					
FRLS	1.8k	470	230	110	61
Hybrid NLMS/FRLS	–	290	150	76	41
NLMS	610	160	76	39	22

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 non-causal taps [21] : 300.

As shown in Section 2.1, 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,

$$\gamma(f) = \frac{|S_{x_1x_2}(f)|}{\sqrt{S_{x_1x_1}(f)S_{x_2x_2}(f)}}, \tag{19}$$

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

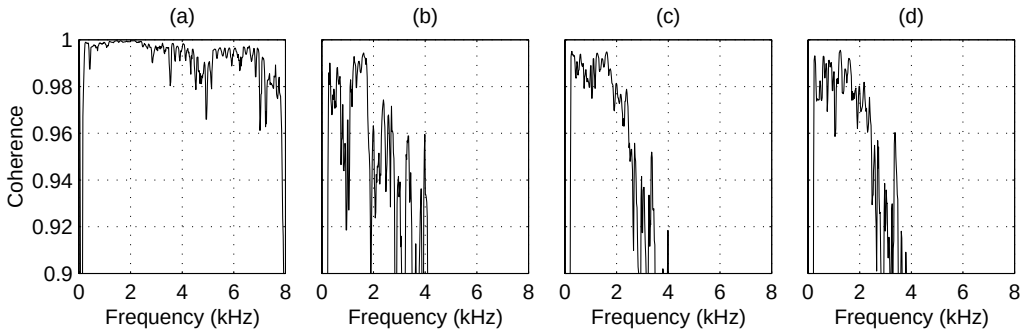

Figure 4: Magnitude coherence between right and left transmission room signal in the echo canceler. Average signal to noise ratio is approximately 40 dB. (a) Unprocessed transmission room signals. (b) Signals pre-processed with the half-wave rectifier, $\alpha = 0.5$. (c) Signal coded/decoded with an MPEG layer III audio coder [11], coded at 32 kbit/s per channel. (d) MTPC coder [12], same conditions as Fig. 4(c).

function is shown for speech signals recorded as described above. The power spectral estimates, $S_{x_i x_j}$, 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. Figure 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 pre-processed with half-wave rectifiers, Section 2.2, 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], [25] and finally in Fig. 4(d) the result using an MTPC coder [12] is shown. Both coders coded the left and right channels separately at 32 kbit/s per channel.

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[1] (MSE) energy of the residual is used. The MSE is given by,

$$\text{MSE} = \frac{P_{e-w}}{P_{y-w}}, \tag{20}$$

$$P_{e-w} = \text{LPF}[e(n) - w(n)]^2, \tag{21}$$

[1] Since we normalize with the power of the echo. We can regard this as the inverse of the echo return loss enhancement (ERLE).

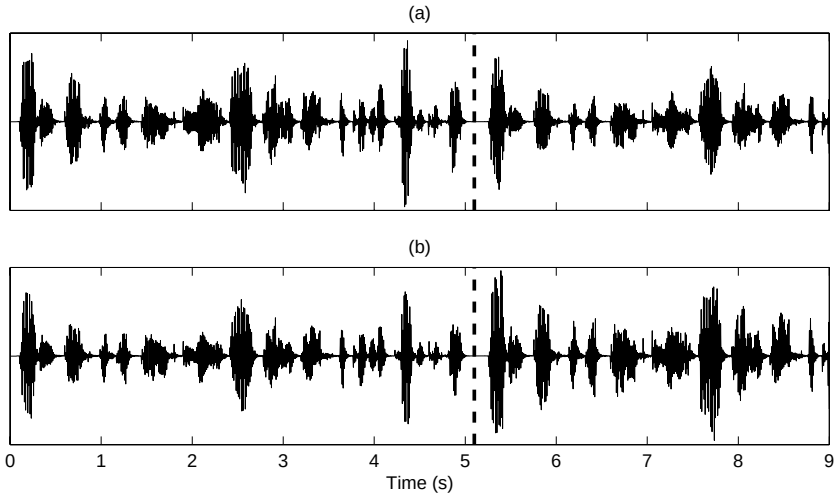

Figure 5: Transmission room signals used as excitation signal in Figs. 6–8. The speaker changes position at 5.1 s. (a) Left channel. (b) Right channel.

where w denotes the receiving room background noise signal and LPF denotes a low-pass filter; in this case it has a single real pole at 0.999. P_{y-w} is analogously calculated. In all examples, except the double-talk example, the background noise signal w is unknown, and cannot be subtracted according to (21). 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 seconds. 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 seconds. 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 towards 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 seconds. Simulations have shown similar behavior for the coded/decoded signals used in Fig. 4(c,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 $\mathbf{R}_{xx}(n)$ (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

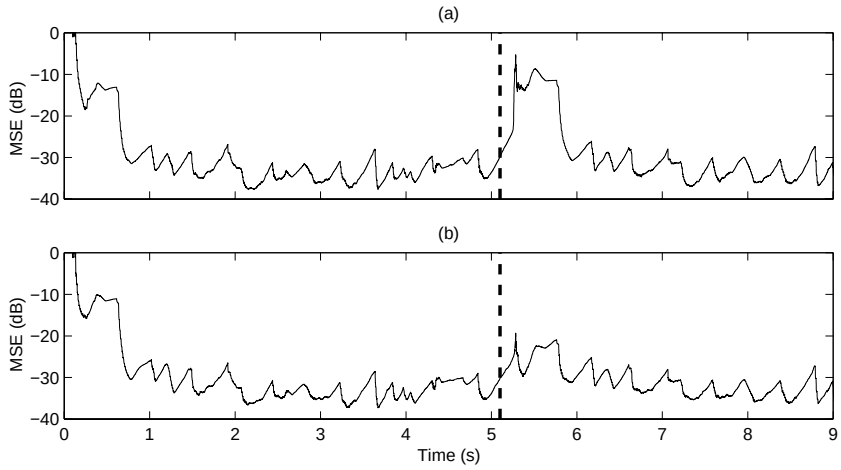

Figure 6: Residual echo after the echo canceler without non-linear residual echo suppression. The clip shows the convergence on a speech signal with 16 kHz sampling rate, and a change of position of the transmission room speaker at 5.1 s. The SNR was approximately 40 dB in both the transmission room and the receiving room. (a) MSE, see (20), of the residual, unprocessed signal. (b) MSE of signal processed with the half-wave rectifier, $\alpha = 0.5$

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 room speaker changes positions. Figure 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.

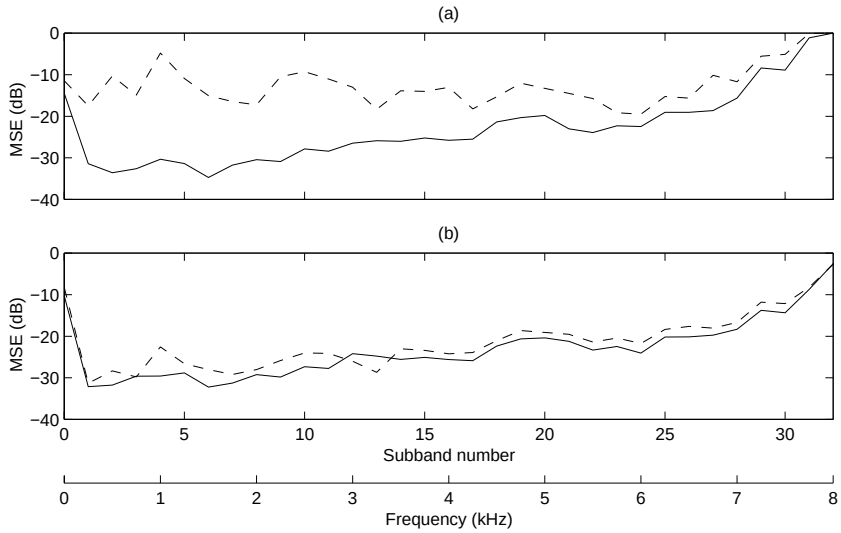

Figure 7: Subband MSE, see (20), at two time instances, solid line at 5.1 s and dashed line at 5.4 s, where the latter is directly after a speaker position change in the transmission room. (a) Unprocessed signal, the same conditions as in Fig. 6(a). (b) Signal processed with the half-wave rectifier, $\alpha = 0.5$, like in Fig. 6(b).

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, (20). 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.

5 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,

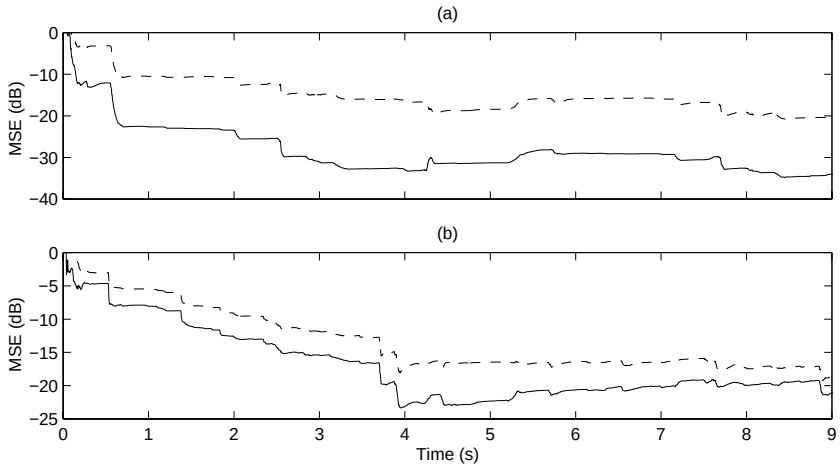

Figure 8: Subband convergence comparison between the two-channel FRLS and NLMS algorithms. The figures show the MSE performance for one typical lower and one typical higher subband. Solid line FRLS, dashed line NLMS. (a) subband 7 (center frequency at 1.75 kHz), (b) subband 18 (center frequency at 4.5 kHz).

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 hours of real-life data was processed, each subband was restarted on average every 18 s.

The real-time system also includes a device to suppress the residual echo after the adaptive filter, see Fig. 2. 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 [26], 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

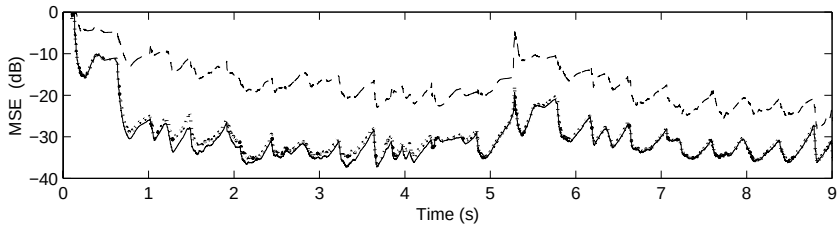

Figure 9: Fullband convergence comparison between the two-channel FRLS and NLMS algorithms. The figure shows the MSE performance for different SAECR setups. Solid line: FRLS is used in all subbands. Dashed line: NLMS is used in all subbands. Dotted line: FRLS in the lower 16 subband, NLMS in the upper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

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, $\hat{h}_{\mathrm{RLS}}$ in Fig. 3, is distributed, whereas the filtering, $\hat{h}$ in Fig. 3, is performed on the DSP which also executes the filterbanks [27]. This way, no extra signal path delay is introduced by the parallel structure, see Appendix E. 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 2.

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.

Acknowledgment

The authors would especially like to thank Dr. Gary W. Elko. Without his enthusiastic engagement, this project would not have been nearly as successful. The authors also highly value the dialogue with Dr. Dennis Morgan during the project.

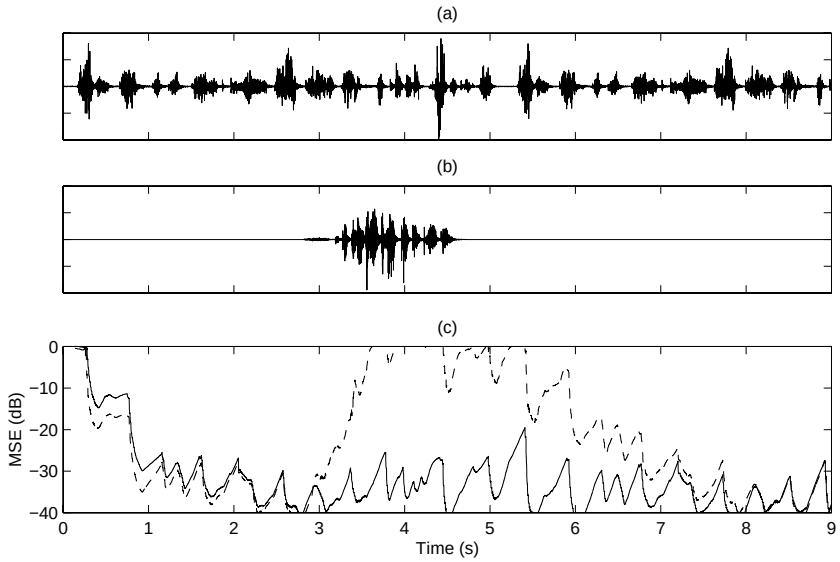

Figure 10: Double-talk situation, only the left channel is shown. (a) Echo-signal, without double-talk. (b) Double-talk signal. (c) MSE performance: SAECR without the two-path structure (dashed line). SAECR With the two-path structure (solid line).

Appendix

A Analysis Filterbank

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

$$x_m(k) = \sum_{l=0}^{K-1} f_m(l)x(rk-l), \tag{22}$$

where the subband filter is denoted $f_m(n)$ with length K and the downsampling factor is r, see Fig. 11. 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, $\mathbf{f}_m^T = \mathbf{w}_m^T\mathbf{F}$, where

$$\mathbf{f}_m = \begin{bmatrix} f_m(0) & f_m(1) & \dots & f_m(K-1) \end{bmatrix}^T, \tag{23}$$

$$\mathbf{w}_m = \begin{bmatrix} 1 & e^{j\frac{2\pi m}{M}(1)} & \dots & e^{j\frac{2\pi m}{M}(M-1)} \end{bmatrix}^T, \tag{24}$$

$$\mathbf{F} = \begin{bmatrix} \text{diag}(\tilde{\mathbf{f}}_0) & \text{diag}(\tilde{\mathbf{f}}_M) & \dots & \text{diag}(\tilde{\mathbf{f}}_{K-M}) \end{bmatrix}. \tag{25}$$

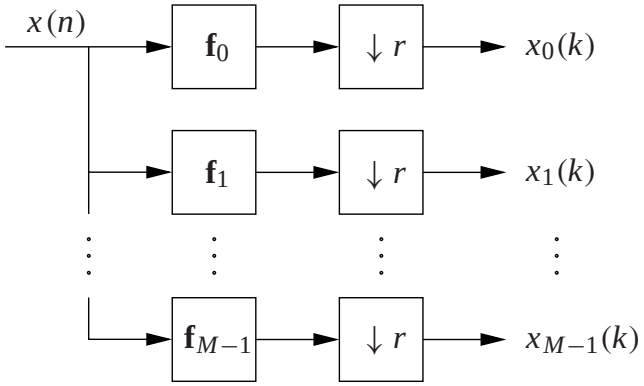

Figure 11: Analysis filterbank: Each subband signal $x_m(k)$ can be found by band-pass filtering the fullband signal $x(n)$ with $f_m(n)$ and then downsampling it by a factor r.

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

$$\text{diag}(\tilde{\mathbf{f}}_i) = \begin{bmatrix} f(i) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & f(i+M-1) \end{bmatrix}. \tag{26}$$

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 $x_m(k) = \mathbf{w}_m^T \mathbf{F}\mathbf{x}(k)$, where the fullband input vector $\mathbf{x}(k)$ is defined as

$$\mathbf{x}(k) = \begin{bmatrix} x(rk) & x(rk-1) & \dots & x(rk-K+1) \end{bmatrix}^T. \tag{27}$$

By exchanging the modulation vector $\mathbf{w}_m$ for the modulation matrix,

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_0 & \mathbf{w}_1 & \dots & \mathbf{w}_{M-1} \end{bmatrix}^T, \tag{28}$$

a vector containing all subband samples at frame k may be calculated as,

$$\begin{bmatrix} x_0(k) & x_1(k) & \dots & x_{M-1}(k) \end{bmatrix}^T = \mathbf{W}\mathbf{F}\mathbf{x}(k). \tag{29}$$

The efficient DFT polyphase filterbank structure can be seen in (29). Since $\mathbf{F}$ is a real sparse matrix, $\mathbf{F}\mathbf{x}(k)$ can be calculated with K real multiplications. Secondly, the modulation matrix divided by the number of subbands, $\mathbf{W}/M$, is also known as the inverse DFT matrix. Therefore, the calculation complexity of (29) can be reduced by

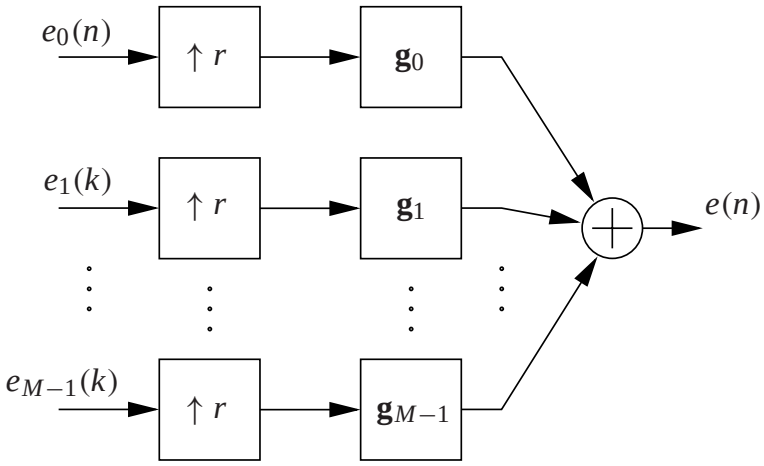

Figure 12: Synthesis filterbank: The reconstructed fullband signal is a sum of upsampled and band-pass filtered subband signals.

using efficient inverse fast Fourier transforms (IFFT). Let us return to (28) once again. Due to the symmetry in the modulation vector (24), the following relation between rows can be seen in (28),

$$\mathbf{w}_i = \mathbf{w}_{M-i}^* \qquad 1 \leq i \leq \frac{M}{2} - 1 \tag{30}$$

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 $\mathbf{F}$ 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.

B Synthesis Filterbank

The synthesis filterbank reconstructs the fullband signal, $e(n)$, by a summation of the interpolated and filtered subband signals $e_m(k)$, see Fig. 12. For reasons that will be seen later in the section, it is advantageous to use a state description of the filterbank filters, illustrated in Fig. 13. The input signal to the filterbank, $e_m(k)$, is upsampled r times. Then, K copies of the interpolated signal are each multiplied with the corresponding synthesis filterbank filter-tap, $g_m(l)$. Each sample factor is added to a state variable $s_{m,l}(n)$, in the state vector $\mathbf{s}_m(n)$. For every fullband sample interval, the state vector is shifted one position. The fullband signal can then be calculated as

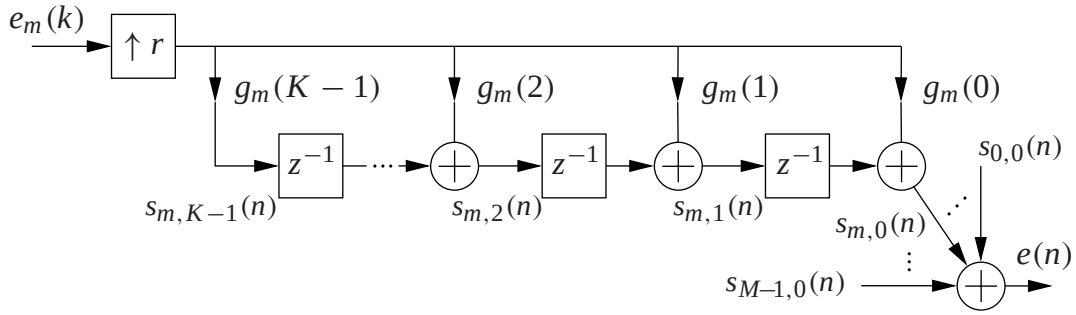

Figure 13: State representation of synthesis filter: The subband signal $e_m(k)$ is first upsampled. L delays of the signal are each multiplied by the corresponding filter-tap $g_m(l)$ and added to the state variable $s_{m,l}(n)$. The state vector $\mathbf{s}_m(n)$ is shifted each sample-interval. The reconstructed signal can then be found as the sum of the state variables, $e(n) = \sum_{m=0}^{M-1} s_{m,0}(n)$.

the sum of the upper state variables, $e(n) = \sum_{m=0}^{M-1} s_{m,0}(n)$. Let us now define the state vector as one upper and one lower vector,

$$\mathbf{s}_m(k) = \begin{bmatrix} \mathbf{s}_m^u(k) \\ \mathbf{s}_m^l(k) \end{bmatrix}, \tag{31}$$

$$\mathbf{s}_m^u(k) = \begin{bmatrix} s_{m,0}(kr) & s_{m,1}(kr) & \dots & s_{m,r-1}(kr) \end{bmatrix}^T, \tag{32}$$

$$\mathbf{s}_m^l(k) = \begin{bmatrix} s_{m,r}(kr) & s_{m,r+1}(kr) & \dots & s_{m,K-1}(kr) \end{bmatrix}^T. \tag{33}$$

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,

$$\begin{bmatrix} e(kr-r+1) & e(kr-r+2) & \dots & e(kr) \end{bmatrix}^T = \sum_{m=0}^{M} \mathbf{s}_m^u(k). \tag{34}$$

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

$$\mathbf{g}_m = \begin{bmatrix} g_m(0) & g_m(1) & \dots & g_m(K-1) \end{bmatrix}^T, \tag{35}$$

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

$$\mathbf{G} = \begin{bmatrix} \mathrm{diag}(\tilde{\mathbf{g}}_0) & \mathrm{diag}(\tilde{\mathbf{g}}_M) & \dots & \mathrm{diag}(\tilde{\mathbf{g}}_{K-M}) \end{bmatrix}^T, \tag{36}$$

where $\mathrm{diag}(\tilde{\mathbf{g}}_0)$ is defined as in (26) and $\mathbf{w}_m$ is defined in (24). As mentioned above, the state vectors can be updated on a frame basis. The state vectors are shifted r positions and the input subband sample, multiplied with $\mathbf{g}_m$, is added,

$$\mathbf{s}_m(k) = \begin{bmatrix} \mathbf{s}_m^l(k-1) \\ \mathbf{0}_{r\times 1} \end{bmatrix} + \mathbf{G}\mathbf{w}_m e_m(k). \tag{37}$$

When reconstructing the output signal $e(n)$, it is enough to know the sum of the state vectors, as is shown in (34). Therefore, the state vectors do not need to be updated individually, instead it is sufficient to update the sum of the state vectors, as

$$\mathbf{s}(k) = \begin{bmatrix} \mathbf{s}^l(k-1) \\ \mathbf{0}_{r\times 1} \end{bmatrix} + \mathbf{G}\mathbf{W}\mathbf{e}^{\mathrm{sub}}(k), \tag{38}$$

where $\mathbf{W}$ is defined in (28) and $\mathbf{e}^{\mathrm{sub}}(k) = \begin{bmatrix} e_0(k) & e_1(k) & \dots & e_{M-1}(k) \end{bmatrix}^T$. Finally, (38) 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.

C Prototype Filter for Non-critical Downsampling

In this Appendix, we will formulate a filterbank design method as a minimization problem. Quadratic programming has been found to be an efficient method for solving this minimization problem, and in Appendix D the minimization criteria is reformulated for a quadratic programming algorithm.

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 non-critical 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,

$$\frac{1}{r}\sum_{m=0}^{M-1} F(e^{j(\omega-\frac{2\pi m}{M})})G(e^{j(\omega-\frac{2\pi m}{M})}) = e^{-j\omega K}, \tag{39}$$

where K is the prototype filter length, M the number of subbands, r downsampling factor, $F(z) = \sum_{k=0}^{K-1} f(k)z^{-k}$ the analysis filter, and $G(z) = \sum_{k=0}^{K-1} g(k)z^{-k}$ the synthesis filter. In order to guarantee linear phase, we let $G(z) = z^{-K}F(z^{-1})$. If we define the non-causal 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 (39) would be,

$$\frac{1}{r}\sum_{m=0}^{M-1} R[e^{j(\omega-\frac{2\pi m}{M})}] = 1. \tag{40}$$

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

$$\Phi_1 = \int_0^{\frac{2\pi}{M}-\frac{\pi}{r}} [R(e^{j\omega}) - r]^2 d\omega. \tag{41}$$

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

$$\Phi_2 = \int_{\frac{2\pi}{M}-\frac{\pi}{r}}^{\frac{\pi}{r}} [R(e^{j\omega}) + R(e^{j(\omega-\frac{2\pi}{M})}) - r]^2 d\omega. \tag{42}$$

Finally, we need to minimize aliasing by enforcing good stopband attenuation,

$$\Phi_3 = \int_{\frac{\pi}{r}}^{\pi} [R(e^{j\omega})]^2 d\omega. \tag{43}$$

The total minimization problem can now be expressed as

$$\min_{R(z)} \quad \alpha_1\Phi_1 + \alpha_2\Phi_2 + \alpha_3\Phi_3, \tag{44}$$

subject to the constraints,

$$\tilde{r}(0) \geq \tilde{r}(n), \quad n \neq 0, \tag{45}$$

$$\tilde{r}(n) = \tilde{r}^*(-n), \quad \forall n, \tag{46}$$

where $\alpha_i \geq 0$ are trade-off parameters and $R(z) = \sum_{k=-K+1}^{K-1} \tilde{r}(k)z^{-k}$. Quadratic programming [28] 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. 14. In [29], a method for calculating $F(z)$ from a given $R(z)$ is presented. An alternative method to design filters that satisfies (39) can be found in [30]. This method has more stringent requirements, forcing Φ_1 and Φ_2 to be equal to zero, but also needs a larger filter length K in order to achieve good stopband attenuation.

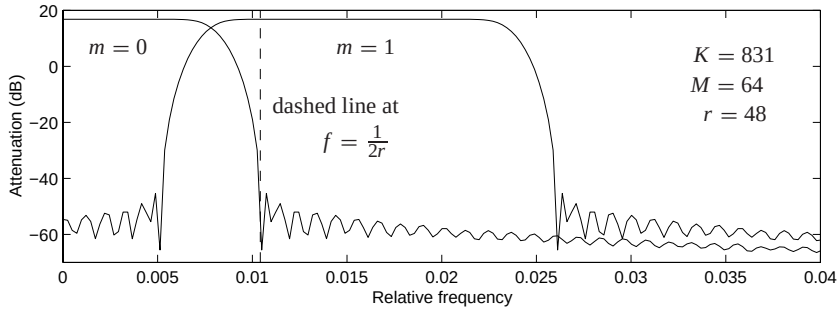

Figure 14: Filterbank for subband $m = 0$ and $m = 1$. Aliasing for subband 0 occurs around $f = \frac{1}{2r}$.

D Prototype Filter Construction

In this section it will be shown how (44) can be numerically determined with a quadratic programming algorithm [28]. Implementations of this algorithm is available in commercial software packages such as in the Optimization toolbox in MATLAB.

In quadratic programming, the following function is minimized,

$$\min_{\mathbf{r}} \quad \frac{1}{2}\mathbf{r}^H\mathbf{H}\mathbf{r} + \mathbf{f}^H\mathbf{r}, \tag{47}$$

under the constraint,

$$\mathbf{A}\mathbf{r} \leq \mathbf{b}, \tag{48}$$

where the number of parameters to minimize is L. The matrices $\mathbf{H}$ and $\mathbf{A}$ are of size $L \times L$, and $\mathbf{r}$ and $\mathbf{f}$ are of size $L \times 1$. We now need to determine the matrices $\mathbf{H}$, $\mathbf{A}$, $\mathbf{r}$ and $\mathbf{f}$ so that the minimization problem in (47) and (48) equals the minimization problem in (44), (45) and (46).

In Appendix C, we have seen that it is sufficient to minimize the non-causal filter $R(z)$. We can evaluate this function on the unit circle, $z = e^{j\omega}$, as

$$R(e^{j\omega}) = \tilde{r}_L e^{j\omega L} + \tilde{r}_{L-1} e^{j\omega(L-1)} + \cdots + \tilde{r}_0 e^{j0} + \cdots + \tilde{r}_L e^{-j\omega L} \tag{49}$$

$$= \tilde{r}_0 + 2\sum_{l=1}^{L} \cos(l\omega)\tilde{r}_l. \tag{50}$$

It is possible to determine the value of $\tilde{r}_0$. By integrating (40) we get,

$$2\pi r = \int_{-\pi}^{\pi} \sum_{m=0}^{M-1} R\left(e^{j\left(\omega - \frac{2\pi m}{M}\right)}\right) d\omega. \tag{51}$$

Now we replace $R(e^{j(\omega-\frac{2\pi m}{M})})$ with the definition in (50), and change the order of integration and summation,

$$2\pi r = \int_{-\pi}^{\pi} \sum_{m=0}^{M-1} \tilde{r}_0 + 2\sum_{l=1}^{L} \cos\left(l\omega - \frac{2\pi ml}{M}\right)\tilde{r}_l d\omega \tag{52}$$

$$= 2\pi M\tilde{r}_0 + \sum_{m=0}^{M-1} 2\sum_{l=1}^{L} \underbrace{\int_{-\pi}^{\pi} \cos\left(l\omega - \frac{2\pi ml}{M}\right)\tilde{r}_l d\omega}_{=0}. \tag{53}$$

From (53) we can determine that $\tilde{r}_0 = \frac{r}{M}$. That is, we now have L parameters to minimize,

$$\mathbf{r} = \begin{bmatrix}\tilde{r}_1 & \tilde{r}_2 & \cdots & \tilde{r}_L\end{bmatrix}. \tag{54}$$

Next we will evaluate the integrals in (41), (42), and (43), starting with (41).

$$\Phi_1 = \int_0^{\frac{2\pi}{M}-\frac{\pi}{r}} \left[R(e^{j\omega}) - r\right]^2 d\omega \tag{55}$$

$$= \int_0^{\frac{2\pi}{M}-\frac{\pi}{r}} \left[\frac{r}{M} - r + 2\sum_{l=1}^{L}\cos(l\omega)\tilde{r}_l\right]^2 d\omega \tag{56}$$

$$= \Bigg[\left(\frac{r}{M} - r\right)^2 \omega + 4\left(\frac{r}{M} - r\right)\sum_{l=1}^{L}\frac{\sin(l\omega)}{l}\tilde{r}_l + 2\sum_{l=1}^{L}\left(\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega\right)\tilde{r}_l^2 + 4\sum_{l=1}^{L-1}\sum_{p=l+1}^{L}\left(\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l}\right)\tilde{r}_l\tilde{r}_p\Bigg]_0^{\frac{2\pi}{M}-\frac{\pi}{r}}. \tag{57}$$

Similarly, Φ_2 is,

$$\Phi_2 = \int_{\frac{2\pi}{M}-\frac{\pi}{r}}^{\frac{\pi}{r}} \left[R(e^{j\omega}) + R(e^{j(\omega-\frac{2\pi}{M})}) - r\right]^2 d\omega = \tag{58}$$

$$= \left[\left(\frac{2r}{M} - r \right)^2 \omega + 4 \left(\frac{2r}{M} - r \right) \sum_{l=1}^{L} \left(\frac{\sin(l\omega)}{l} + \frac{\sin(l\omega - \frac{2\pi l}{M})}{l} \right) \tilde{r}_l + \right.$$
$$2 \sum_{l=1}^{L} \left(\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega + 2\cos\left(\frac{2\pi l}{M}\right)\omega + \frac{\sin(2l\omega - \frac{2\pi l}{M})}{l} + \right.$$
$$\left. \frac{\cos(l\omega - \frac{2\pi l}{M})\sin(l\omega - \frac{2\pi l}{M})}{l} + \omega - \frac{2\pi}{M} \right) \tilde{r}_l^2 +$$
$$4 \sum_{l=1}^{L-1} \sum_{p=l+1}^{L} \left(\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l} + \frac{\sin(p\omega - l\omega - \frac{2\pi p}{M})}{p-l} + \right.$$
$$\frac{\sin(p\omega + l\omega - \frac{2\pi p}{M})}{l+p} + \frac{\sin(p\omega - l\omega + \frac{2\pi l}{M})}{p-l} + \frac{\sin(l\omega + p\omega - \frac{2\pi l}{M})}{l+p} +$$
$$\left. \left. \frac{\sin(p\omega - l\omega - \frac{2\pi(p-l)}{M})}{p-l} + \frac{\sin(l\omega + p\omega - \frac{2\pi(l+p)}{M})}{l+p} \right) \tilde{r}_l \tilde{r}_p \right]_{\frac{2\pi}{M} - \frac{\pi}{r}}^{\frac{\pi}{r}}. \quad (59)$$

Finally Φ_3 is,

$$\Phi_3 = \int_{\frac{\pi}{r}}^{\pi} \left(R(e^{j\omega}) \right)^2 d\omega \quad (60)$$
$$= \left[(\frac{r}{M})^2 \omega + \frac{4r}{M} \sum_{l=1}^{L} \frac{\sin(l\omega)}{l} \tilde{r}_l + 2 \sum_{l=1}^{L} \left(\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega \right) \tilde{r}_l^2 + \right.$$
$$\left. 4 \sum_{l=1}^{L-1} \sum_{p=l+1}^{L} \left(\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l} \right) \tilde{r}_l \tilde{r}_p \right]_{\frac{\pi}{r}}^{\pi}. \quad (61)$$

Let us start with determining the matrix $\mathbf{H}$. Since the matrix $\mathbf{H}$ in (47) is pre-multiplied with $\mathbf{r}^H$ and post-multiplied with $\mathbf{r}$, the elements in $\mathbf{H}$ correspond to the coefficients in front of all $\tilde{r}_l^2$ and all $\tilde{r}_l \tilde{r}_p$ in (57), (59) and (61). We will start to derive a non-symmetric matrix $\mathbf{H}'$, and later transform it to the symmetric matrix $\mathbf{H}$.

Let us begin with determining the elements on the diagonal of $\mathbf{H}'$. In (47), the operation $\mathbf{r}^H \mathbf{H} \mathbf{r}$ lead to that all diagonal elements are multiplied by $\tilde{r}_l^2$, where l is given by the position and $\mathbf{r}$ is defined in (54). That is, the top left element in $\mathbf{H}$ is multiplied by $\tilde{r}_1^2$ and so on. By examining (57), (59) and (61), we can determine the

diagonal elements,

$$[\mathbf{H}']_{(l-1,l-1)} = \alpha_1 \cdot \underbrace{2\left[\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega\right]_0^{\frac{2\pi}{M}-\frac{\pi}{r}}}_{\text{from } \phi_1 \text{ in (57)}} + \alpha_2 \cdot \underbrace{2\left[\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega + 2\cos\left(\frac{2\pi l}{M}\right)\omega + \frac{\sin(2l\omega - \frac{2\pi l}{M})}{l} + \frac{\cos(l\omega - \frac{2\pi l}{M})\sin(l\omega - \frac{2\pi l}{M})}{l} + \omega - \frac{2\pi}{M}\right]_{\frac{2\pi}{M}-\frac{\pi}{r}}^{\frac{\pi}{r}}}_{\text{from } \phi_2 \text{ in (59)}} + \alpha_3 \cdot \underbrace{2\left[\frac{\cos(l\omega)\sin(l\omega)}{l} + \omega\right]_{\frac{\pi}{r}}^{\pi}}_{\text{from } \phi_3 \text{ in (61)}}, \qquad 1 \le l \le L, \tag{62}$$

where α_1, α_2 and α_3 are from (44), and $[\mathbf{H}']_{(l,p)}$ denotes the element in row l, column p of the matrix $\mathbf{H}'$. Next we determine the rest of elements in $\mathbf{H}'$. If we study the upper left triangular part of $\mathbf{H}$, we can conclude that these elements are multiplied with $\tilde{r}_l \tilde{r}_p$ in the operation $\mathbf{r}^H\mathbf{H}\mathbf{r}$ in (47), where $\mathbf{r}$ is defined in (54). Once again, we examine (57), (59) and (61), and determine that, the upper left triangular part of $\mathbf{H}'$ is

$$[\mathbf{H}']_{(l-1,p-1)} = \alpha_1 \cdot \underbrace{4\left[\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l}\right]_0^{\frac{2\pi}{M}-\frac{\pi}{r}}}_{\text{from } \phi_1 \text{ in (57)}} +$$

$$\alpha_2 \cdot 4 \underbrace{\left[\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l} + \frac{\sin(p\omega - l\omega - \frac{2\pi p}{M})}{p-l} + \right.}_{\text{from } \phi_2 \text{ in (59)}}$$

$$\underbrace{\frac{\sin(p\omega + l\omega - \frac{2\pi p}{M})}{l+p} + \frac{\sin(p\omega - l\omega + \frac{2\pi l}{M})}{p-l} + \frac{\sin(l\omega + p\omega - \frac{2\pi l}{M})}{l+p} +}_{\text{from } \phi_2 \text{ in (59)}}$$

$$\underbrace{\left. \frac{\sin(p\omega - l\omega - \frac{2\pi(p-l)}{M})}{p-l} + \frac{\sin(l\omega + p\omega - \frac{2\pi(l+p)}{M})}{l+p} \right]_{\frac{2\pi}{M} - \frac{\pi}{r}}^{\frac{\pi}{r}}}_{\text{from } \phi_2 \text{ in (59)}} +$$

$$\alpha_3 \cdot 4 \underbrace{\left[\frac{\sin(p\omega - l\omega)}{p-l} + \frac{\sin(p\omega + l\omega)}{p+l} \right]_{\frac{\pi}{r}}^{\pi}}_{\text{from } \phi_3 \text{ in (61)}},$$

$$1 \le l \le L-1, \quad l < p \le L. \tag{63}$$

So far the lower left triangular part of $\mathbf{H}'$ is zero. The quadratic programming algorithm requires the matrix $\mathbf{H}$ to be symmetric. We can fix this and at the same time compensate for the factor $\frac{1}{2}$ in (47),

$$\mathbf{H} = \mathbf{H}'^{T} + \mathbf{H}'. \tag{64}$$

The elements in the vector $\mathbf{f}$ correspond to the terms in (57), (59), and (61) which include the non-quadratic coefficient $\tilde{r}_l$. Like before, we examine (57), (59) and (61), and determine that,

$$[\mathbf{f}]_{(l-1,0)} = \alpha_1 \cdot \underbrace{4\left(\frac{r}{M} - r\right)\left[\frac{\sin(l\omega)}{l}\right]_0^{\frac{2\pi}{M} - \frac{\pi}{r}}}_{\text{from } \phi_1 \text{ in (57)}} +$$

$$\underbrace{\alpha_2 \cdot 4\left(\frac{2r}{M} - r\right)\left[\frac{\sin(l\omega)}{l} + \frac{\sin(l\omega - \frac{2\pi l}{M})}{l}\right]_{\frac{2\pi}{M} - \frac{\pi}{r}}^{\frac{\pi}{r}}}_{\text{from } \phi_2 \text{ in (59)}} +$$

$$\underbrace{\alpha_3 \cdot \frac{4r}{M}\left[\frac{\sin(l\omega)}{l}\right]_{\frac{\pi}{r}}^{\pi}}_{\text{from } \phi_3 \text{ in (61)}}, \qquad 1 \le l \le L. \tag{65}$$

Finally, we will determine the constraints in (48). From (45) we know that $\tilde{r}(0) \geq \tilde{r}(n)$. This is satisfied if $\mathbf{A}$ is the identity matrix of size $L \times L$, and all elements in $\mathbf{b}$ equal $\tilde{r}_0$, i.e., all elements in $\mathbf{b}$ equal $\frac{r}{M}$.

The total response of the filterbank, $R(z)$, can now be determined with a quadratic programming algorithm. This filter can then be divided into one analysis and one synthesis filter, $F(z)$ and $G(z)$ respectively, so that $R(z) = F(z)G(z)z^{-K}$. This relation is described in Appendix C and an appropriate method to split the filter $R(z)$ is given in [29].

E Real-time Implementation

In this section, the real-time implementation of the stereophonic acoustic echo canceler is described. All vital parts of an echo canceler have been implemented and the underlying hardware is a parallel floating-point DSP-board. Each board is equipped with four DSPs, the TMS320C44 from Texas Instruments. Special care was taken to create a general implementation. That is, by the means of a re-compilation, it is possible to change variables such as the sampling rate, the number of subbands, and the downsampling factor. The main program was written in C. However, all time critical code sections are hand optimized TMS320C44 assembly code. The echo canceler has also been ported to run natively on a PC, under MATLAB using MATLAB's standard C interface. All simulations in this paper have been performed using this latter port of the program.

Section 3 describes the basic stereophonic acoustic echo canceler depicted in Fig. 1. As described in Section 1 and Section 3.4, the calculation complexity of a fullband implementation is high and therefore a subband realization was chosen, illustrated by a block scheme in Fig. 2. All parts, except the adaptive filter, require dual units, one for each channel, whereas the adaptive filter is an integrated unit for the two channels, since the same Kalman gain calculation is needed for both channels.

Decomposition of Code for Parallel Execution

A subband approach is not only advantageous from a calculation complexity perspective, it also usually results in code that can execute efficiently in parallel on multiple execution units. That is, with a reasonable amount of data transmission, the code can be distributed over a number of DSPs. The two-channel FRLS adaptive filters used in the implementation represent the majority of the calculation complexity, and it was therefore decided that out of the four DSPs, three would be used to execute the adaptive filters. These three DSPs are denoted B,C and D, and each one estimates the

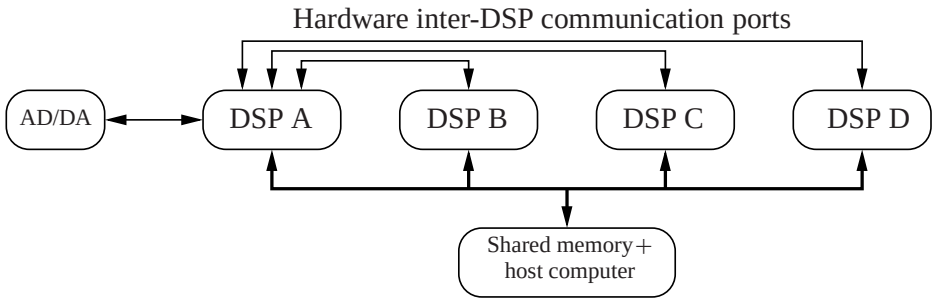

Figure 15: DSP-board: The AD/DA converter is connected to the master DSP, DSP A. This DSP has bi-directional parallel communication ports to the other DSPs. All DSPs are connected to the shared memory and the host computer with a data buss. Each DSP is also equipped with two banks of zero wait state local memory. In the two-board configuration, the boards communicate over a communication port connected between the two DSP A's.

impulse response for one third of the subbands. That leaves the four analysis filterbanks, the two synthesis filterbanks, the filtering part of the two-path structure and post-processing of the residual signals for DSP A. Figure 15 shows the communication paths between the different units. All data transfers between DSPs are done over the bi-directional parallel communication ports that are available on the TMS320C4x. The direct memory access (DMA) coprocessors on this DSP can move blocks of data over the communication ports without interaction from the central processing unit (CPU), almost eliminating the overhead of the the inter-processor communication.

In Fig. 16, a timing diagram shows where and when the different parts of the echo canceler are executed. Let us study the process in DSP A during one frame. When one frame of samples has been received from the AD converter at t_0 the far-end and near-end signals are decomposed into subbands. After execution of each analysis filterbank, DMAs move the subband samples over the communication ports to the other DSPs. One third of the subbands are moved to DSP B, etc. As can be seen in Fig. 16, the results from the adaptive filters are not available until the next frame. Therefore DSP A must use the echo path estimates, $\hat{h}_{(\cdot)}$, from earlier frames in the actual echo cancellation, at time t_1. After the filtering, a post-processor reduces the residual echo in the subbands. Finally, at t_3, the synthesis filterbanks reconstruct the fullband echo canceled signals.

The DSPs executing the adaptive filter are shown in the lower half of Fig. 16. When all subband signals are received, at t_2, the execution of a new frame starts. First, the adaptive filters update the echo path estimates and calculate the residual signals. During this process, residuals from the same frame of input signals, but calculated with the previous echo-path estimates, are received from DSP A, at time t_4. These residuals

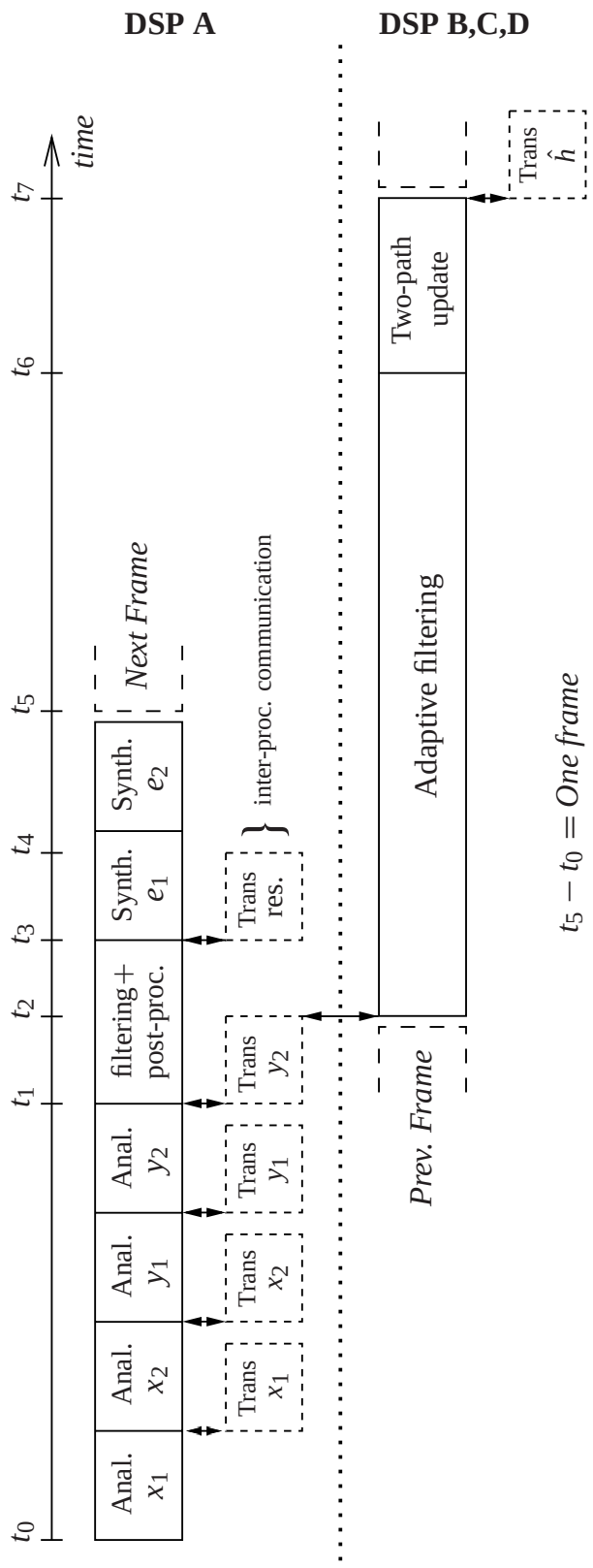

Figure 16: Timing diagram of execution. The left half of the figure describes DSP A, and the right half DSP B,C,D. Solid boxes show tasks performed by the CPU and dashed boxes tasks performed by the DMAs.

are compared according to the two-path structure at t_6 and for those subbands where the echo-path estimates have improved, the new estimates are sent back to DSP A at time t_7.

Note that even though the echo-path estimates are delayed in the structure, there is no extra delay in the actual signal path due to the parallel structure. The delay of the estimates is actually needed in the two-path structure, see Section 3.2.

Time Synchronization

During initialization, the AD/DA converter is set to its sampling rate and released from its halt state. After this point, it becomes the master timing device, operating independently of the DSPs. For each sample received, a hardware interrupt is generated by the AD/DA converter on DSP A, Fig. 15, and this interrupt is received by one of the DMAs. The DMA moves the samples to buffers in the memory, and moves samples from the memory buffers to the DA converter. When one frame of samples has been received, the DMA interrupts the CPU and the analysis filterbank starts executing. This point is denoted t_0 in Fig. 16. The DMA continues without interaction from the CPU to move the next frame of samples to another buffer in the memory. In a dual DSP board configuration, the DMA also moves samples directly to the other board over a communication port.

The DSPs processing the adaptive filters uses DMAs to move subband samples from the communication port to buffers in the memory. When samples from all four analysis filterbanks have been received, the DMA interrupts the CPU, and execution of the adaptive filters is started (see time t_2 in Fig. 16). The adaptive filters must complete this frame before the next frame has arrived at time $t_2 + \textit{one frame}$.

Analysis and Synthesis Filterbank Implementation

The theoretical description of the filterbank is given in Appendix A, and the analysis filterbank is implemented in assembler according to (29). The Fourier transform was initially implemented as a DFT, but it has later been converted to an FFT implementation. This has been done also for the synthesis filterbank, of which the implementation corresponds to (34) and (38).

Numerous complexity analyses of filterbanks can be found in the literature, e.g. in [17], [18], [19], [20]. However, the analyses usually include implementation dependent constants. Below follows an analysis of the number of clock cycles needed to execute one analysis filterbank on a TMS320C44. In this analysis, memory is assumed to be accessed without wait states.

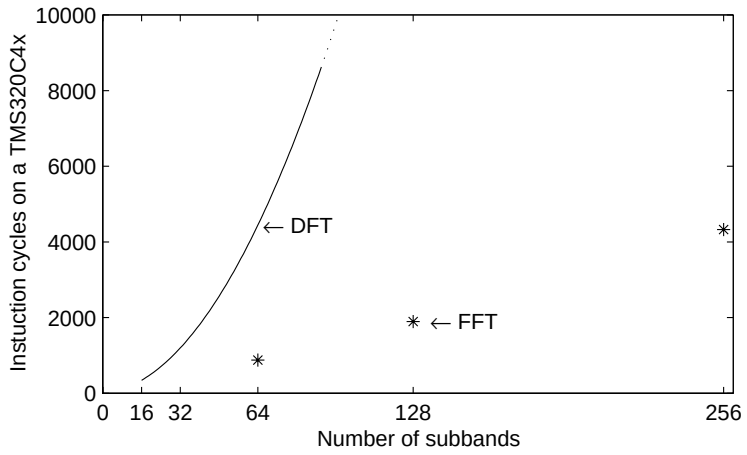

Figure 17: Comparison between a DFT and a radix-2 FFT with real-valued input signals on a TMS320C4x DSP, [31].

The filtering part involves a convolution with K elements. Due to the data storage structure imposed by the filterbank, the actual number of clock cycles needed for the operation are $K + 5M$, where K denotes the length of the prototype filter and M the number of subbands.

Secondly, an inverse discrete Fourier Transform must be performed. Actually, only the $M/2 + 1$ lowest subbands need to be calculated, as shown in Appendix A. Taking this into account plus the fact that the fullband signal is real-valued, a DFT implementation needs $M^2 + 5.5M$ cycles. FFTs are computationally efficient realizations of DFTs, especially for large transforms. However, for small transforms, the DFT is faster due to overhead in the FFT structure. Using the real FFT implementation presented in [31], the number of cycles needed, including bit-reversing, can be approximated with $2M \log_2 M + 2M$ cycles. The FFT structure requires the number of subbands to be a multiple of 2, and the complexity approximation represents an upper limit for transforms of size 64 to 1024, the sizes used in [31]. In Fig. 17, a comparison between the DFT and the FFT implementation is shown. Using an FFT, the total number of cycles needed can now be written as,

$$\text{Cycles}_{Analysis} = K + 2M \log_2 M + 7M + C \tag{66}$$

where C is a reasonably small number due to initialization overhead. A similar analysis can be performed for the synthesis filterbank.

Two-channel FRLS Implementation

The two-channel FRLS is the most complex block in the echo canceler, considering both structural and computational complexity. The implementation closely follows the equations given in Table 1 in Section 3.1, with the exception that Table 1 only describes cancellation of one return signal. The main structure was written in C, but all matrix primitives are coded in assembler for maximum performance.

In the implementation, the state of the two-channel FRLS in each subband is stored in a data structure. This structure is too big to fit in the internal memory in the TMS320C44, and using external memory for all the matrix operations would significantly increase the number of clock cycles needed. Therefore a DMA was set up to move data to and from external memory in a just-in-time manner. More specifically, the DMA operates on one of the external and one of the internal memory blocks while the CPU operates on the other external and the other internal memory blocks, maximizing the memory bandwidth usage.

References

[1] J. Benesty, D. R. Morgan, J. Hall, and M. M. Sondhi, "Synthesized stereo combined with acoustic echo cancellation for desktop conferencing," *Bell Labs Tech. J.*, vol. 3, no. 3, pp. 148–158, July-Sept. 1998.

[2] M. M. Sondhi and D. R. Morgan, "Acoustic echo cancellation for stereophonic teleconferencing," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio Acoustics*, 1991.

[3] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation — An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[4] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[5] T. Gänsler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1998, pp. 3649–3652.

[6] A. Gilloire and V. Turbin, "Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers," in *Proc. IEEE ICASSP*, 1998, pp. 3681–3684.

[7] S. Shimauchi, Y. Haneda, S. Makino, and Y Kaneda, "New configuration for a stereo echo canceller with nonlinear pre-processing," in *Proc. IEEE ICASSP*, 1998, pp. 3685–3688.

[8] M. Ali, "Stereophonic echo cancellation system using time-varying all-pass filtering for signal decorrelation," in *Proc. IEEE ICASSP*, 1998, pp. 3689–3692.

[9] S. Haykin, *Adaptive Filter Theory*, Prentice Hall International, 1996.

[10] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. on Speech Audio Processing*, submitted.

[11] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to* MPEG-*2*, chapter 4, pp. 55–79, Digital Multimedia Standards Series. Chapman & Hall, 1997.

[12] S. A. Ramprashad, "A multimode transform predictive coder (MTPC) for speech and audio," in *IEEE Speech Coding Workshop*, June 1999.

[13] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.

[14] K. Ochiai, T. Araseki, and T. Ogihara, "Echo cancellation with two path models," *IEEE Trans. on Commun.*, vol. COM-25, no. 6, pp. 589–595, June 1977.

[15] M. G. Bellanger, *Adaptive Digital Filters and Signal Analysis*, Marcel Dekker, 1987.

[16] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[17] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall PTR, 1995.

[18] G. Strang and T. Nguyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1996.

[19] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall PTR, 1993.

[20] N. J. Fliege, *Multirate Digital Signal Processing*, John Wiley & Sons, 1994.

[21] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. of ICASSP*, 1988, pp. 2570–2573.

[22] H. Sorensen, D. Jones, M. Heideman, and S. Burrus, "Real-values fast Fourier transform algorithms," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, pp. 849–863, June 1987.

[23] D. A. Berkley and J. L. Flanagan, "HuMaNet: an experimental human-machine communications network based on ISDN wideband audio," *AT&T Tech. J.*, vol. 69, pp. 87–99, Sept./Oct. 1990.

[24] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms.," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, pp. 70–73, June 1967.

[25] Fraunhofer IIS, "MPEG-1 LAYER III shareware audio coder," 1995, Am Weichselgarten 3 D-91058 Erlangen Germany, encoder and decoder code: http://www.iis.fhg.de/amm/techinf/layer3/index.html, Public domain decoder source code (ANSI c): ftp://ftp.fhg.de/pub/iis/layer3/public_c/.

[26] E. J. Diethorn, "An algorithm for subband echo suppression in speech communications," Private Communication, 1998.

[27] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "An implementation of a stereophonic acoustic echo canceler on a general purpose DSP," in *Proc. ICSPAT*, 1999.

[28] T. F. Coleman and Y. Li, "A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM J. on Optimization*, vol. 6, no. 4, pp. 1040–1058, 1996.

[29] R. Boite and H. Leich, "A new procedure for the design of high order minimum phase FIR digital or CCD filters," *Signal Processing*, pp. 101–108, 1981.

[30] G. Wackersreuther, "On the design of filters for ideal QMF and polyphase filter banks," *AEÜ*, vol. 39, no. 2, pp. 123–130, 1985.

[31] Texas Instruments, *TMS320C4x General-Purpose Applications User's Guide*, chapter 6, pp. 56–86, Texas Instruments, Mar. 1996.

Paper II

Paper II

Comparison of Different Adaptive Algorithms for Stereophonic Acoustic Echo Cancellation

Abstract

In this paper, different adaptive algorithms for stereophonic acoustic echo cancellation are compared. The algorithms include the simple LMS algorithm and two specialized two-channel adaptive algorithms. Due to the high calculation complexity needed in stereophonic acoustic echo cancellation applications, the time domain algorithms are applied in a subband structure. The comparison include aspects as convergence rate, calculation complexity, signal path delay and memory usage. Real-life recordings are used in the evaluation.

Based on: P. Eneroth, J. Benesty, T. Gänsler, and S. Gay, "Comparison of Different Adaptive Algorithms for Stereophonic Acoustic Echo Cancellation," *European Signal Processing Conference*, September 2000.

This work was partly done at Bell Labs, Lucent Technologies.

1 Introduction

The increasing use of teleconferencing systems and desktop conferencing where the acoustic echo canceler (AEC) plays a central role, has led to the requirement of faster and better performing algorithms. In these applications there is a desire to have far better sound quality and sound localization than what has been provided previously. These quality improvements can be achieved by increasing the signal bandwidth and also by adding more audio channels to the system. This last fact spurred the need for multi-channel acoustic echo cancelers of which the two channel (stereo) AEC is the the most interesting since only complexity issues differ in the more general multi-channel case. Figure 1 illustrates the concept of stereophonic echo cancellation between a transmission room and a receiving room. As is depicted in the figure, the echo is due to the acoustic coupling between the loud-speakers and the microphones in the receiving room. The solution is to estimate this coupling, and subtract an estimated echo from the return signal.

Stereophonic acoustic echo cancellation (SAEC) is fundamentally different from traditional mono echo cancellation. Four mono echo cancelers 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 is described in [1]. The fundamental problem is that the two channels may carry linearly related signals which in turn may 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. As a result, intensive studies have been made of how to handle this properly.

A complete theory of non-uniqueness and characterization of the SAEC solution was presented in [2]. It was shown that the only solution to the non-uniqueness problem is to reduce the correlation between the stereo signals and an efficient low complexity method for this purpose was also given, [2]. Several decorrelation methods has since then been presented [3], [4], [5], and even if these decrease the correlation between the channels, the normal equations to be solved will still represent an ill-conditioned problem.

The performance of the SAEC is also more severely affected by the choice of algorithm than the monophonic counterpart. This is easily recognized since the performance of most adaptive algorithms depends on the condition number of the input

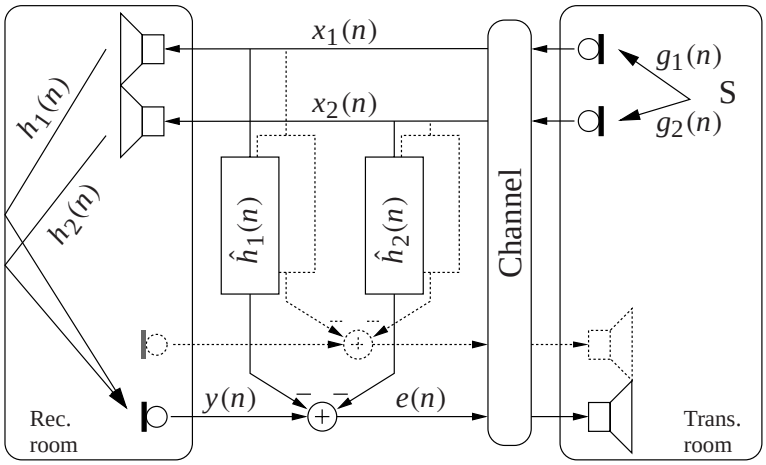

Figure 1: Schematic diagram of a stereophonic echo canceler.

signal. In the stereo case, the condition number is very high and algorithms that do not take the cross-correlation between the channels into account, such as the standard Least Mean Square (LMS) or Normalized LMS (NLMS), converge slowly to the true solution.

More sophisticated algorithms such as the APA (Affine Projection Algorithm) or RLS (Recursive Least Squares), that are less affected by a high condition number, handles the stereo case much better. This is even more true for algorithms that are specially derived for the two-channel situation. In the following we will study two different types of two-channel algorithms, the two-channel fast RLS algorithm, and a frequency domain adaptive algorithm.

2 Adaptive Algorithms

A few adaptive algorithms, which exploit the correlation between the two channels in SAEC, have been derived. In this section, we will investigate important properties of such algorithms. The algorithms that have been chosen are the two-channel fast recursive least mean squares (FRLS), Table 1 [6], applied in a subband scheme and the unconstrained versions of a two-channel frequency domain adaptive filter, Table 2 [7]. For comparison, the normalized least mean square (NLMS), is also studied. In Table 3 calculation complexity, signal path delay and memory usage of these algorithms are exemplified, and finally in Sec. 3, convergence and tracking properties are shown in simulations. All complexity and memory usage numbers are calculated for the full

SAEC, i.e. with *four* adaptive filters and *two* return signals, as depicted in Fig. 1.

2.1 Normalized Least Mean Square

The NLMS is without competition the most common adaptive filter, and its strengths include robust behavior, and its structure allows for simple implementation. The error signal and the filter updates are calculated as

$$e(n) = y(n) - \mathbf{x}_1^H \hat{\mathbf{h}}_1(n-1) - \mathbf{x}_2^H \hat{\mathbf{h}}_2(n-1), \tag{1}$$

$$\hat{\mathbf{h}}_i(n) = \hat{\mathbf{h}}_i(n-1) + \frac{\mu}{\mathbf{x}_1^H \mathbf{x}_1 + \mathbf{x}_2^H \mathbf{x}_2} \mathbf{x}_i e(n), \quad i = 1, 2 \tag{2}$$

where $\mathbf{x}_i$ is the transmission room signal vector containing the last L samples for channel i, and $\hat{\mathbf{h}}_i(n)$ the filter estimate vector. The main disadvantage with this algorithm is the slow convergence rate on signals with large correlation matrix eigenvalue spread [8], as speech signals, and with correlated channels in the two channel situation. For the long adaptive filters needed in SAEC, calculation complexity is also significant. The number of real-valued multiplications per sample needed for the four adaptive filters depicted in Fig. 1 is $8L$ for real-valued signals and $32L$ for complex-valued signals. The calculation complexity can be reduced by applying the adaptive filters in a subband structure, but this will also introduce signal transmission delay to the otherwise delayless NLMS algorithm. This is described in Sec. 2.4.

2.2 Two-channel Fast Recursive Least Mean Squares

A complete analysis of the two-channel FRLS algorithm, Table 1, is beyond the scope of this paper. However, a general analysis of the RLS algorithm can be found in [8] and stabilized two-channel versions are described in [6], [9], [10].

In contrast to the NLMS algorithm, estimates of the signal correlation and the cross-correlation between the two channels are incorporated in the Kalman gain vector $\mathbf{G}$, and used in the update of the adaptive filter $\hat{\mathbf{h}}$. This improves the convergence rate significantly over the NLMS algorithm, but it is also the cause of the well-known instability problem of RLS type algorithms.

In this version, stability is improved with the stability parameter k. But for operation on non-stationary signals, like speech-signals, further enhancements are needed. First of all, by monitoring φ, 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 a suitable start value, while the adaptive filter estimate, $\hat{\mathbf{h}}$, can be left unchanged. Secondly, the algorithm can be applied in a two-path structure [11], where the adaptive filter $\hat{\mathbf{h}}$ is

Table 1: Two-channel FRLS algorithm for complex arithmetic. The transposition and the Hermitian transposition operators are denoted ^{T} and ^{H}, respectively, and the conjugate is denoted *. $\lambda \in (0, 1]$ is the forgetting factor and $k \in [1, 2.5]$ a stabilization parameter.

Input signals	**Matrix sizes**
$\boldsymbol{\chi}(n) = \begin{bmatrix} x_1(n) & x_2(n) \end{bmatrix}^T$	(2×1)
$\mathbf{x}(n) = \begin{bmatrix} \boldsymbol{\chi}^T(n) & \dots & \boldsymbol{\chi}^T(n-L+1) \end{bmatrix}^T$	$(2L \times 1)$
Prediction	
$\mathbf{e}_\mathrm{A}(n) = \boldsymbol{\chi}(n) - \mathbf{A}^H(n-1)\mathbf{x}(n-1)$	(2×1)
$\varphi_1(n) = \varphi(n-1) + \mathbf{e}_\mathrm{A}^H(n)\mathbf{E}_\mathrm{A}^{-1}(n-1)\mathbf{e}_\mathrm{A}(n)$	(1×1)
$\begin{bmatrix} \mathbf{M}(n) \\ \mathbf{m}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{G}(n-1) \end{bmatrix} + \begin{bmatrix} \mathbf{I}_2 \\ -\mathbf{A}(n-1) \end{bmatrix} \mathbf{E}_\mathrm{A}^{-1}(n-1)\mathbf{e}_\mathrm{A}(n)$	$((2L+2) \times 1)$
$\mathbf{e}_{\mathrm{B}_2}(n) = \boldsymbol{\chi}(n-L) - \mathbf{B}^H(n-1)\mathbf{x}(n)$	(2×1)
$\varphi(n) = \varphi_1(n) - \mathbf{e}_{\mathrm{B}_2}^H(n)\mathbf{m}(n)$	(1×1)
$\mathbf{A}(n) = \mathbf{A}(n-1) + \mathbf{G}(n-1)\mathbf{e}_\mathrm{A}^H(n)/\varphi(n-1)$	$(2L \times 2)$
$\mathbf{E}_\mathrm{A}(n) = \lambda[\mathbf{E}_\mathrm{A}(n-1) + \mathbf{e}_\mathrm{A}(n)\mathbf{e}_\mathrm{A}^H(n)/\varphi(n-1)]$	(2×2)
$\mathbf{G}(n) = \mathbf{M}(n) + \mathbf{B}(n-1)\mathbf{m}(n)$	$(2L \times 1)$
$\mathbf{e}_{\mathrm{B}_1}(n) = \mathbf{E}_\mathrm{B}(n-1)\mathbf{m}(n)$	(2×1)
$\mathbf{e}_\mathrm{B}(n) = k\mathbf{e}_{\mathrm{B}_2}(n) + (1-k)\mathbf{e}_{\mathrm{B}_1}(n)$	(2×1)
$\mathbf{B}(n) = \mathbf{B}(n-1) + \mathbf{G}(n)\mathbf{e}_\mathrm{B}^H(n)/\varphi(n)$	$(2L \times 2)$
$\mathbf{E}_\mathrm{B}(n) = \lambda[\mathbf{E}_\mathrm{B}(n-1) + \mathbf{e}_{\mathrm{B}_2}(n)\mathbf{e}_{\mathrm{B}_2}^H(n)/\varphi(n)]$	(2×2)
Filtering	
$e(n) = y(n) - \hat{\mathbf{h}}^H(n-1)\mathbf{x}(n)$	(1×1)
$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \mathbf{G}(n)e^*(n)/\varphi(n)$	$(2L \times 1)$
Definition	
$\hat{\mathbf{h}}(n) = \begin{bmatrix} \hat{h}_{1,0}(n) & \hat{h}_{2,0}(n) & \cdots & \hat{h}_{1,L-1}(n) & \hat{h}_{2,L-1}(n) \end{bmatrix}^T$	

copied to a static filter, if it is a better estimate than a previous copy. The static filter is then used for the actual filtering.

Another disadvantage with the two-channel FRLS in most situations is the high calculation complexity. The number of real-valued multiplications per sample needed for the four adaptive filters depicted in Fig. 1 is $32L$ and $128L$ for complex-valued signals. Most likely this algorithm will be used in subband structures, in order to decrease the complexity, but also to improve stability by having shorter adaptive filters. Like for the NLMS, this will introduce signal transmission delay.

2.3 Two-channel Frequency Domain Algorithm

Frequency domain algorithms perform better than the NLMS on signals with large correlation matrix eigenvalue spread, like speech signals. In the adaptive filter update of the NLMS, (2), the same normalization factor, $\mathbf{x}_1^H\mathbf{x}_1+\mathbf{x}_2^H\mathbf{x}_2$, is used for all frequencies, independent of the spectral characteristic of the excitation signals. In the two-channel frequency domain filter, Table 2, each frequency bin has an independent normalization factor. In the update of $\hat{\underline{\mathbf{h}}}_{1,k}$, in Table 2, this normalization is represented by the spectrum estimate vector $\mathbf{S}_1^{-1}$. In the same equation, the channel cross-correlation is taken into account by the term $\rho\mathbf{S}_{1,2}\tilde{\mathbf{S}}_{2,2}^{-1}\mathbf{X}_2^*(m-k)$, where the parameter ρ controls the amount of cross-correlation in the adaptation. This improves the performance when the two channels x_1 and x_2 are correlated. $\tilde{\mathbf{S}}_{i,i}$, $i=1,2$, are regularized versions of $\mathbf{S}_{i,i}$, since this greatly decreases the misalignment in poorly excited frequency bands.

The algorithm given in Table 2 is, except for the regularization, an exact derivative of a time-domain block LMS, and is denoted the constrained version. Large complexity reduction is possible by replacing $\mathbf{F}\mathbf{W}_2\mathbf{F}^{-1}$ with the identity matrix, without significant reduction of the performance [7], and this version is denoted unconstrained. For the unconstrained version, the optimal step-size parameter is $\mu_b = 2(1-\beta)$. In the algorithm in Table 2, the input data blocks have no overlap ($\alpha_o = 1$). By overlapping the input data, it is possible to increase the convergence rate, and in the simulation section, we use only 25% new input-data for each iteration ($\alpha_o = 4$). It is also possible to have several adaptive filter taps per frequency bin, however in this paper, we will only consider the situation with 1 tap per frequency bin, i.e. $K = 1$.

For $\alpha_o = 1$, the unconstrained version of the algorithm in Table 2, with two return signals (e_1,e_2), requires 3 FFTs on vectors with $2L$ elements plus $148L$ real-valued multiplications and $8L$ real-valued divisions per L samples. The fact that 2 real-valued FFTs can be calculated with one complex-valued FFT, have been used. For the overlapped version, the complexity is approximately increased with the factor α_o. The signal transmission delay is equal to the block size L. In contrast to the FRLS, this algorithm is very robust. A complete analysis of the algorithm is given in [7].

Table 2: Two-channel frequency domain adaptive filter, i denotes channel 1 or 2. The block size is denoted L, and the number of adaptive coefficients per frequency bin K, resulting in a total adaptive filter length of KL. Forgetting factor $0 < \beta < 1$, channel cross-correlation dependence $0 \leq \rho \leq 1$ and adaptive step size μ_b. $\mathbf{F}$ denotes the Fourier matrix.

Input signals	**Matrix sizes**
$\mathbf{X}_i(m) = \text{diag}\left\{\mathbf{F}\left[x_i(mL-L) \quad \cdots \quad x_i(mL+L-1)\right]^T\right\}$	$(2L \times 2L)$
$\mathbf{y}(m) = \left[\mathbf{0}_{1\times L} \quad y(mL) \quad \cdots \quad y(mL+L-1)\right]^T$	$(2L \times 1)$
Power spectrum estimation with regularization	
$\mathbf{S}_{i,j}(m) = \beta\mathbf{S}_{i,j}(m-1) + (1-\beta)\mathbf{X}_i^*(m)\mathbf{X}_j(m)$	$(2L \times 2L)$
$\tilde{\mathbf{S}}_{i,i}(m) = \mathbf{S}_{i,i}(m) + \text{diag}\left\{\mathbf{e}_{\text{reg}}\right\}$	$(2L \times 2L)$
$\mathbf{S}_i(m) = \tilde{\mathbf{S}}_{i,i}(m)\left[\mathbf{I}_{2L\times 2L} - \rho^2\mathbf{S}_{1,2}^*(m)\mathbf{S}_{1,2}(m)\left\{\tilde{\mathbf{S}}_{1,1}(m)\tilde{\mathbf{S}}_{2,2}(m)\right\}^{-1}\right]$	$(2L \times 2L)$
Filtering	
$\underline{\hat{\mathbf{y}}}_i'(m) = \sum_{k=0}^{K-1}\mathbf{X}_i(m-k)\underline{\hat{\mathbf{h}}}_{i,k}(m-1)$	$(2L \times 1)$
$\mathbf{e}(m) = \mathbf{y}(m) - \mathbf{W}_1\mathbf{F}^{-1}\left[\underline{\hat{\mathbf{y}}}_1'(m) + \underline{\hat{\mathbf{y}}}_2'(m)\right]$	$(2L \times 1)$
$\underline{\hat{\mathbf{h}}}_{1,k}(m) = \underline{\hat{\mathbf{h}}}_{1,k}(m-1) + \mu_b\mathbf{F}\mathbf{W}_2\mathbf{F}^{-1}\mathbf{S}_1^{-1}\cdot \left[\mathbf{X}_1^*(m-k) - \rho\mathbf{S}_{1,2}\tilde{\mathbf{S}}_{2,2}^{-1}\mathbf{X}_2^*(m-k)\right]\mathbf{F}\mathbf{e}(m)$	$(2L \times 1)$
$\underline{\hat{\mathbf{h}}}_{2,k}(m) = \underline{\hat{\mathbf{h}}}_{2,k}(m-1) + \mu_b\mathbf{F}\mathbf{W}_2\mathbf{F}^{-1}\mathbf{S}_2^{-1}\cdot \left[\mathbf{X}_2^*(m-k) - \rho\mathbf{S}_{2,1}\tilde{\mathbf{S}}_{1,1}^{-1}\mathbf{X}_1^*(m-k)\right]\mathbf{F}\mathbf{e}(m)$	$(2L \times 1)$
Definitions	
$\mathbf{e}(m) = \left[\mathbf{0}_{1\times L} \quad e(mL) \quad \cdots \quad e(mL+L-1)\right]^T$	$(2L \times 1)$
$\underline{\hat{\mathbf{h}}}_{i,k}(m) = \mathbf{F}\left[\hat{\mathbf{h}}_{i,k}^T(m) \quad \mathbf{0}_{1\times L}\right]^T$	$(2L \times 1)$
$\hat{\mathbf{h}}_{i,k}(m) = \left[\hat{h}_{i,kL}(m) \quad \cdots \quad \hat{h}_{i,kL+L-1}(m)\right]^T$	$(L \times 1)$
$\mathbf{W}_1 = \text{diag}\left\{\left[\mathbf{0}_{1\times L} \quad \mathbf{1}_{1\times L}\right]\right\}$, $\mathbf{W}_2 = \text{diag}\left\{\left[\mathbf{1}_{1\times L} \quad \mathbf{0}_{1\times L}\right]\right\}$	

Table 3: Calculation complexity as the number of real-valued multiplication per sample, the transmission delay in ms and the memory usage in words, for the algorithms used in the Simulation section.

	NLMS	FRLS	NLMS in subbands	FRLS in subbands	Frequency LMS
complex.	8.2k	33k	0.72k	2.6k	1.0k
delay (ms)	0	0	64	64	64
memory	6.1k	18k	17k	40k	66k

2.4 Adaptive Filtering in Filterbank Structures

In subband structures, an analysis filterbank divides a signal into M down-sampled subband signals, which each represent a frequency region. Then the adaptive filtering is performed on the subband signals, and the fullband residual echo signal $e(n)$ is finally reconstructed with a synthesis filterbank. Since down-sampling aliasing has a very negative effect on the convergence of the adaptive filter [12], the downsampling factor r is usually less than M. For FFT based filterbanks we will only need to have $M/2+1$ complex-valued adaptive filters since the $M/2-1$ upper subbands differ only by the conjugate from the lower counterpart. The calculation complexity reduction comes from the fact that each adaptive filter is r times shorter and only updated once for each r fullband sample.

The biggest disadvantage with a filterbank is the introduced signal path delay. For linear phase filterbanks, this delay is equal to the length of the prototype filter, which depends on several factors, including: aliasing, attenuation, and the ratio r/M. For 60 dB stopband suppression and $r/M = 0.75$ we need approximately $11M$ filter coefficients. In addition to this we need approximately 5 non-causal taps per subband [13]. If the the inherent flat delay in the receiving room is less, we can artificially delay the signal y.

Four analysis filterbanks, decomposing x_1, x_2, y_1, y_2 in Fig. 1, and two analysis filterbanks, e_1, e_2, are needed. The number of real-valued multiplication for the four analysis filterbanks is $\frac{1}{r}(4K + 4M\log_2 M - 14M + 24)$ per sample, and $\frac{1}{r}(2K+2M\log_2 M-7M+12)$ for the two synthesis filterbanks, where K is the filter length [10].

3 Simulations

The convergence rate and the ability of the algorithm to track echo path changes highly effects the performance of SAECs. In this section we will exemplify these properties

with real-life data recorded in a quiet office-like room. Both the recordings and the simulations are performed with 16 kHz sampling rate. In order to reduce the correlation between the two channels, the data was processed with a non-linear function, [2] $\alpha_n = 0.5$. As performance index the mean square error (MSE) energy of the residual is used. The MSE is given by,

$$\text{MSE} = \frac{P_e}{P_y}, \qquad P_e = \text{LPF}\Big[e^2(n)\Big], \tag{3}$$

where LPF denotes a lowpass filter; in this case it has a single pole in 0.9996. Four different systems are used in the simulation. The classical NLMS algorithm, with each adaptive filter having a length of 1024 taps. Both the NLMS and the two-channel FRLS algorithms in a subband structure, with 64 subbands and a downsample factor of 48. Each adaptive filter have 28 taps, corresponding to 1024 fullband taps plus 320 non-causal adaptive filter taps. Finally, the frequency based algorithm has a block-size $L = 1024$, $K = 1$ filter taps per frequency bin and the input-data overlap factor $\alpha_o = 4$.

In Fig. 2 the initial convergence and the tracking ability of the algorithms are shown. There is an instantaneously change of *the receiving room impulse responses*, $h_i(t)$, after 20 s. Both the FRLS in subbands and the frequency domain algorithm perform very well. In Fig. 3 there is an instantaneously change of *the transmission room impulse responses*, $g_i(t)$, after 5.1 s. As can be seen in the figures, the two algorithms specially derived for the two-channel situation are less affected than the other two.

Finally in Fig. 4, it is shown that even if the fullband NLMS appear to perform reasonably well in Fig. 2 and 3, it does only perform well in the region with the most signal energy, 200–2500 Hz. As the human ear is sensitive also for higher frequencies, Fig. 4 tells us more about the perceptual quality than previous figures.

In order to study the stability of the FRLS algorithm, 24 hours of real-life data was processed. On average each subband was restarted every 18 s. These restarts are handled with little effect on the performance. In Fig. 2 there are 33 restarts, on average one per subband, and 13 of them occur in the time-interval 16–19 s.

4 Conclusions

From the simulations it can be concluded that the two-channel FRLS and the two-channel frequency domain based echo canceler has significantly faster convergence rate than a NLMS based echo canceler. As is depicted in Fig. 3, the two two-channel algorithms are also less affected by the fundamental problem in stereophonic acoustic echo cancellation, i.e., that the two channels are highly correlated.

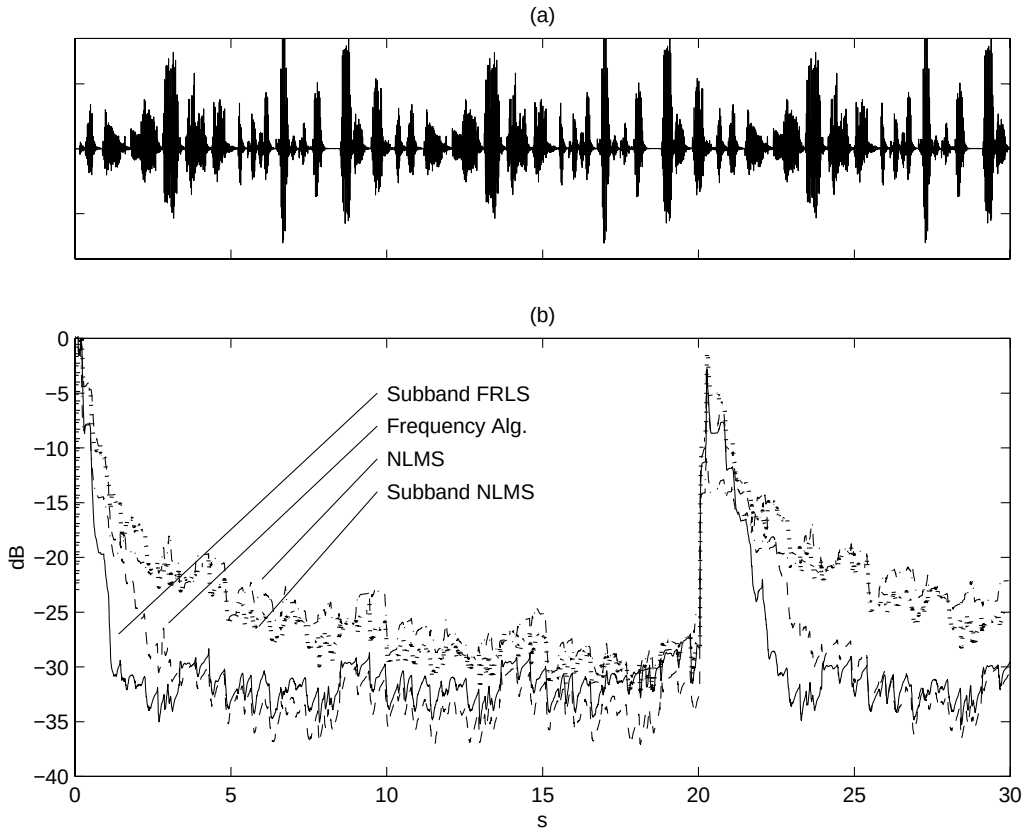

Figure 2: (a) Transmission room speech signal (left channel). Instantaneous receiving room impulse responses, $h_i(t)$, change after 20 s. (b) Mean square error performance for four adaptive algorithms: NLMS (dash-dotted line), subband NLMS (dotted line), two-channel FRLS (solid line), two-channel frequency algorithm (dashed line).

Table 3 illustrates that the calculation complexity of the fullband versions of both the NLMS and the two-channel FRLS algorithm is significantly higher than the other systems evaluated, and that the calculation complexity can be reduced by more than 10 times by using a 64 band filterbank. The disadvantage is the signal transmission delay imposed with a filterbank based echo canceler. The system designer can though make the compromise between calculation complexity reduction and signal transmission delay by altering the number of subbands.

Finally, it should be remembered that two-channel FRLS needs to be stabilized. As pointed out in the paper, the algorithm may need frequent restarts in order achieve system stability. The frequency-domain algorithm on the other hand be guaranteed to be stabile [14].

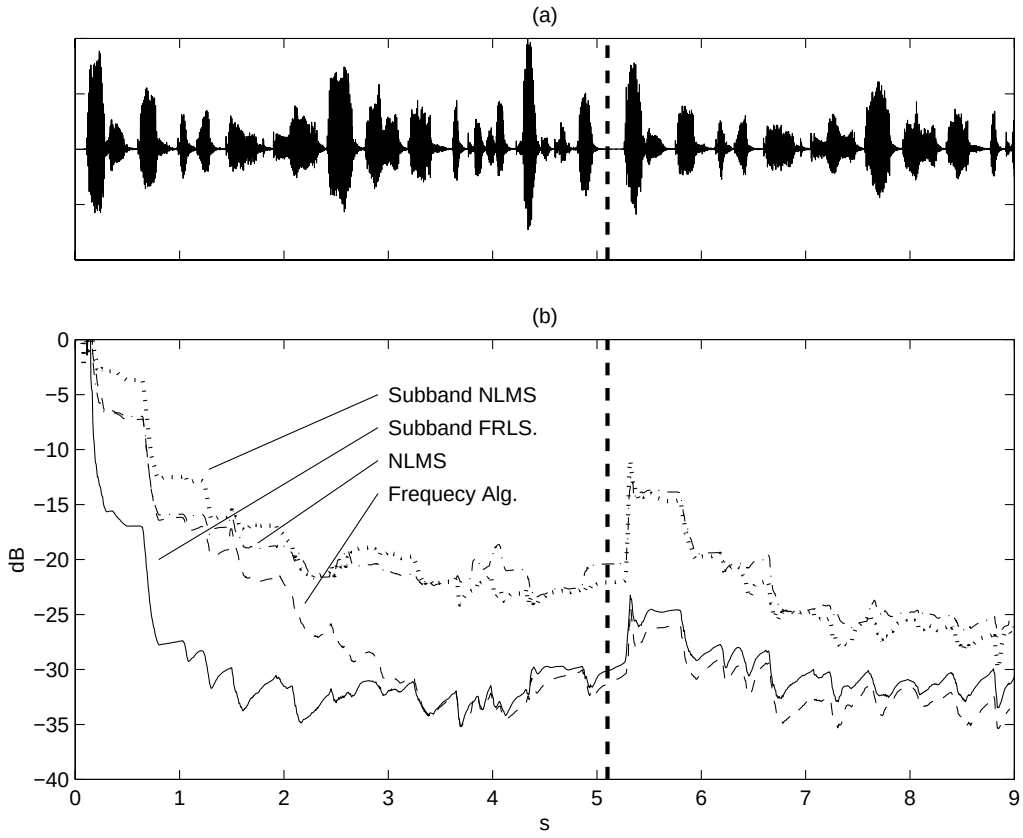

Figure 3: Instantaneous transmission room impulse response, $g_i(t)$, change after 5.1 s. Other conditions as in Fig. 2.

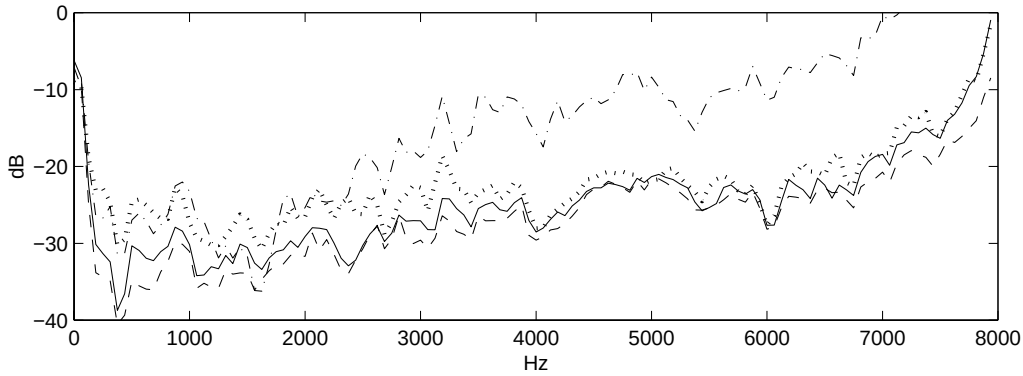

Figure 4: Echo canceler suppression in the frequency domain. Line types as in Fig. 2.

References

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation — An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[2] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[3] T. Gänsler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1998, pp. 3649–3652.

[4] A. Gilloire and V. Turbin, "Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers," in *Proc. IEEE ICASSP*, 1998, pp. 3681–3684.

[5] S. Shimauchi, Y. Haneda, S. Makino, and Y Kaneda, "New configuration for a stereo echo canceller with nonlinear pre-processing," in *Proc. IEEE ICASSP*, 1998, pp. 3685–3688.

[6] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.

[7] J. Benesty, A. Gilloire, and Y. Grenier, "A frequency domain stereophonic acoustic echo canceler exploiting the coherence between the channels," *J. Acoust. Soc. Am.*, vol. 106, pp. L30–L35, Sept. 1999.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice Hall International, 1996.

[9] M. G. Bellanger, *Adaptive Digital Filters and Signal Analysis*, Marcel Dekker, 1987.

[10] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A real-time stereophonic acoustic echo canceler," in *Acoustic Signal Processing For Telecommunication*, S. L. Gay and J. Benesty, Eds., vol. 551, chapter 8, pp. 135–152. Kluwer Academic Publishers, 2000, ISBN 0-7923-7814-8.

[11] K. Ochiai, T. Araseki, and T. Ogihara, "Echo cancellation with two path models," *IEEE Trans. on Commun.*, vol. COM-25, no. 6, pp. 589–595, June 1977.

[12] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[13] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. of ICASSP*, 1988, pp. 2570–2573.

[14] D. Mansour and A. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 30, no. 5, pp. 726–734, Oct. 1982.

Paper III

Paper III

Analysis of Subband Impulse Responses in Subband Echo Cancelers

Abstract

Several types of subband echo cancelers have been presented and used in acoustic applications such as hands-free mobile communication and video conferencing. In these structures, adaptive filtering is performed in the subband domain, and in this paper we analyze the resulting subband impulse responses. It has previously been shown that a causal fullband impulse response is best modeled by non-causal subband impulse responses. In this paper, we analyze the rationale of the non-causality in various situations: ideal infinitely long filters as well as realizable finite filters. The theoretical results are also exemplified with simulations. To compensate for non-causal subband filter taps, a signal path delay can be introduced in the signal. Formulas for calculating the minimum mean square error for the optimal subband impulse response as a function of this introduced signal path delay are presented. This way, a suitable value of the signal path delay can be determined.

Based on: P. Eneroth, and T. Gänsler, "Analysis of Subband Impulse Responses in Subband Echo Cancelers," *IEEE Trans. on Signal Processing*, submitted 2000.

1 Introduction

Adaptive algorithms are today in use in numerous applications. One of the most used algorithm in practise is the least mean square (LMS) adaptive algorithm. It has several advantages, such as robustness, a very simple structure, and, in most applications, the convergence properties of the LMS algorithm is sufficient. In some applications, such as acoustic echo cancellation, where the number of filter taps to estimate is large, up to several thousands, the computational complexity of the LMS algorithm is high. Also, the LMS algorithm converges slowly on signals with a large correlation matrix eigenvalue spread [1], as for example with speech signals. These drawbacks can be reduced by dividing the input and the desired signal into subband signals, where each subband corresponds to one mostly non-overlapping frequency interval. The subband signals are downsampled and since adaptation is performed at a lower rate on the subband signals, the computational complexity can be reduced. As an adaptive subband structure is parallel by nature, implementation on high-performance parallel digital signal processing systems should be efficient. The convergence rate is improved because the eigenvalue spread in the *effective* frequency range of the subbands is reduced [2], where "effective" is the frequency range used in the estimation of the subband impulse response. The major disadvantages of subband structures include aliasing due to downsampling [3], the transmission delay introduced in the signal path, and the necessity to model a few non-causal impulse response taps [4], as discussed in this article.

For the most commonly used subband echo canceler structure, illustrated in Fig. 1, the echo cancellation takes place in the subband domain. The residual echo signals, $e_m(k)$, are then used to reconstruct the fullband signal by the synthesis filterbank. If the filterbanks are not properly designed, aliasing can be a major source of distortion. Even though perfect reconstruction synthesis filterbanks cancel aliasing, aliasing still exists in the subband signals, and this will significantly decrease the performance of the adaptive filters, as shown in [3]. A second way to decrease distortion caused by aliasing is to use filterbanks with non-critical downsampling, where it is possible to suppress aliasing sufficiently by designing filters with high stopband attenuation. This is the preferred method in echo cancellation; thus the analysis in this paper will consider filterbanks with non-critical downsampling.

An alternative subband structure, which does not introduce extra delay to the signal path, has been introduced in [5]. Like the structure described above, the adaptive filters estimate the impulse response in subbands. Instead of performing the compensation in the subbands, the fullband impulse response estimate is reconstructed from the estimated subband impulse responses. Using this reconstructed fullband impulse response, echo cancellation is performed in the time domain without introducing extra signal path delay in the near-end signal. The price payed for zero signal path delay is increased calculation complexity, though it is still lower than for the fullband LMS

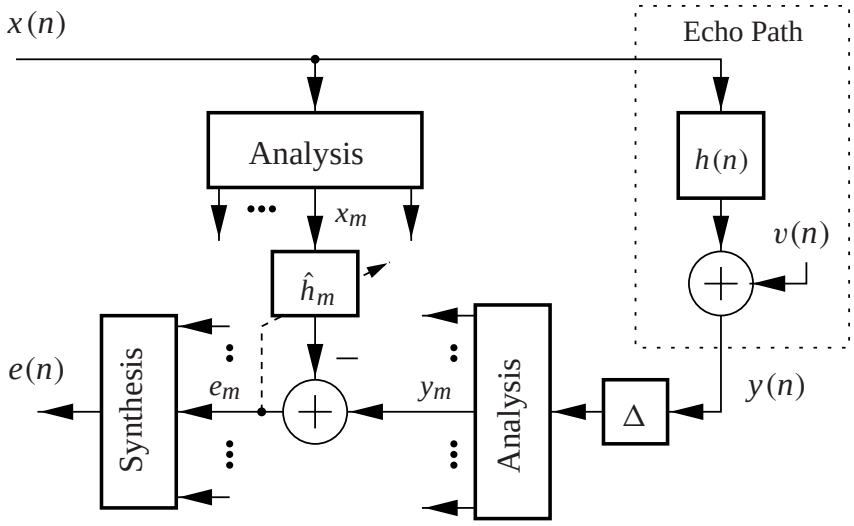

Figure 1: Subband echo canceler and model of acoustic echo path. In each subband, an adaptive algorithm estimates the subband impulse response $\hat{h}_m$, and the subband residual echo signal $e_m(k)$ is generated.

case [5].

Adaptive filtering is a way of solving a minimization problem, and in subband structures we can usually choose between two minimization problems. In echo cancellation, the optimum solution is to minimize the mean square error (MSE) of the fullband error signal, $\mathrm{E}\{|e(n)|^2\}$, and structures minimizing this signal are called *closed-loop*. Closed-loop structures permit the adaptive filters to partly compensate for errors introduced by the non-ideal analysis and synthesis filterbanks [5]. The other method, *open-loop*, is a sub-optimal structure, where the MSE of every subband error signal, $e_m(k)$, is minimized independently of the other subbands. The latter structure is commonly used, since the delayed error signal in the closed-loop structure reduces the convergence rate [6], [7].

In this paper, we analyze the subband impulse responses for open-loop subband structures, i.e., the error in each subband is minimized separately. First, the true subband impulse response in an ideal system is derived. It is shown that, due to the bandpass characteristics of the subband signal, the ideal subband impulse response is non-causal and of infinite length. Then the Wiener solution for the subband impulse response is derived in the frequency domain. Due to aliasing, this solution has a very large variance close to the band edges. Next the optimal finite length MSE solution is calculated. In subband systems with non-critical downsampling, this solution can be quite different from the true subband impulse response since the subband signal

energy in some frequency regions are near zero. Finally, we investigate convergence properties of the LMS algorithm on non-critically downsampled subband signals. In contrast to the optimal MSE impulse response, the LMS estimates in the frequency regions with near zero signal energy will remain near the initial value [8]. Therefore the LMS estimate will have bandpass characteristics similar to the true impulse response in the ideal system. All these analyzes will show that *causal* fullband impulse responses are best modeled with *non-causal* subband impulse responses. Therefore, when implementing echo cancelers in subband schemes, it is important to be able to model a few non-causal subband impulse response taps. This can be done by introducing a delay Δ in the near-end signal, Fig. 1. Also presented in this paper are formulas for calculation of the minimum mean square error as a function of the introduced delay Δ for a given fullband impulse response and analysis filterbank. By using these formulas, the optimum delay Δ for a given system can be calculated. The non-causality has earlier been investigated by [4], and the introduction of a near-end signal delay Δ to compensate for this model error has been used in [4], [9], [10]. The main contributions of this article are the analysis of the problem in different system situations, especially for realizable situations, and the formulas presented, by which the optimum delay Δ in the optimal MSE solution can be determined. The influence that the LMS class of adaptive filters have on this problem is also of interest.

2 System Description

In order to derive an expression for the subband representation of a fullband impulse response and the minimum mean square error, we will analyze the subbands separately. In Fig. 2, a model of the signal path for subband number m is presented. The fullband impulse response to be estimated is denoted $h(n)$ and the estimated subband impulse response by $\hat{h}_m(k)$. For analysis purpose, the far-end signal, $x(n)$, is assumed to be a white Gaussian noise signal. The near-end signal is represented by a filtered version of $x(n)$ with the addition of noise,

$$y(n) = \sum_{i=0}^{N-1} h(i)x(n-i) + v(n), \tag{1}$$

where N is the length of the fullband impulse response $h(n)$. The near-end and background noise signal, $v(n)$, is assumed to be independent of the far-end signal, $x(n)$, and to be of zero mean. The far-end signal is delayed Δ samples, in order to artificially increase the flat delay of the system impulse response $h(n)$. Both the far-end and the near-end signals are then bandpass filtered by the analysis filterbank, $a_m(n)$, and downsampled by a factor r times. Downsampling is defined as $x_m(k) = \tilde{x}_m(rk)$, where $\tilde{x}_m(rk)$ is the subband signal before downsampling, and the downsampled sub-

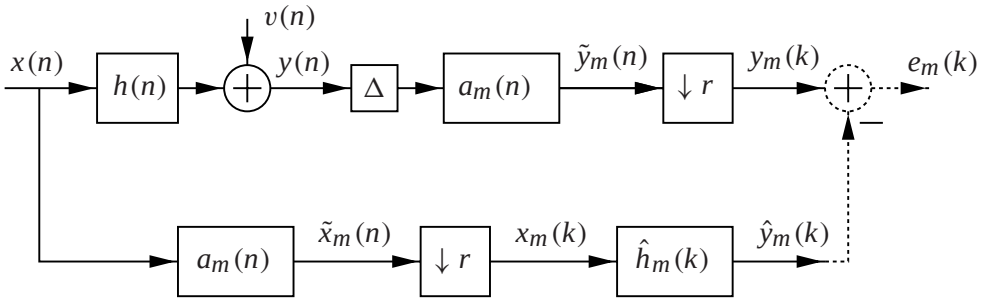

Figure 2: Signal path model for subband m. The estimated subband impulse response $\hat{h}_m(k)$ should minimize the subband residual signal $e_m(k)$.

band signals can be expressed as,

$$x_m(k) = \sum_{i=0}^{L-1} a_m(i)x(rk-i), \tag{2}$$

$$\begin{aligned} y_m(k) &= \sum_{j=0}^{L-1} a_m(j) \sum_{i=0}^{N-1} h(i)x(rk-\Delta-i-j) \\ &\quad + \sum_{q=0}^{L-1} a_m(q)v(rk-\Delta-q), \\ &= \sum_{i=0}^{N-1} \sum_{p=i}^{L-1+i} a_m(p-i)h(i)x(rk-\Delta-p) \\ &\quad + \sum_{q=0}^{L-1} a_m(q)v(rk-\Delta-q), \end{aligned} \tag{3}$$

where L is the length of the analysis filter $a_m(n)$.

3 Analysis of the Subband Impulse Response

In this section, we will analyze the subband impulse response, given an analysis filterbank and a fullband impulse response. Of special interest is the subband impulse responses in systems with non-critical downsampling, i.e. $r < M$, where r is the downsampling factor and M is the number of subbands. The analysis is divided into four parts. First we will derive the true subband impulse response, given an ideal system. Then we will consider the infinitely long Wiener solution for a filterbank with

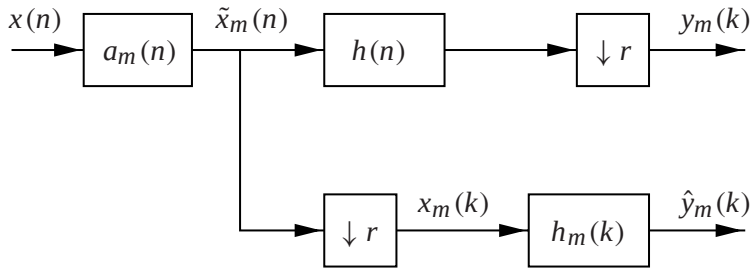

Figure 3: In an ideal system, this model is equivalent to the model presented in Fig. 2 for $\Delta = 0$ and $v(n) = 0$.

finitely long filters. Aspects such as downsample aliasing will be taken into account. Thereafter we will limit the subband impulse response to a causal FIR filter and the optimal subband impulse response in a minimum mean-square error sense is calculated. Finally, we will examine what influence the use of the LMS algorithm will have on the subband impulse response.

3.1 Ideal System

In an ideal system, the analysis filterbank has infinitely long non-causal bandpass filters with perfect frequency separation. Both the near-end signal, $v(n)$, and the introduced signal path delay, Δ, are set to zero, in Fig. 2. Now $\tilde{y}_m(n)$ can be written exactly as a linear combination of the input signal samples $x(n-i)$, $i \in \mathbf{Z}$. The two filters, $h(n)$ and $a_m(n)$, are interchangeable, and an equivalent structure is shown in Fig. 3. Since the signal $\tilde{x}_m(n)$ is an ideal bandpass signal, it is possible to reconstruct it from the downsampled signal $x_m(k)$, if the downsampling rate satisfies the sampling theorem. Let $u(n)$ denote the upsampled version of $x_m(k)$,

$$u(n) = \begin{cases} x_m(k), & n = rk \\ 0, & n \neq rk \end{cases} . \tag{4}$$

Then the fullband signal $\tilde{x}_m(n)$ can be written as

$$\tilde{x}_m(n) = u(n) * h_m^{\mathrm{BP}}(n), \tag{5}$$

where $h_m^{\mathrm{BP}}(n)$ is an ideal bandpass filter,

$$h_m^{\mathrm{BP}}(n) = \operatorname{sinc}\left(\frac{n}{r}\right) e^{j\frac{2\pi m}{M}n}, \qquad n \in \mathbf{Z}, \tag{6}$$

and $\operatorname{sinc} x \equiv \frac{\sin \pi x}{\pi x}$. The center frequency of the bandpass filter is determined by the subband number, m, and the bandwidth by the downsampling factor, r. Using (5), the

true subband near-end signal can be expressed as a function of $u(n)$,

$$y_m(k) = \sum_{i=-\infty}^{\infty} \sum_{j=0}^{N-1} u(rk - i - j) h_m^{\mathrm{BP}}(i) h(j). \tag{7}$$

Since the input signal in (7), $u(n)$, has non-zero samples only for $n = rk$, (4), the noble identity [11] is valid and the filtering and the downsampling can be interchanged:

$$\begin{aligned} y_m(k) &= \sum_{p=-\infty}^{\infty} u(rk - rp) \sum_{q=0}^{N-1} h_m^{\mathrm{BP}}(rp - q) h(q) \\ &= x_m(k) * \left[(h_m^{\mathrm{BP}} * h) \downarrow r \right]. \end{aligned} \tag{8}$$

The true subband impulse response, now directly given by (8), as

$$h_m(k) = \left[h_m^{\mathrm{BP}}(n) * h(n) \right] \downarrow r, \tag{9}$$

shows that a *causal* fullband impulse response of finite length transforms into a *non-causal* subband impulse responses with *infinite* length, [4], if $h(n) \neq 0$ for any $n \neq rk$, $k \in \mathbf{Z}$.

3.2 Wiener Solution

In this section, we will derive an expression for the transfer function of the subband impulse response when the analysis filterbank, a_m, is *non-ideal*. An illustration of such filters are shown in Fig. 4. Like in the previous section, the corresponding impulse response may be infinitely long and non-causal.

The optimal transfer function for subband m can be derived from the power spectra and cross-power spectra of the subband signals $x_m(k)$ and $y_m(k)$ [12],

$$H_m^{\mathrm{wien}}(f) = \frac{S_{y_m,x_m}(f)}{S_{x_m,x_m}(f)}. \tag{10}$$

Note that the corresponding impulse response, $h_m^{\mathrm{wien}}(k)$, can be of infinite length and non-causal. The power spectra of the subband signals are derived in Appendix A. By setting the variable delay $\Delta = 0$, see (3), they can be expressed as,

$$S_{x_m x_m}(f) = \frac{\sigma_x^2}{r} \sum_{d=0}^{r-1} \left| A_m\left(\frac{f-d}{r} \right) \right|^2, \tag{11}$$

$$S_{y_m x_m}(f) = \frac{\sigma_x^2}{r} \sum_{d=0}^{r-1} H\left(\frac{f-d}{r} \right) \left| A_m\left(\frac{f-d}{r} \right) \right|^2, \tag{12}$$

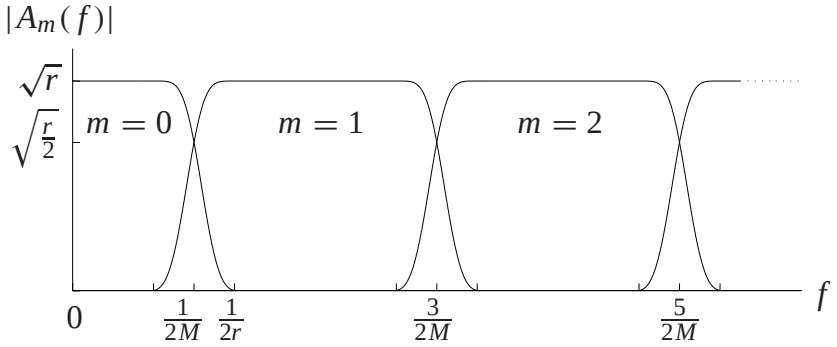

Figure 4: An example of a filterbank suitable for the structure presented in Fig. 1. The filters approximately prevent aliasing by sufficient stopband attenuation; see filter $m = 0$ for $f > \frac{1}{2r}$.

where H and A_m are the Fourier transforms of $h(n)$ and a_m, respectively. Using (10), (11), and (12), the optimal subband transfer function is

$$H_m^{\text{wien}}(f) = \frac{H(\frac{f-d_m}{r})\left|A_m(\frac{f-d_m}{r})\right|^2 + \sum_{d=0,d\neq d_m}^{r-1} H(\frac{f-d}{r})\left|A_m(\frac{f-d}{r})\right|^2}{\left|A_m(\frac{f-d_m}{r})\right|^2 + \sum_{d=0,d\neq d_m}^{r-1}\left|A_m(\frac{f-d}{r})\right|^2}, \qquad (13)$$

where d_m denotes an integer for which the analysis filter $A_m(\frac{f-d_m}{r})$ has its passband somewhere in the region $-\frac{1}{2} < f < \frac{1}{2}$. That is, the first term in the numerator in (13) corresponds to the expected impulse response estimate and the following sum corresponds to aliasing errors. In the passband of $A_m(f)$, Fig. 5, the summations in both the numerator and denominator are small and equation (13) can be written as,

$$H_m^{\text{wien}}(f) = H(\frac{f-d_m}{r}) + \varepsilon_{\text{PB}}(f), \qquad \frac{2m-1}{2M} < \frac{f-d_m}{r} < \frac{2m+1}{2M}, \qquad (14)$$

where $\varepsilon_{\text{PB}}(f)$ is the error due to aliasing. The error has the subscript PB since this is in the passband region of subband m. With properly designed analysis filters, a_m, this error should be small, typically $|\varepsilon_{\text{PB}}(f)| \leq (r-1)A_{\text{SB}}H(\frac{f-d_m}{r})$, where A_{SB} is the stopband attenuation of a_m. This since ε_{PB} is a sum of $r-1$ aliasing errors that have been suppressed by the filter a_m. In the transition bands of a_m, $\frac{m}{M} - \frac{1}{2r} < \frac{f-d_m}{r} < \frac{2m-1}{2M}$ and $\frac{2m+1}{2M} < \frac{f-d_m}{r} < \frac{m}{M} + \frac{1}{2r}$, all terms in (13) are small, especially for $\frac{f-d_m}{r} \approx \frac{m}{M} - \frac{1}{2r}$ and $\frac{f-d_m}{r} \approx \frac{m}{M} + \frac{1}{2r}$, Fig. 5. Let us rewrite (13) for one specific

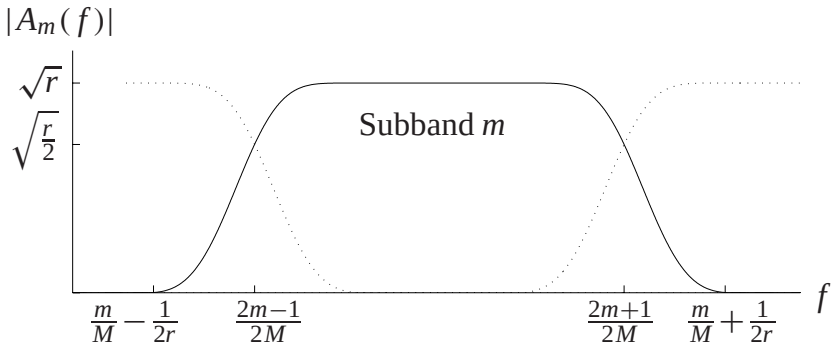

Figure 5: Frequency response of the analysis filter for subband m. The region between the $\frac{2m-1}{2M}$ and $\frac{2m+1}{2M}$ is denoted the passband, and between $\frac{m}{M}-\frac{1}{2r}$ and $\frac{2m-1}{2M}$ the transition band. At frequencies close to $\frac{m}{M}-\frac{1}{2r}$ and $\frac{m}{M}+\frac{1}{2r}$, the subband signals $x_m(k)$ and $y_m(k)$ are suppressed, and are therefore disturbed by aliasing.

frequency, f_0,

$$H_m^{\text{wien}}(f_0) = \frac{\alpha_{d_m} H(\frac{f_0-d_m}{r}) + \sum_{d=0,d\neq d_m}^{r-1} \alpha_d H(\frac{f_0-d}{r})}{\sum_{d=0}^{r-1} \alpha_d}, \tag{15}$$

where $\alpha_{(\cdot)}$ can been seen as attenuation constants. α_{d_m} correspond to the attenuation constant for the expected subband, and all other $\alpha_{(\cdot)}$ for alias components. For f_0 close to the band edges of the analysis filter, Fig. 5, α_{d_m} is small and there exist other $\alpha_{(\cdot)}$ with the same magnitude. How many and for which frequencies f_0 that other $\alpha_{(\cdot)}$ are in the same magnitude as α_{d_m} depend on the characteristics of the stopband region of a_m. Moreover, magnitude and phase changes of $H(f)$ also affect how $H_m^{\text{wien}}(f_0)$ is effected by the alias components.

Let's start with examine a special case, namely, we restrict $H(f)$ so that all alias components in (15) have the same magnitude,

$$\left|H\left(\frac{f-d_m}{r}\right)\right| = \left|H\left(\frac{f-d}{r}\right)\right| \qquad d \in [0,\ldots,r-1], \tag{16}$$

and a linear phase-function, with phase depending on the downsampling factor r,

$$\arg\{H(f)\} = 2\pi f r i, \qquad i \in \mathbf{Z}. \tag{17}$$

One function satisfying (16) and (17) is the pure delay function, $h(n) = \delta(n-ir)$, where the delay is a multiple of the downsampling factor. If (16) and (17) are satisfied,

(15) can be reduced to,

$$H_m^{\text{wien}}(f_0) = H\left(\frac{f_0 - d_m}{r}\right), \qquad -\frac{1}{2} < f_0 < \frac{1}{2}. \tag{18}$$

That is, $H_m^{\text{wien}}(f_0)$ has the expected value for all frequencies, and $H_m^{\text{wien}}(f)$ is a causal function if $H(f)$ is causal.

In a second interesting special case we assume that several of the terms $\alpha_d H(\frac{f_0-d}{r})$, in (15), have the same order of magnitude as $\alpha_{d_m} H(\frac{f_0-d_m}{r})$. If, in addition to this, the phase function of the terms are changing fast enough, (15) can be reduced to,

$$H_m^{\text{wien}}(f_0) \approx 0, \tag{19}$$

for f_0 close to the band-edges. In this case $H_m^{\text{wien}}(f)$ approximately follows the behavior of the analysis filter, thus spreading of the impulse response (inverse Fourier transform of $H_m^{\text{wien}}(f)$) occurs, thereby necessitating non-causal taps in the model filter.

Finally we examine (15) for arbitrary $H(f)$. Even if the uncertainty is large, one can conclude that for f_0 close the the band edges, $H_m^{\text{wien}}(f_0)$ is likely to have a value that significantly differ from the expected value, i.e.

$$H_m^{\text{wien}}(f_0) = \frac{\alpha_{d_m} H(\frac{f_0-d_m}{r})}{\sum_{d=0}^{r-1} \alpha_d} + \varepsilon_{\text{TB}}(f_0), \tag{20}$$

where $\varepsilon_{\text{TB}}(f_0)$ can have the same order of magnitude as $H(\frac{f_0-d_m}{r})$. The subscript TB on the error is used since this is in the transition band of subband m. In many situations, this behavior at the band edges, are better modeled with non-causal rather than causal $H_m^{\text{wien}}(f)$.

3.3 Minimum Mean Square Error Solution of Non-ideal Systems

In this section, we will study the minimum mean square error (MSE) solution. In contrast to the previous section, a realizable system with finite length filters is assumed. As has been shown in Section 3.1, the true subband impulse response is non-causal and of infinite length. In order to model a number of non-causal impulse response taps, a delay Δ is introduced in the near-end signal path, Fig. 2.

The optimum subband response is found by minimizing the mean squared error in each subband,

$$\min_{\hat{h}_m,\Delta} \mathrm{E}\,|e_m(k)|^2 = \min_{\hat{h}_m,\Delta} \mathrm{E}\left| y_m(k) - \sum_{i=0}^{N_m-1} \hat{h}_m(i) x_m(k-i) \right|^2, \tag{21}$$

where N_m is the length of the estimated subband impulse response $\hat{h}_m$. The solution to (21), for a given Δ, is the well-known finite length Wiener filter [1]. The optimum subband transfer function can be written in matrix form as

$$\mathbf{h}_m^{\text{mse}} = \mathbf{R}_{x_m x_m}^{-1} \boldsymbol{\gamma}_{y_m x_m}, \tag{22}$$

where $\mathbf{R}_{x_m x_m}$ is the correlation matrix,

$$\mathbf{R}_{x_m x_m} = \begin{bmatrix} \gamma_{x_m x_m}(0) & \gamma_{x_m x_m}(1) & \cdots & \gamma_{x_m x_m}(N_m-1) \\ \gamma^*_{x_m x_m}(1) & \gamma_{x_m x_m}(0) & \cdots & \gamma_{x_m x_m}(N_m-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^*_{x_m x_m}(N_m-1) & \gamma^*_{x_m x_m}(N_m-2) & \cdots & \gamma_{x_m x_m}(0) \end{bmatrix} \tag{23}$$

and $\mathbf{h}_m^{\text{mse}}$ and $\boldsymbol{\gamma}_{y_m x_m}$ are, respectively, the optimal impulse response and the cross-correlation vector,

$$\mathbf{h}_m^{\text{mse}} = \begin{bmatrix} h_m^{\text{mse}}(0) & h_m^{\text{mse}}(1) & \cdots & h_m^{\text{mse}}(N_m-1) \end{bmatrix}^T, \tag{24}$$

$$\boldsymbol{\gamma}_{y_m x_m} = \begin{bmatrix} \gamma_{y_m x_m}(0) & \gamma_{y_m x_m}(1) & \cdots & \gamma_{y_m x_m}(N_m-1) \end{bmatrix}^T. \tag{25}$$

The estimation error resulting from using the optimal filter $\mathbf{h}_m^{\text{mse}}$ is then [1],

$$\varepsilon_{min} = \min_{\hat{h}_m, \Delta} \mathrm{E}\,|e_m(k)|^2 = \min_{\Delta} \left[\gamma_{y_m y_m}(0) - \boldsymbol{\gamma}^H_{y_m x_m} \mathbf{h}_m^{\text{mse}} \right]. \tag{26}$$

In order to determine the optimal MSE subband impulse response, we have to derive expressions for the auto- and cross-correlation functions of the subband signals $x_m(k)$ and $y_m(k)$, as defined in Section 2 and illustrated in Fig. 2. The calculations are presented in Appendix A. The autocorrelation of the downsampled far-end signal in subband m, $x_m(k)$, can be expressed as,

$$\gamma_{x_m x_m}(l) = \sum_{i=0}^{L-1} a_m(i) a_m^*(i - rl) \sigma_x^2, \tag{27}$$

where a_m denotes the analysis filter for subband m, σ_x is the standard deviation of the white input-signal $x(n)$, and r the downsampling factor. The cross-correlation between the near-end signal $y_m(k)$ and far-end signal $x_m(k)$ in subband m is

$$\gamma_{y_m x_m}(l) = \sum_{i=0}^{N-1} \sum_{j=-\infty}^{\infty} a_m(j-i) h(i) a_m^*(j + \Delta - rl) \sigma_x^2 \tag{28}$$

where h denotes the fullband system impulse response to be estimated and Δ the delay introduced in the near-end signal $y(n)$, as defined in Section 2. Finally, the variance of the near-end subband signal,

$$\gamma_{y_m y_m}(0) = \sum_{i=0}^{N-1} \sum_{p=0}^{N-1} \sum_{j=-\infty}^{\infty} a_m(j-i) a_m^*(j-p) h(i) h^*(p) \sigma_x^2 + \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} a_m(i) a_m^*(j) \gamma_{vv}(j-i), \tag{29}$$

where $\gamma_{vv}(l)$ denotes the autocorrelation function of the zero-mean disturbance $v(n)$. By using (22), (27) and (28) we can now determine the optimal subband impulse response for a given fullband response and a given analysis filterbank. The subband impulse response can be determined for a given number of non-causal taps by selecting the near-end signal delay value Δ, Fig. 2. By using (26), we can also determine the optimal near-end signal delay value Δ for a given fullband response and analysis filterbank.

For infinitely long filters, the MSE solution presented in (22) is identical to the Wiener solution in (10). Next we will show how $\mathbf{h}_m^{\text{mse}}$ in (22) is affected by reducing the filter length, i.e. how the MSE solution differ from the Wiener solution. The length of $\mathbf{h}_m^{\text{mse}}$ in (22) is determined by the sizes of $\mathbf{R}_{x_m x_m}$ and $\boldsymbol{\gamma}_{y_m x_m}$, and this will be used in deriving an expression for the relation between the MSE solution and the Wiener solution. Let's start by defining $\mathbf{h}_m^{\text{wien}}$ as a $(2C+1)N_m$ long vector with indices $-CN_m$ to $(C+1)N_m - 1$, which approaches the Wiener solution (10) as $C \to \infty$. Then, by using (22), we can find the relation between $\mathbf{h}_m^{\text{wien}}$ and $\mathbf{h}_m^{\text{mse}}$,

$$\mathbf{R}_{x_m x_m} \mathbf{h}_m^{\text{mse}} = \boldsymbol{\gamma}_{y_m x_m} = \begin{bmatrix} \mathbf{R}_{C^-} & \mathbf{R}_{x_m x_m} & \mathbf{R}_{C^+} \end{bmatrix} \mathbf{h}_m^{\text{wien}}, \tag{30}$$

where $\mathbf{R}_{C^-}$ and $\mathbf{R}_{C^+}$ are matrices of size $(N_m \times CN_m)$ with the following elements,

$$[\mathbf{R}_{C^-}]_{(p,q)} = \gamma_{x_m x_m}^*(CN_m + p - q), \tag{31}$$

$$[\mathbf{R}_{C^+}]_{(p,q)} = \gamma_{x_m x_m}(N_m - p + q), \tag{32}$$

where $[\cdot]_{(p,q)}$ denotes the element in row $p \in [0, \ldots, N_m - 1]$, and column $q \in [0, \ldots, CN_m]$. The matrix $\begin{bmatrix} \mathbf{R}_{C^-} & \mathbf{R}_{x_m x_m} & \mathbf{R}_{C^+} \end{bmatrix}$ denotes row 0 to $N_m - 1$ of the large correlation matrix that belongs to $\mathbf{h}_m^{\text{wien}}$. That is, $\mathbf{R}_{C^-}$ is the symmetric extension on the left side of the correlation matrix $\mathbf{R}_{x_m x_m}$, see (23), and $\mathbf{R}_{C^+}$ the extension on the right side. The MSE estimate can now be written as,

$$\mathbf{h}_m^{\text{mse}} = \tilde{\mathbf{h}}_m^{\text{wien}} + \mathbf{R}_{x_m x_m}^{-1} \begin{bmatrix} \mathbf{R}_{C^-} & \mathbf{0}_{N_m \times N_m} & \mathbf{R}_{C^+} \end{bmatrix} \mathbf{h}_m^{\text{wien}}, \tag{33}$$

where the first term in (33), $\tilde{\mathbf{h}}_m^{\text{wien}}$, is a sub-vector of $\mathbf{h}_m^{\text{wien}}$ with elements 0 to $N_m - 1$. The second term in (33) expresses how instances of $\mathbf{h}_m^{\text{wien}}$ outside rows 0 to $N_m - 1$ affects $\mathbf{h}_m^{\text{mse}}$, which is only defined for rows 0 to $N_m - 1$. This term in a way describe what we need to sacrifice when we reduce the filter from the size of $\mathbf{h}_m^{\text{wien}}$ to the size of $\mathbf{h}_m^{\text{mse}}$. To gain a better understanding of this second term we will transfer the whole expression to the frequency domain. The transformation is performed by pre-multiplying (33) with the Fourier matrix $\mathbf{Q}^H$,

$$\left[\mathbf{Q}^H\right]_{(p,q)} = \frac{1}{\sqrt{N_m}} e^{-j\frac{2\pi pq}{N_m}}, \quad p, q \in [0, \ldots, N_m - 1]. \tag{34}$$

As the Fourier transformation of correlation matrices requires Fourier matrices on both the left and the right side, $\mathbf{Q}\mathbf{Q}^H$ will be inserted. The fact that $\begin{bmatrix}\mathbf{R}_{C-} & \mathbf{0}_{N_m \times N_m} & \mathbf{R}_{C+}\end{bmatrix}$ is not a quadratic matrix, and that $\mathbf{h}_m^{\text{wien}}$ is bigger than $\mathbf{h}_m^{\text{mse}}$ also need to be addressed. Instead of performing one large Fourier transform, we will calculate the average over several shorter transforms, by inserting $\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^H/(2C+1)$, where $\tilde{\mathbf{Q}}$ is a block-matrix of size $((2C+1)N_m \times N_m)$ with elements $\mathbf{Q}$. The transform is then,

$$\begin{aligned}\mathbf{Q}^H\mathbf{h}_m^{\text{mse}} &= \mathbf{Q}^H\tilde{\mathbf{h}}_m^{\text{wien}} + \mathbf{Q}^H\mathbf{R}_{x_m x_m}^{-1}\mathbf{Q}\mathbf{Q}^H \\ &\quad \cdot \begin{bmatrix}\mathbf{R}_{C-} & \mathbf{0}_{N_m \times N_m} & \mathbf{R}_{C+}\end{bmatrix} \frac{\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^H}{2C+1}\mathbf{h}_m^{\text{wien}} \qquad (35) \\ &= \mathbf{Q}^H\tilde{\mathbf{h}}_m^{\text{wien}} + \mathbf{Q}^H\mathbf{R}_{x_m x_m}^{-1}\mathbf{Q}\mathbf{Q}^H\mathbf{R}_t\mathbf{Q}\tilde{\mathbf{Q}}^H\mathbf{h}_m^{\text{wien}}, \qquad (36)\end{aligned}$$

where $\mathbf{Q}^H\mathbf{h}_m^{\text{mse}}$ is the Fourier transform of $\mathbf{h}_m^{\text{mse}}$, denoted $\mathbf{H}_m^{\text{mse}}$. In (36), the averaging is incorporated into $\mathbf{R}_t$, which is a $(N_m \times N_m)$ matrix with the following elements,

$$[\mathbf{R}_t]_{(p,q)} = \frac{1}{2C+1}\sum_{l=0}^{C-1}\left\{\gamma_{x_m x_m}((l+1)N_m - p + q) + \gamma^*_{x_m x_m}((C+l)N_m + p - q)\right\}. \tag{37}$$

As is shown in [13], $\mathbf{Q}^H\mathbf{R}_{x_m x_m}\mathbf{Q}$ is a good approximation of a diagonal matrix where the elements on the diagonal are the sampled power spectra, $S_{x_m x_m}(q)$, defined in (11). The signal x_m is filtered by the bandpass filter a_m, Fig. 5, and has therefore near zero energy close to the band edge. Correspondingly, the power spectra is approximately zero close to the band edge, and that frequency bin is denoted q_0, i.e. $S_{x_m x_m}(q_0) \approx 0$. Therefore the inverse, $\mathbf{Q}^H\mathbf{R}_{x_m x_m}^{-1}\mathbf{Q}$, will have small values except for a sharp peak in and possibly around the element $[\mathbf{Q}^H\mathbf{R}_{x_m x_m}^{-1}\mathbf{Q}]_{(q_0,q_0)}$.

Due to the bandpass filter a_m, the correlation function $\gamma_{x_m x_m}(l)$ is bounded by a decreasing function. Therefore, $\mathbf{R}_t$ will have its biggest elements at or close to the upper right and lower left corner. As is examined in [13], this cannot be approximated by

a diagonal matrix. Hence, $\mathbf{Q}^H \mathbf{R}_{x_m x_m}^{-1} \mathbf{R}_t \mathbf{Q}$ in (36) will, due to the peak in $\mathbf{Q}^H \mathbf{R}_{x_m x_m}^{-1} \mathbf{Q}$, have one or a few adjacent rows, at row q_0, that are dominant. Therefore, the value of $\mathbf{H}_m^{\text{mse}}(q_0)$ can be significantly different from the Wiener solution, $\mathbf{H}_m^{\text{wien}}(q_0)$. Actually, $\mathbf{H}_m^{\text{mse}}$ can have a near arbitrary shape for frequencies close to the band edges of the analysis filter a_m.

3.4 Adaptive LMS Solution

The LMS algorithm is known to converge to the MSE solution as the time index $k \to \infty$, except for a small error denoted excess mean-squared error [1]. Of more interest is knowledge about the convergence over a shorter time interval. Also in [1], it is shown how a large eigenvalue spread of the correlation matrix decreases the convergence rate. In this section it will be shown how this affects the convergence of the LMS in subband structures.

Denoting the optimal impulse response by $\mathbf{h}_m^{\text{mse}}$ and the estimated impulse response at time k by $\hat{\mathbf{h}}_m(k)$, the weight-error vector is defined as,

$$\boldsymbol{\epsilon}_m(k) = \hat{\mathbf{h}}_m(k) - \mathbf{h}_m^{\text{mse}}. \tag{38}$$

In [1] the following LMS weight-error vector update formula is given,

$$\boldsymbol{\epsilon}_m(k+1) = [\mathbf{I} - \mu \mathbf{x}_m(k)\mathbf{x}_m^H(k)]\boldsymbol{\epsilon}_m(k) + \mu \mathbf{x}_m(k) e_{\text{o}}^*(k), \tag{39}$$

where $\mathbf{I}$ denotes the identity matrix, $\mathbf{x}_m(k)$ the input vector $[x_m(k) \ldots x_m(k - N_m + 1)]^T$, μ the LMS step-size parameter, and e_{o} the estimation error produced in the Wiener solution. The ensemble average of input-signal matrix $\mathbf{x}_m(k)\mathbf{x}_m^H(k)$ equals the correlation-matrix,

$$\mathbf{R}_{x_m x_m} = \text{E}\{\mathbf{x}_m(k)\mathbf{x}_m^H(k)\}. \tag{40}$$

The correlation-matrix can approximately be decomposed into two Fourier matrices (34) and one diagonal matrix [13],

$$\mathbf{R}_{x_m x_m} \approx \mathbf{Q} \boldsymbol{\Lambda}_m \mathbf{Q}^H, \tag{41}$$

where $\boldsymbol{\Lambda}_m$ is the diagonal power spectral matrix,

$$\boldsymbol{\Lambda}_m = \text{diag}\left[S_{x_m x_m}(0), S_{x_m x_m}(1), \ldots, S_{x_m x_m}(N_m - 1)\right]. \tag{42}$$

Now the expected value of (39) can then be written as,

$$\text{E}\{\boldsymbol{\epsilon}_m(k+1)\} = \mathbf{Q}[\mathbf{I} - \mu \boldsymbol{\Lambda}_m]\mathbf{Q}^H \text{E}\{\boldsymbol{\epsilon}_m(k)\} + \mu \text{E}\{\mathbf{x}_m(k) e_{\text{o}}^*(k)\}. \tag{43}$$

In order to study the convergence of the LMS algorithm in the frequency domain, we transform (43), using the Fourier transform operator $\mathbf{Q}^H$,

$$\begin{aligned}\mathbf{\Gamma}_m(k+1) &= \mathbf{Q}^H \mathrm{E}\{\boldsymbol{\epsilon}_m(k+1)\} \\ &= [\mathbf{I} - \mu\mathbf{\Lambda}_m]\mathbf{\Gamma}_m(k) + \mu\mathrm{E}\{\mathbf{X}_m(k)e_o^*(k)\},\end{aligned} \tag{44}$$

where $\mathbf{X}_m(k) = \begin{bmatrix} X_m(k,0) & \dots & X_m(k,N_m-1)\end{bmatrix}^T$ is the discrete Fourier transform of $\mathbf{x}_m(k)$ and $\mathbf{\Gamma}_m(k)$ the discrete Fourier transform of the weight-error vector $\mathrm{E}\{\boldsymbol{\epsilon}_m(k)\}$. When the LMS algorithm is operating in a subband structure with non-critical downsampling, $\mathbf{x}_m(k)$ will be a bandpass signal, since the analysis filterbank a_m will suppress signals in the frequency range outside subband m. In those suppressed frequency regions, the corresponding power spectrum $S_{x_m x_m}(q)$ is approximately zero. If we examine the frequency domain LMS weight-error vector update formula, (44), in the stopband frequency region of the analysis filter, these modes will hardly be updated since

$$1 - \mu S_{x_m x_m}(q) \approx 1, \tag{45}$$

$$X_m(k,q) \approx 0. \tag{46}$$

Therefore the LMS algorithm will on the average converge extremely slowly in those frequency regions, i.e. the estimates will remain close to the initial value, usually zero. Due to this, the estimated subband transfer function will have a bandpass characteristic, approximated by

$$\hat{H}_m(f) \approx H_m^{\mathrm{mse}}(f) G_m^{\mathrm{BP}}(f), \tag{47}$$

where G_m^{BP} is an ideal bandpass filter, and transformed into the time domain, we find,

$$\hat{h}_m(k) \approx h_m^{\mathrm{mse}}(k) * g_m^{\mathrm{BP}}(k), \tag{48}$$

where the bandpass filter is,

$$g_m^{\mathrm{BP}}(k) = \frac{1}{W}\,\mathrm{sinc}\left(\frac{k}{W}\right) e^{j\frac{2\pi m}{M}k}, \qquad k \in \mathbf{Z}. \tag{49}$$

The center frequency of the bandpass filter is given by the subband, m, and the bandwidth by W, where W is dependent on the analysis filter a_m and is in the range $r < W < M$, Fig. 5.

From (48), it is clear that subband impulse responses estimated by the LMS algorithm in systems with non-critical downsampling will be similar to the true subband impulse response, given in (9). However, $g_m^{\mathrm{BP}}(k)$ has smaller bandwidth than the bandpass filter in the ideal solution, (9), since $r < M$. Consequently, the LMS estimated

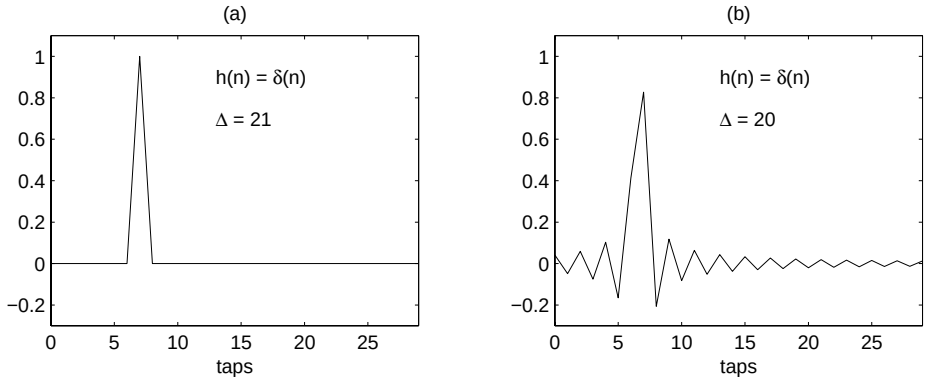

Figure 6: Ideal subband impulse response for the fullband impulse response $h(n) = \delta(n)$, (a) Introduced signal path delay $\Delta = 21$, see Fig. 1. Note that Δ is a multiple of the downsampling factor $r = 3$. (b) $\Delta = 20$.

subband impulse response may need *more non-causal taps* than the ideal subband impulse response. In Section 4, the validity of (48) is confirmed by simulation examples. In the design of systems, it is important to compensate for the non-causal subband impulse response taps, especially when the flat delay of the fullband impulse response is short. As presented in earlier sections, non-causal taps are made causal by delaying the near-end signal Δ taps [4], see Fig. 1 and 2.

4 Examples

The first example will show two delayed and truncated *ideal* subband impulse responses, as derived in Section 3.1. Consider a system with the fullband impulse response $h(n) = \delta(n)$ and a downsampling factor of $r = 3$. In latter examples, we consider systems with $M = 4$ subbands, but this information is not needed in order to calculate the ideal impulse response. In order to model a number of non-causal filter taps with a causal FIR filter, a delay Δ is introduced in the near-end signal, Fig. 1 and 2. If Δ is a multiple of the downsampling factor r, the total impulse response, $h(n) = \delta(n - \Delta r)$, can be perfectly modeled with a single tap. This is shown in Fig. 6(a). If the delay is not a multiple of the downsampling factor, the subband impulse response is an infinitely long non-causal filter. A truncated example is shown in Fig. 6(b). Since the lowest subband, $m = 0$, is real-valued, it will be used in all figures.

In Section 3.3, formulas to calculate the minimum mean square error subband impulse response were derived. Consider the fullband impulse response from the previous example. As above, the filterbank has $M = 4$ subbands and the downsample

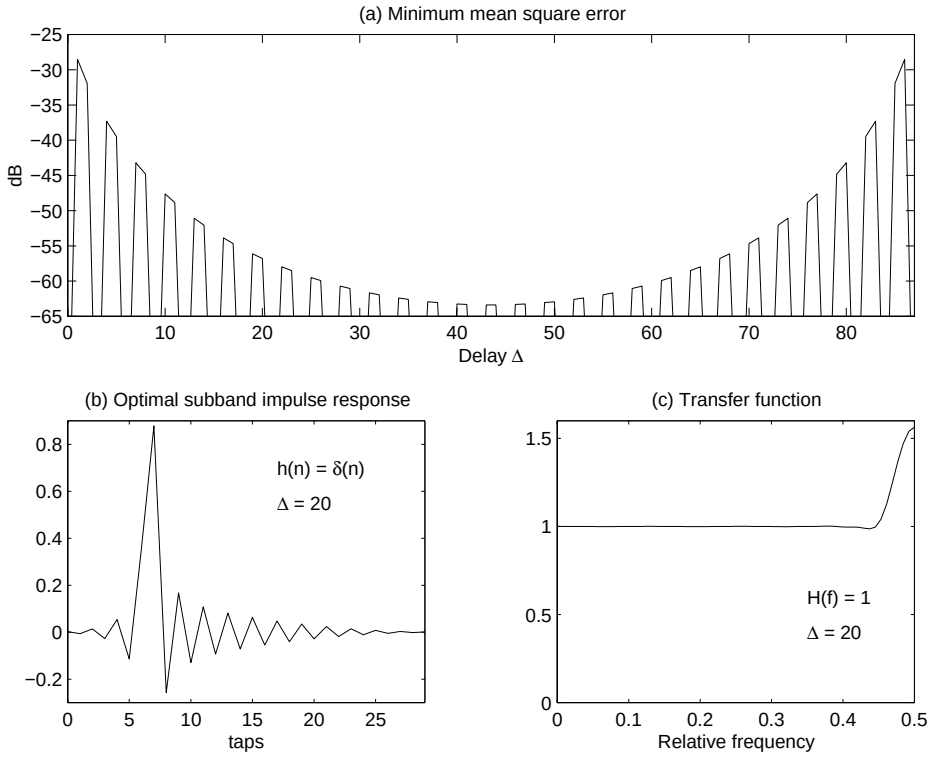

Figure 7: Optimal subband characteristics of subband 0. Fullband impulse response $h(n) = \delta(n)$. (a) Minimum mean square error as a function of introduced near-end signal delay Δ. (b) Optimal subband impulse response for $\Delta = 20$. (c) Corresponding transfer function.

factor is $r = 3$. The analysis filters are designed to have a flat passband region with an amplification of $\sqrt{r}$. In the transition bands, the sum of two adjacent filters will also have an amplification of $\sqrt{r}$ and in the stopband regions the filters have at least 80 dB attenuation. The filters are illustrated in Fig. 4. The adaptive subband filter length is 30. The minimum mean-squared error of the optimal subband impulse response, ε_{min}, is then calculated using (26) for different values of the near-end delay Δ. In Fig. 7(a), ε_{min} is plotted as a function of Δ. Especially notice that when Δ is a multiple of r, the optimum subband impulse response has zero minimum mean-squared error. This occurs since the fullband impulse response $h(n) = \delta(n-rk)$, k positive integer, has an exact representation in the subbands for downsampling factor r. This also corresponds to the special case (16), (17) and (18), analyzed in Section 3.2. The optimal subband impulse response, $\mathbf{h}_m^{\text{mse}}$ in (22), for $\Delta = 20$ is shown in Fig. 7(b). Since the delay introduced in the near-end signal equals 20 and the downsampling factor is 3, subband impulse response taps 0 to 6 in Fig. 7(b) represent non-causal filter taps.

The result in Fig. 7(a) shows the error when the optimum subband impulse response is calculated according to (22). Usually, we need to estimate the subband impulse response with an adaptive filter. Using the same setup as in Fig. 7, we estimate the subband impulse response with the normalized LMS algorithm instead of calculating the optimum response by (22). A white Gaussian noise signal is used as far-end signal, and the NLMS algorithm is adapting on 20000 samples. The last 1000 samples are used to calculate the variance of the error signal and this value is averaged over 100 simulations. The results are shown in Fig. 8. As can be seen in Fig. 8(a), the mean square error is reduced when a delay Δ is introduced to the near-end signal. If Fig. 8(c) is compared against Fig. 7(c), it is obvious that the LMS solutions has not converged to a value close to the minimum mean square solution for frequencies $f \approx 0.5$. This is due to the fact that the NLMS algorithm converges extremely slowly in frequency regions with near zero signal energy, as is shown in Section 3.4. Because of this, a slightly bigger Δ may be needed when the NLMS algorithm is used, in comparison with studies of the MSE solution or when adaptive algorithms which take the signal correlation matrix into account are used.

For comparison, the previous simulation is repeated with the recursive least squares (RLS) adaptive algorithm. The RLS algorithm is adapting from a zero state on 1000 samples. The variance of the residual error is calculated using the last 500 samples, and an average over 100 iterations is presented in Fig. 9(a). Note that the RLS solution is very close to the optimal solution presented in Fig. 7, including the frequency response for $f \approx 0.5$ in Fig. 9(c).

In Fig. 10, $h(n)$ is replaced with its continuous counterpart, $h(t) = \delta(t - t_0)$, and the sample rate is normalized to 1. In the Figure, the minimum mean square error is shown for $0 \leq t_0 \leq 60$, and the artificial flat delay, $\Delta = 0$. The non-causal characteristics of subband impulse responses can also be explained as an interpolation phenomena, and this figure resembles the figures in a study of interpolation of ideal band-limited signals [14].

Finally, shown in Fig. 11, is the size of the delay Δ needed to get a minimum mean square error less than -50 dB, for the impulse response $h(n) = \delta(n)$. In average, this corresponds to 4.5 non-causal taps in the adaptive filters in the subbands. It should be noted that this is dependent on the stopband attenuation of the filters used in the filterbank, and that adaptive filters like the NLMS may converge extremely slowly to the minimum mean square error solution, as shown in Sec. 3.4. In practical situations, the impulse response usually include a flat delay, i.e. $h(n) = 0$ for $0 < n < n_0$, and the size of Δ can be reduced correspondingly.

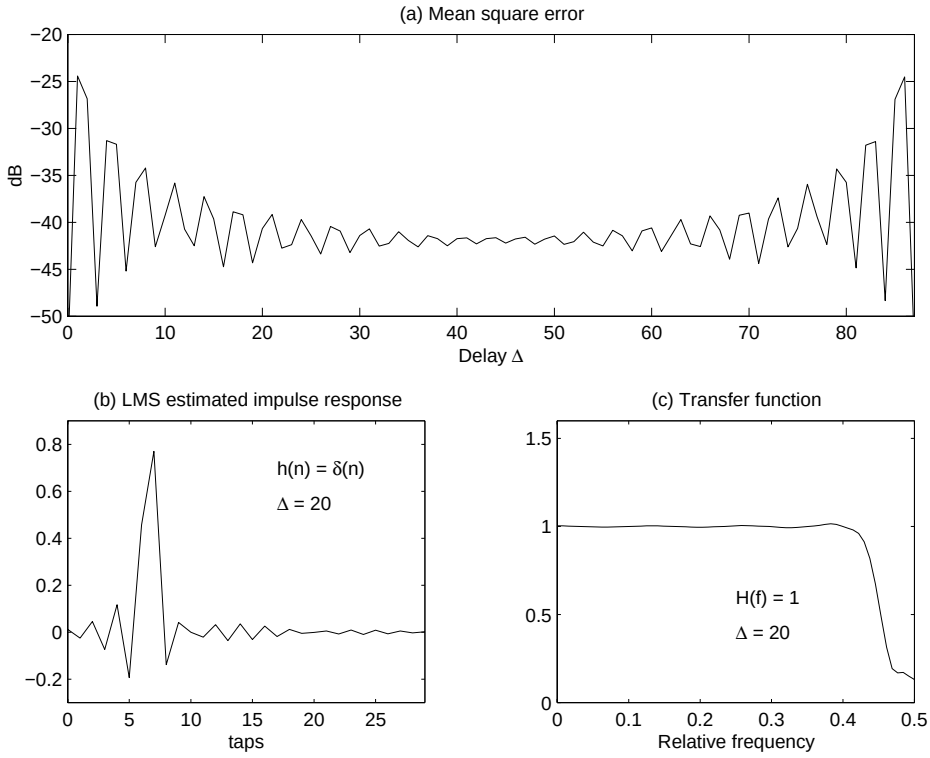

Figure 8: LMS-estimated subband impulse response for a non-critically downsampled system, subband 0. Fullband impulse response $h(n) = \delta(n)$. (a) Mean square error, averaged over 100 simulations, as a function of the introduced near-end signal delay Δ. (b) LMS estimated impulse response for $\Delta = 20$. (c) Corresponding transfer function.

5 Discussion

We have studied the subband impulse responses in subband echo cancelers. The analysis shows that *non-causal* subband impulse responses can model causal fullband impulse responses better than *causal* subband impulse responses. This is shown in four different ways in the paper. First an ideal system is studied, and it is shown that due to the limited bandwidth, the subband impulse response is not limited in time. Then the Wiener solution in the frequency domain is derived, and it is shown that the value of the transfer function estimate at the subband edges can have values that radically differ from the expected value. If an inverse Fourier transform was to be performed on this transfer function, it is likely that it would be non-causal. After this, we derive formulas for calculating the optimal subband impulse response in the mean square error sense, given a fullband impulse response and a filterbank. Using these formulas, the optimal number of non-causal taps can be calculated if we have knowledge of the full-

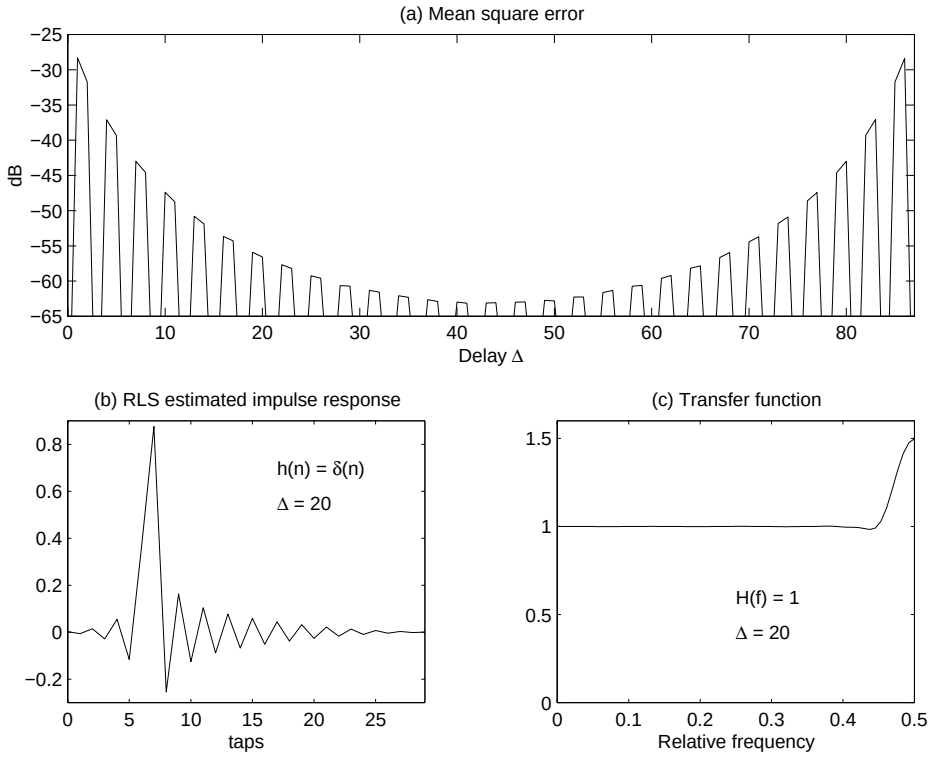

Figure 9: Like Fig. 8 with the exception that an RLS adaptive algorithm is used instead of an LMS.

band impulse response. Finally it is shown that due to the convergence properties of the LMS algorithm, the number of non-causal taps may need to be increased slightly, when the subband impulse responses are to be estimated with an LMS algorithm.

To conclude, when a subband system is designed and the impulse response of the system to be estimated has a short flat delay, the design needs to compensate for non-causal subband impulse response taps. This is especially important for systems where the LMS adaptive algorithm is used.

Acknowledgments

The authors would like to thank Dr Dennis R. Morgan, Bell Labs, and Dr Göran Salomonsson, Lund University, for constructive discussions.

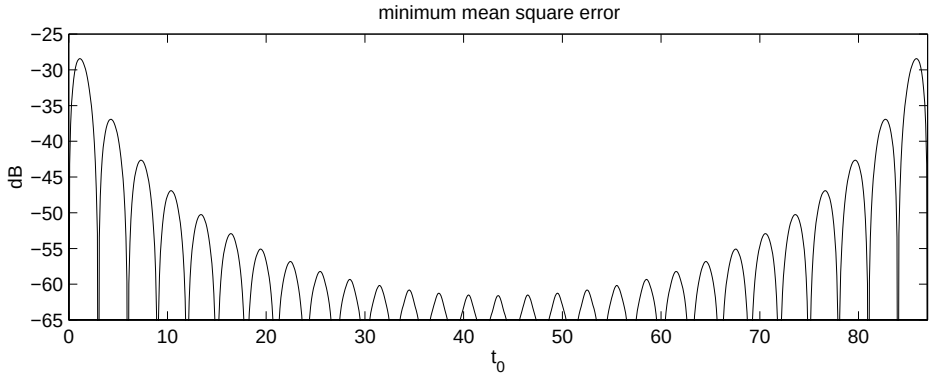

Figure 10: Minimum mean square error for the continues echo path $h(t) = \delta(t - t_0)$ as a function on t_0. Sample rate normalized to 1.

Appendix

A Correlation Calculations

The autocorrelation of $x_m(k)$ can be calculated as

$$\begin{aligned}\gamma_{x_m x_m}(l) &= \mathrm{E}\left\{x_m(k)x_m^*(k-l)\right\} \\ &= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} a_m(i)a_m^*(j)\mathrm{E}\left\{x(rk-i)x^*(rk-rl-j)\right\} \\ &= \sum_{i=\max(0,rl)}^{\min(L-1,L-1-rl)} a_m(i)a_m^*(i-rl)\sigma_x^2, \end{aligned} \tag{50}$$

where in the last step we have used the assumption of white noise so that $\mathrm{E}\{x(n)x^*(n-l)\} = \delta(l)\sigma_x^2$, σ_x is the standard deviation of $x(n)$, and δ is the Kronecker delta.

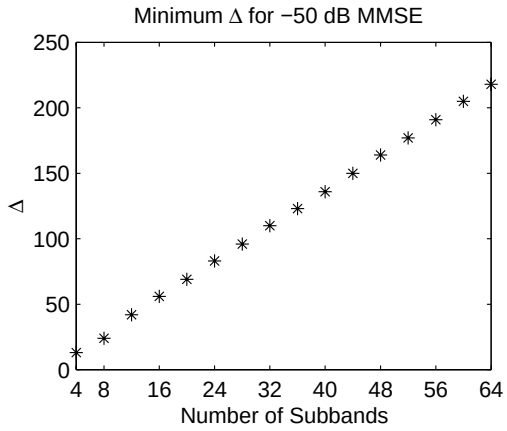

Figure 11: Minimum Δ as a function of number of subbands, M, for minimum mean square error < -50 dB. The impulse response is as before $h(n) = \delta(n)$, and the downsampling factor is $\frac{3}{4}M$.

Similarly, the cross-correlation between y_m and x_m is

$$\begin{aligned}
\gamma_{y_m x_m}(l) &= \mathrm{E}\left\{y_m(k)x_m^*(k-l)\right\} \\
&= \sum_{i=0}^{N-1}\sum_{j=i}^{L-1+i}\sum_{p=0}^{L-1} a_m(j-i)h(i)a_m^*(p) \\
&\quad \cdot \mathrm{E}\left\{[x(rk-\Delta-j)+v(rk-\Delta-j)]\,x^*(rk-rl-p)\right\} \\
&= \sum_{i=0}^{N-1}\sum_{j=\max(i,rl-\Delta)}^{A} a_m(j-i)h(i)a_m^*(j+\Delta-rl)\sigma_x^2, \qquad (51) \\
A &= \min(L-1+i, L-1+rl-\Delta),
\end{aligned}$$

where in the last step we have used the fact that the input signal $x(n)$ and the disturbance $v(n)$ are independent signals. We also need the autocorrelation of $y_m(k)$ to

determine the minimum filter error,

$$
\begin{aligned}
\gamma_{y_m y_m}(0) &= \mathrm{E}\left\{y_m(k) y_m^*(k)\right\} \\
&= \mathrm{E}\Bigg\{\Bigg[\sum_{i=0}^{N-1}\sum_{j=i}^{L-1+i} a_m(j-i)h(i)x(kr-\Delta-j) \\
&\quad + \sum_{s=0}^{L-1} a_m(s)v(kr-\Delta-s)\Bigg] \\
&\quad \cdot \Bigg[\sum_{p=0}^{N-1}\sum_{q=p}^{L-1+p} a_m^*(q-p)h^*(p)x^*(kr-\Delta-q) \\
&\quad + \sum_{t=0}^{L-1} a_m^*(t)v^*(kr-\Delta-t)\Bigg]\Bigg\} \\
&= \sum_{i=0}^{N-1}\sum_{p=0}^{N-1}\sum_{j=\max(i,p)}^{B} a_m(j-i)a_m^*(j-p)h(i)h^*(p)\sigma_x^2 \\
&\quad + \sum_{s=0}^{L-1}\sum_{t=0}^{L-1} a_m(s)a_m^*(t)\gamma_{vv}(t-s), \qquad (52) \\
B &= \min(L-1+i, L-1+p).
\end{aligned}
$$

The power spectral density function of the far-end subband signal $x_m(k)$ is found by Fourier transforming the autocorrelation $\gamma_{x_m x_m}$,

$$
\begin{aligned}
S_{x_m x_m}(f) &= \sum_{i=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_m(i)a_m^*(i-rl)\sigma_x^2 e^{-j2\pi fl} \\
&= \sum_{i=-\infty}^{\infty}\sum_{l'=-\infty}^{\infty} a_m(i)a_m^*(i-l')\left[\frac{1}{r}\sum_{d=0}^{r-1} e^{j\frac{2\pi l' d}{r}}\right]\sigma_x^2 e^{-j\frac{2\pi f l'}{r}} \\
&= \frac{\sigma_x^2}{r}\sum_{d=0}^{r-1}\sum_{i=-\infty}^{\infty} a_m(i)e^{j\frac{2\pi i(d-f)}{r}} \sum_{k=-\infty}^{\infty}\left(a_m(k)e^{j\frac{2\pi k(d-f)}{r}}\right)^* \\
&= \frac{\sigma_x^2}{r}\sum_{d=0}^{r-1}\left|A_m\left(\frac{f-d}{r}\right)\right|^2. \qquad (53)
\end{aligned}
$$

where $A_m(f)$ denotes the Fourier transform of the analysis filter a_m. Similarly the cross power spectral density is derived as,

$$
\begin{aligned}
S_{y_m x_m}(f) &= \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} a_m(k-i)h(i) \\
&\quad \cdot a_m^*(k+\Delta-rl)\sigma_x^2 e^{-j2\pi fl} \\
&= \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l'=-\infty}^{\infty} a_m(k-i)h(i)a_m^*(k+\Delta-l') \\
&\quad \cdot \left[\frac{1}{r}\sum_{d=0}^{r-1} e^{j\frac{2\pi l' d}{r}}\right]\sigma_x^2 e^{-j\frac{2\pi f l'}{r}} \\
&= \frac{\sigma_x^2}{r}\sum_{d=0}^{r-1}\sum_{i=-\infty}^{\infty} h(i)e^{-j\frac{2\pi i(f-d)}{r}} \sum_{k=-\infty}^{\infty} a_m(k-i)e^{-j\frac{2\pi(k-i)(f-d)}{r}} \\
&\quad \cdot \sum_{l'=-\infty}^{\infty} \left(a_m(k+\Delta-l')e^{-j\frac{2\pi(k+\Delta-l')(f-d)}{r}}\right)^* e^{-j\frac{2\pi\Delta(f-d)}{r}} \\
&= \frac{\sigma_x^2}{r}\sum_{d=0}^{r-1} H\left(\frac{f-d}{r}\right)\left|A_m\left(\frac{f-d}{r}\right)\right|^2 e^{-j\frac{2\pi\Delta(f-d)}{r}}. \qquad (54)
\end{aligned}
$$

References

[1] S. Haykin, *Adaptive Filter Theory*, Prentice Hall International, 1996.

[2] M. M. Sondhi and W. Kellermann, *Advances in Signal Processing*, chapter 11, pp. 327–356, Marcel-Dekker, 1992.

[3] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[4] W. Kellermann, *Zur Nachbildung physikalischer Systeme durch parallelisierte digitale Erzatzsysteme im Hinblick auf die Kompensation akustischer Echos*, Ph.D. thesis, Darmstadt, 1989.

[5] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Trans. on Signal Processing*, vol. 43, no. 8, pp. 1819–1830, Aug. 1995.

[6] G. Long, F. Ling, and J. G. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1397–1405, Sept. 1989.

[7] G. Long, F. Ling, and J. G. Proakis, "Correction to 'the LMS algorithm with delayed coefficient adaptation'," *IEEE Trans. on Signal Processing*, vol. 40, pp. 230–232, Jan. 1992.

[8] D. R. Morgan, "Slow asymtotic convergence of LMS acoustic echo cancelers," *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 2, pp. 126–136, Mar. 1995.

[9] M. Dörbecker and P. Vary, "Reducing the delay of an acoustic echo canceller with subband adaptation," in *The Int. Workshop on Acoust. Echo and Noise Control*, 1995, pp. 103–106.

[10] S. L. Gay, *Fast Projection Algorithms with Application to Voice Echo Cancellation*, Ph.D. thesis, State University of New Jersey, Oct. 1994.

[11] G. Strang and T. Nguyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1996.

[12] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, chapter 10, pp. 323–324, McGraw-Hill Inc., 3rd edition, 1991.

[13] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. on Information Theory*, vol. IT-18, pp. 725–730, Nov. 1972.

[14] D. R. Morgan and A. Aridgides, "Interpolation and extrapolation of an ideal band-limited random process," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 1, pp. 43–47, Jan. 1987.

Paper IV

Paper IV

Influence of Audio Coding on Stereophonic Acoustic Echo Cancellation

Abstract

Stereophonic acoustic echo cancellation has been found more difficult than echo cancellation in mono due to a high correlation between the two audio channels. Different methods to decorrelate the channels have been proposed so that the stereophonic echo canceller identifies the true echo paths and its convergence rate increases. In this paper it is shown that the use of a perceptual audio coder effectively reduces the correlation between the channels and thus convergence to the true echo paths is insured. Furthermore, in those frequency regions where the encoder introduced quantization noise is below the global perceptual masking threshold, an extra amount of inaudible noise can be added to the channels. Thereby the channel correlation is further decreased and the solution is stabilized. In subband audio coders with high frequency resolution only minor modifications are needed in the decoder.

Based on: Tomas Gänsler and Peter Eneroth "Influence of Audio Coding on Stereophonic Acoustic Echo Cancellation," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 3649–3652, May 1998.

This work was supported by Telia Research AB.

1 Introduction

Two emerging applications for stereophonic acoustic echo cancellation are high quality videoconferencing and tele-gaming. In the future, desktop based conference systems will also need stereophonic acoustic echo cancellers (SAEC). These systems have different quality demands influencing the bandwidth and bitrates etc.

Stereophonic acoustic echo cancellation, however, has been found far more complicated than its monophonic counterpart. This is due to the fact that the two channels carry linearly related signals, [1], which leads to convergence problems of the echo canceller. Due to the linear relation between the channels there is, in theory, no unique solution for the echo canceller to identify. Moreover, the non-unique solutions that exist are all dependent on the echo paths in the *far-end* (remote) room, [1], [2], [3]. In real situations, however, the solution to the problem is not truly singular only extremely ill-conditioned due to uncorrelated microphone noise and infinite impulse responses of the remote room's echo paths, [4], [5]. The convergence rate of the NLMS algorithm is highly dependent on the condition number of the correlation matrix thus more sophisticated algorithms must be used in stereophonic echo cancelling, e.g. [2], [6], [7].

Despite using more sophisticated algorithms there are still problems with unstable estimates of the echo paths, [3]. In order to stabilize the solution the correlation between the stereo channels has to be reduced without introducing annoying distortion. A number of solutions to this problem has been suggested, see e.g. [1], but rejected for different reasons. The most promising solution so far is to distort the stereo channels non-linearly as proposed in [5] where a half-wave rectified portion (α) of the signal is added to the signal itself. This distortion does not destroy the stereophonic perception but introduces a noise that most often is inaudible but may be perceived depending on the level of introduced non-linearity, [8].

The objectives for this paper is to study perceptual *audio coding* as another option to reduce the correlation between the channels. A perceptual audio coder, depending on bitrate etc., introduces a quantization noise that most often is below the hearing threshold. The question is if this distortion is strong enough in order to make the solution to the stereophonic acoustic echo canceller problem "well-conditioned."

2 Problem Formulation

The circumstances under which convergence to the true echo paths of an SAEC is achieved has been thoroughly analyzed in [5]. This section summarize some of their results that are used in this paper to formulate the problem and analyze the performance of the SAEC.

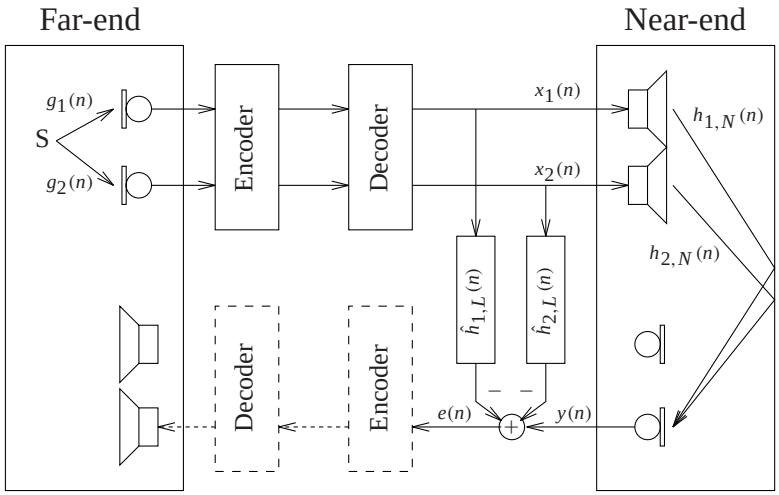

Figure 1: Audio coder and Stereophonic AEC. Only one return part is shown.

Assume that the far-end microphone signals are given by, Fig. 1,

$$x_i(n) = g_i(n) * s(n),\ i = 1, 2, \tag{1}$$

where $s(n)$ is the source signal and $g_i(n),\ i = 1, 2$ are the far-end echo paths of length M. "$*$" denotes convolution. The residual echo $e(n)$ after the EC is

$$e(n) = y(n) - \hat{\mathbf{h}}_{1,L}^T \mathbf{x}_{1,L} - \hat{\mathbf{h}}_{2,L}^T \mathbf{x}_{2,L} \tag{2}$$

$$y(n) = \mathbf{h}_{1,N}^T \mathbf{x}_{1,N}(n) + \mathbf{h}_{2,N}^T \mathbf{x}_{2,N}(n) \tag{3}$$

$$\mathbf{h}_{i,N} = [h_{i,0} \cdots h_{i,N-1}]^T, \tag{4}$$

$$\mathbf{x}_{i,N}(n) = [x_i(n) \cdots x_i(n - N + 1)]^T. \tag{5}$$

$\mathbf{h}_{i,N},\ i = 1, 2$ are the true responses of length N of the near-end room and $\hat{\mathbf{h}}_{i,L},\ i = 1, 2$ are the estimated responses of length L.

Minimization of the weighted least squares criterion

$$J(n) = \sum_{l=1}^{n} \lambda^{n-l} |e(n)|^2,\ 0 < \lambda \leq 1, \tag{6}$$

results in solving the system of linear equations

$$\mathbf{R}_{xx}(n) \begin{bmatrix} \hat{\mathbf{h}}_{1,L} \\ \hat{\mathbf{h}}_{2,L} \end{bmatrix} = \mathbf{r}_{yx}(n), \tag{7}$$

where $\mathbf{r}_{yx}(n)$ is the estimated cross-correlation vector and $\mathbf{R}_{xx}(n)$ is the correlation matrix,

$$\mathbf{R}_{xx}(n) = \sum_{l=1}^{n} \lambda^{n-l} \begin{bmatrix} \mathbf{x}_{1,L}(l)\mathbf{x}_{1,L}^T(l) & \mathbf{x}_{2,L}(l)\mathbf{x}_{1,L}^T(l) \\ \mathbf{x}_{1,L}(l)\mathbf{x}_{2,L}^T(l) & \mathbf{x}_{2,L}(l)\mathbf{x}_{2,L}^T(l) \end{bmatrix}. \tag{8}$$

The challenging problem with stereophonic echo cancelling all lies in the condition number of this matrix. It is also shown in [5] that,

$$\begin{aligned} L \geq M \quad &\Rightarrow \quad \mathbf{R}_{xx}(n) \text{ is singular } \forall n, \\ L < M \quad &\Rightarrow \quad \mathbf{R}_{xx}(n) \text{ is ill-conditioned }, \\ L \geq N \quad &\Rightarrow \quad \text{misalignment, } \varepsilon(n) \to 0,\ n \to \infty, \\ L < N \quad &\Rightarrow \quad \text{misalignment, } \varepsilon(n) \neq 0,\ \forall n, \end{aligned} \tag{9}$$

where the misalignment is $\varepsilon(n) = ||\mathbf{h} - \hat{\mathbf{h}}||^2/||\mathbf{h}||^2$ and $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_{1,L}^T\ \hat{\mathbf{h}}_{2,L}^T]^T$, $\mathbf{h} = [\mathbf{h}_{1,L}^T\ \mathbf{h}_{2,L}^T]^T$. An ill-conditioned $\mathbf{R}_{xx}(n)$ increases the misalignment in (9). Thus there is a contradiction, if $L << M$ the solution of (7) is better conditioned, on the other hand $L = N$ reduces misalignment, but practically $L < M \approx N$. The solution to this misalignment problem is therefore to *decrease the correlation between the stereo channels, thus reducing the condition number of* $\mathbf{R}_{xx}(n)$.

The eigenvalues of the correlation matrix can be lower bounded by $[1 - |\gamma(f)|^2]$, where $\gamma(f)$ is the coherence between the stereo channels [5]. Ill-conditioning can therefore be monitored by the coherence function which serves as a measure of the achieved decorrelation. The next section explains how decorrelation can be achieved by having a perceptual audio coder in the transmission path, Fig. 1.

3 Audio Coding

The Moving Picture Experts Group (MPEG) has developed two of the first international high-quality audio-visual coding standards, known as MPEG-1 and MPEG-2. Both these standards include high-quality stereophonic audio coders and MPEG-2 even includes multi-channel audio coding. These features are needed in todays and tomorrows interactive multi-media applications.

The MPEG-1 Layer III audio coder, the most advanced audio coder in MPEG-1, typically compresses stereophonic audio up to 12 times with insignificant audio quality loss. It is included in communications standards as H.310 Broadband Audio-visual Communications systems and H.323 Visual Telephone Systems and Equipment for Local Area Networks. The Layer III coder is also commonly used as a high-quality audio coder on the World Wide Web.

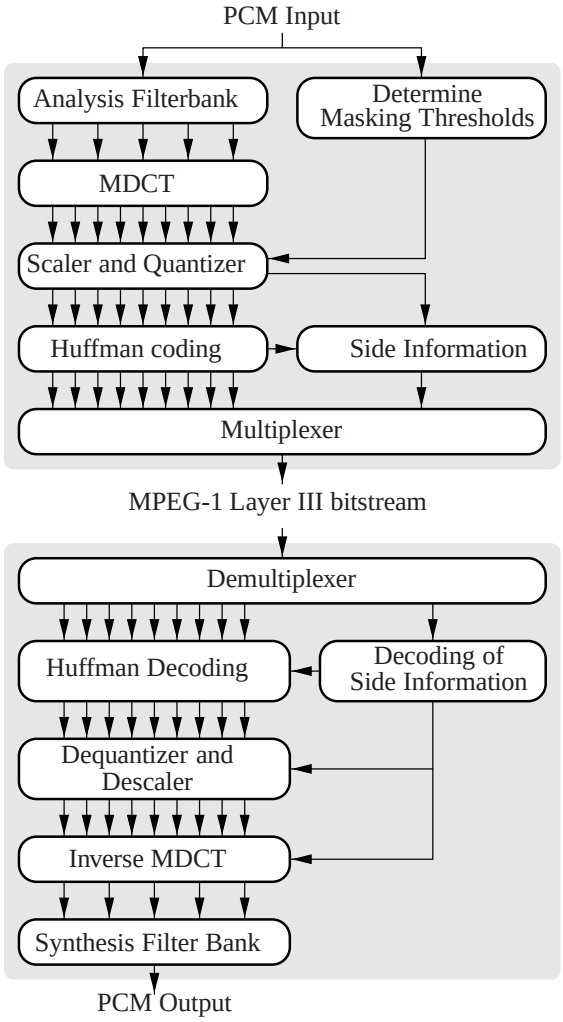

Figure 2: MPEG-1 Layer III encoder and decoder.

The high compression ratio is possible by removing components of the source signal that are perceptually irrelevant to the ear. In *Simultaneous masking*, a large frequency component will mask smaller ones in a nearby frequency band, whereas in *temporal masking*, components just before or right after (in the time domain) a large audio component are masked. Using this knowledge, the audio-encoder dynamically estimates the *global masking threshold*, that describe the just noticeable distortion as a function of frequency and time segment [9], Fig. 3.

The actual audio encoder operates in parallel with the global mask estimation algorithm. The audio source signal is decomposed into 32 critically downsampled bandpass signals by a filter bank. The frequency resolution is increased by processing each

bandpass signal with a Modified Discrete Cosine Transform (MDCT) in the Layer III coder. Depending on the input signal, each bandpass signal is decomposed in either 6 or 18 MDCT components, where the shorter window (generating 6 MDCT components) may be used during transients in the audio source. After this decomposition the MDCT components are scaled and quantized [10]. The main key in perceptual coders is to select enough quantization levels in the every subband, so that the introduced quantization noise level is below the global masking threshold. Data redundancy is reduced by Huffman coding the signal before transmitting it over the channel, Fig. 2. The decoder operates almost like an encoder in reverse, as illustrated in Fig. 2.

When the channels are not identical, the quantization noise introduced to the two channels are almost independent. As a result, the correlation between the two channels is decreased.

Correlation between the two channels can be decreased even more if independent noise is added to the channels. Due to large overhead, every single DCT-band cannot be optimally quantized. Instead they are divided into five regions, with specific numbers of quantization levels. Define quantization-noise to mask ratio (QMR) as the difference between the level of quantization noise and the level where distortion may become just audible in a given MDCT band. Then it is possible to add non-perceivable noise in those MDCT bands where QMR is positive. That is, for all MDCT bands in the frequency region where the channel correlation need to be reduced, perform

$$\begin{aligned} \mathrm{QMR}(j) > 0 \quad &\Rightarrow \quad \tilde{X}^j_{\mathrm{MDCT}} = X^j_{\mathrm{MDCT}} + f(\mathrm{QMR}(j)) \cdot v \\ \mathrm{QMR}(j) < 0 \quad &\Rightarrow \quad \tilde{X}^j_{\mathrm{MDCT}} = X^j_{\mathrm{MDCT}} \end{aligned}$$

where X^j_{MDCT} is the MDCT component in band j and $f(\cdot)$, given by the global masking threshold, amplifies the noise component v to be added, Fig. 3. A block implementing this channel decorrelation is added to the decoder right before the Inverse MDCT, Fig. 2. The global masking information is not available in the decoder, but because of the high frequency resolution of the MDCT, a simplified global masking estimate can be calculated with low complexity.

4 Measurement Studies

The influence audio coding has on the convergence of the SAEC is exemplified by simulations. The far-end speech is recorded in a room of size 330×500×272 cm, having a reverberation time of 0.3 seconds. The two far-end microphones are positioned 60 cm apart. Recordings were made using 48 kHz sampling rate.

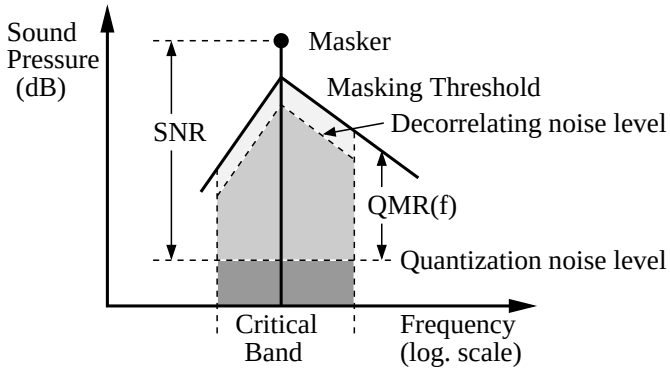

Figure 3: Masking Threshold: The grey area is masked by the tone.

To be able to access the misalignment of the estimated echo paths we use synthetic near-end room responses. The lengths of these responses are $N = 4096$ samples when the sample rate is 16 kHz and they have been estimated using data from a room of size $460 \times 670 \times 272$ cm. Echo cancellation is performed using a sample rate of 16 kHz where the length of the filters are $L = 2048$ each. No ambient near-end noise is added in the simulations.

As adaptive algorithm the two-channel FRLS, [11], is used, which is in principle the same algorithm as in [2] without numerical stabilization. The algorithm, [11], has been modified in order to remain stable when speech is used as input. The MPEG-1 Layer III coder used can be found in [12].

Results from four far-end cases, original recording, MPEG Layer III encoded/decoded speech, MPEG Layer III encoded/decoded with modified decoder, non-linearly modified far-end, $\alpha = 0.5$, are shown. The last case is shown as a reference since this technique has been proved effective, [5]. The MPEG coder is set to produce a bitrate of 192 kbits/s at a compression ratio of 8:1.

Possible ill-conditioning of the correlation matrix is indicated by the coherence function. Figure 4 shows the coherence between the far-end channels of the four cases. It is clearly seen that introduction of a non-linearity as well as audio coding results in smaller coherence, especially at higher frequencies. However, the coherence is still fairly close to one in lower frequencies. By modifying the decoding procedure according to Section 3 the coherence can be made smaller also in the lower frequencies without noticeable distortion, Fig. 4c.

Convergence of the FRLS algorithm is shown in Fig. 5. The convergence is presented in Fig. 5c as the Echo Return Loss Enhancement (ERLE) and in Fig. 5d as misalignment. In all cases the algorithm achieve the same high ERLE. The misalignment is on the other hand highly dependent on preprocessing of the far-end speech.

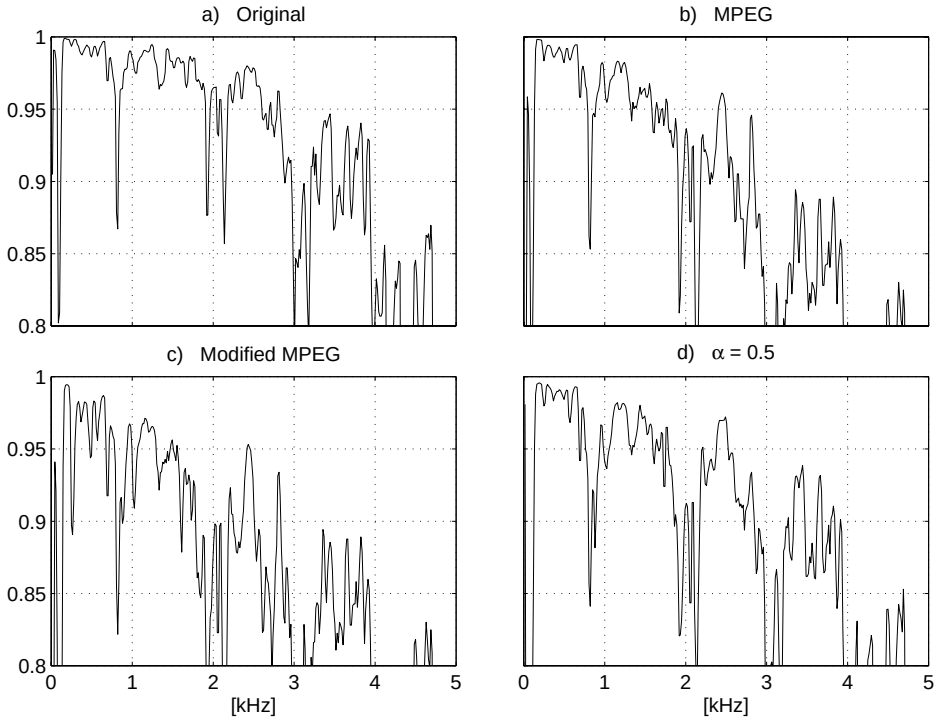

Figure 4: Coherence of the four studied cases.

5 Conclusions

A perceptual audio coder indeed improves the ability of an SAEC algorithm to converge to the true solution. Further improvement of the misalignment can be made by decoding the data utilizing the QMR-margin that often exists after coding. Very good perceptual quality is achieved while the complexity is maintained low since only a few simple operations need to be added in the decoder.

Acknowledgments

The authors thank N. Johansson, B. Rodger and O. Till, Telia Research AB, for supplying measured data and performing listening tests. J. Benesty should also be acknowledged for sharing his paper [5].

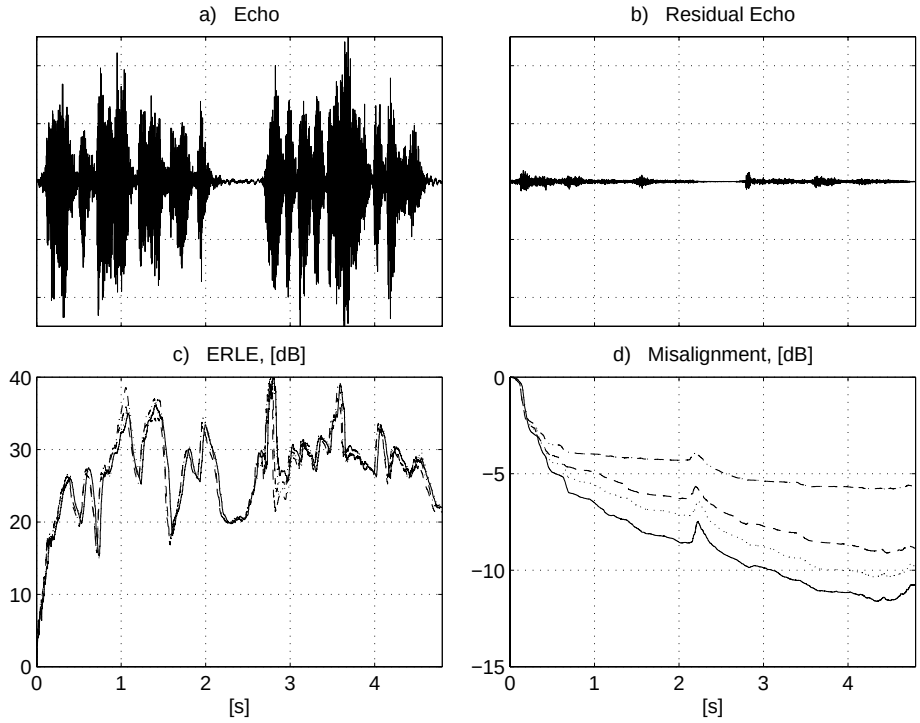

Figure 5: (a) Echo. (b) Residual echo. (c) ERLE. (d) Misalignment of the SAEC. Dashed-dotted line: Result from original far-end speech. Dashed line: Non-linearly modified. Dotted line: MPEG encoded/decoded. Solid line: MPEG encoded/decoded with modification. Each adaptive filter has 2048 taps.

References

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation — An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[2] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.

[3] S. L. Gay, "Algorithms for acoustic echo cancellation," Keynote Talk, The Int. Workshop on Acoust. Echo and Noise Control, Sept. 1997.

[4] F. Amand, A. Gilloire, and J. Benesty, "Identifying the true echo path impulse response in stereophonic acoustic echo cancellation," in *Proc. of EUSIPCO*, 1996, pp. 1119–1122.

[5] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[6] J. Benesty, P. Duhamel, and Y. Grenier, "A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation," *IEEE Signal Processing Lett.*, vol. 3, no. 2, pp. 35–37, Febr. 1996.

[7] S. Makino et. al., "Subband stereo echo canceller using the projection algorithm with fast convergence to the true echo path," in *Proc. of ICASSP*, 1997, pp. 299–302.

[8] O. Till et al., "Statusrapport 961215: Akustisk ekosläckning 7/0363-FCPA 109 0004," Tech. Rep., Telia Research AB, 1996.

[9] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, vol. 14, no. 5, pp. 59–81, Sept. 1997.

[10] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to* MPEG-*2*, chapter 4, pp. 55–79, Digital Multimedia Standards Series. Chapman & Hall, 1997.

[11] M. G. Bellanger, *Adaptive Digital Filters and Signal Analysis*, Marcel Dekker, 1987.

[12] Fraunhofer IIS, "MPEG-1 LAYER III shareware audio coder," 1995, Am Weichselgarten 3 D-91058 Erlangen Germany, encoder and decoder code: http://www.iis.fhg.de/amm/techinf/layer3/index.html, Public domain decoder source code (ANSI c): ftp://ftp.fhg.de/pub/iis/layer3/public_c/.

Paper V

Paper V

Joint Filterbanks for Echo Cancellation and Audio Coding

Abstract

In this paper, joint structures for audio coding and echo cancellation are investigated, utilizing standard audio coders. Two types of audio coders are considered, coders based on cosine modulated filterbanks and coders based on the modified discrete cosine transform (MDCT). For the first coder type, two methods for combining such a coder with a subband echo canceler are proposed. The two methods are: a modified audio coder filterbank that is suitable for echo cancellation but still generates the same final decomposition as the standard audio coder filterbank, and another that converts subband signals between an audio coder filterbank and a filterbank designed for echo cancellation. For the MDCT based audio coder, a joint structure with a frequency-domain adaptive filter based echo canceler is considered. Computational complexity and transmission delay for the different coder/echo canceler combinations are presented. Convergence properties of the proposed echo canceler structures are shown using simulations with real-life recorded speech.

Based on: P. Eneroth, "Joint Filterbanks for Echo Cancellation and Audio Coding," *IEEE Trans. on Signal Processing*, submitted January 2001.

1 Introduction

Echoes in telephone systems became severe with the introduction of long distance telephone services. In the 1960's, it was found that adaptive filtering was an efficient method to reduce the echoes caused by the electrical coupling in the 4 wire to 2 wire hybrids [1]. By means of adaptive filtering, it is possible to estimate the impulse response of this hybrid, and subtract an estimate of the echo from the return signal, thereby reducing the annoying echo. When handsfree communication systems like speaker phones and video conferencing became popular, a new source of echo was introduced. In these systems, the echo originates from the acoustic coupling between the loud-speaker and the microphone in the receiving room. The characteristics of acoustic echoes differ from echoes due to the electrical coupling in that the impulse responses are considerably longer. This results in a large increase of calculation complexity, since adaptive filters have a calculation complexity that is proportional to the adaptive filter length.

The calculation complexity of the long filters used in acoustic echo cancellation can be reduced by applying the adaptive filter in a subband filterbank structure [2]. In such a system, the signals are decomposed into several subband signals by an analysis filterbank. Then one adaptive filter in each subband suppresses the echo in the subband signal, before the fullband residual echo signal is reconstructed with a synthesis filterbank. If we have M subbands, and each subband is downsampled r times, M adaptive filters are needed. The calculation complexity reduction comes from the fact that each subband adaptive filter is r times shorter than the fullband filter and for r new fullband signal samples there will only be one new subband signal sample. That is, the calculation complexity of the adaptive filters will approximately be reduced by a factor M/r^2. The filterbank will generate some overhead calculations, but since efficient filterbank implementations exist, the adaptive filters are the major contributor of calculation complexity. Also, the convergence rate can be improved. For non-white signals, some frequency regions will have more signal energy than others, and accordingly, the signal correlation matrix will have large eigenvalues corresponding to frequency regions with strong signal energy and smaller eigenvalues corresponding to other regions. For certain adaptive filters, such as the normalized least mean square (NLMS) algorithm, the convergence rate is slow in frequency regions with small eigenvalues [3], [4]. In a subband system, the eigenvalue spread in each subband is reduced compared to the fullband signal and consequently, some adaptive algorithms, will perform better on non-white signals in a subband structure. Other advantages with subband structures are increased adaptive filter stability, due to shorter adaptive filters, and a structure that allows for efficient implementations on parallel systems, since the adaptive algorithms in each subband operate independently of one another. The two major disadvantages are the transmission path delay that is introduced and possible aliasing due to downsampling.

Another way to reduce the calculation complexity of the adaptive filter is to use a frequency-domain algorithm. In a frequency-domain algorithm based echo canceler, blocks of the signals are transformed into the frequency-domain with a discrete Fourier transform [5]. The echo transfer function is then estimated in the frequency-domain. As the case with subband echo cancelers, a frequency-domain adaptive filter (FDAF) algorithm usually achieves a fast convergence rate also in frequency regions with small correlation matrix eigenvalues. This is due to the fact that the adaptive filter in each frequency bin can have a normalization factor that corresponds to the energy of the signal in that frequency bin [6]. Being a block based adaptive algorithm, the transmission path delay is the biggest disadvantage of this method.

In a communication system, the audio signals need to be transmitted from one end to the other. As sampled speech and audio are extremely redundant, coders are commonly used to reduce the redundancy. One class of audio coders, denoted perceptual audio coders, uses a model of the human ear to determine which components of the sound that are audible to humans [7], [8]. Only audio components that the coder determines as audible should be coded and transmitted to the other end. The cochlea in the human ear actually acts as an octave band filterbank, dividing the sound into several subbands, and what is audible is mostly affected by the audio components within one frequency region. Subbands in the lower frequency region are narrower than subbands in the higher frequency regions, since it is an octave filterbank. Because efficient algorithms exist for linear filterbanks [9], some coder designs choose a linear filterbank when decomposing the signal. Examples of such coders are MPEG 1 and 2 audio layer 1, 2 and 3 [10], [11], where MPEG is an acronym for the Moving Picture Experts Group. In a system with both a subband echo canceler and a subband audio coder, it seems unnecessary to have two independent filterbanks. This paper describes methods for how to use one filterbank or how to combine the two filterbanks. In more recent perceptual audio coders, higher compression ratio has been made possible by increasing frequency resolution of the subband signals. In e.g. the Advanced Audio Coding (AAC) coder [12], an optional audio coder in MPEG 2, the subband filterbanks are exchanged for high resolution modified discrete cosine transforms (MDCTs). Since the MDCT has many similarities with the discrete Fourier transform used in FDAF, this paper will also investigate the possibility of joint transforms between a MDCT based audio coder and a FDAF based echo canceler.

The paper is organized as follows. In Section 2, the traditional filterbank design for echo cancellation is described. The cosine modulated QMF filterbank is also described, as this filterbank is commonly used in audio coders. In Section 3, we will show possible modifications in order to combine the two types of filterbanks. Then, in Section 4, we will switch our interest to audio coders based on the MDCT transform and FDAF based echo cancelers, and show possible joint designs for this type of systems. Finally, simulations, calculation complexity and signal transmission delay examples are given in Section 5.

2 Problem Formulation

In Fig. 1, a typical setup for a communication system including a filterbank based audio coder and subband echo canceler is shown. In such a system, data from the transmission side is received in a coded format, and the first step is to decode the received signal. Of interest here is the final stage in the decoder, namely a synthesis filterbank, which reconstructs the fullband audio signal from several subband signals. This is depicted as filterbank number 1 in Fig. 1. The cause of echo is the acoustic coupling between the received signal $x(n)$, and the signal to be transmitted $y(n)$. This coupling is denoted $h(n)$ in the figure. Any additional speech or noise in the receiving room is denoted $v(n)$. In a system without an echo canceler, $y(n)$ would be encoded with the audio coder, and the analysis filterbank of this encoder is denoted filterbank number 2 in the Fig. 1.

With an echo canceler, an estimated echo signal, $\hat{y}(n)$, is subtracted from the return signal, $y(n)$. The new return signal, now with less echo, is usually denoted the residual echo signal, $e(n)$ in the figure. In a subband echo canceler, all fullband signals are decomposed into downsampled narrow band signals. Echo cancellation is then performed on these signals, and the fullband residual echo signal, $e(n)$, is reconstructed from the subband residual echo signals, $e_m(n)$.

As is shown in Fig. 1, the system uses 5 filterbanks. These filterbanks not only have a calculation complexity cost, but what is worse, they also introduce transmission delay to the signals. The fundamental issue is that the aim of the audio coder and the echo canceler filterbank differ; the echo canceler cannot be applied directly to the subband signal from the audio coder's filterbank because it is common to have critical downsampling (hence alias) in the design. On the other hand, the echo canceler's filterbank is usually not efficient enough for coding purposes. Furthermore, the system designer usually has little control over the audio coder's filterbank whereas the echo canceler filterbank, which is independent of the rest of the system, can be freely designed. The question is now: can the subband decomposition done by the audio coder be modified and utilized by the echo canceler in order to reduce overall complexity, system delay, memory etc? To answer this question, we need to look in to some details regarding adaptive filter algorithms.

2.1 Normalized Least Mean Square Adaptive Algorithm

The NLMS is the most commonly used adaptive filter. Its strengths are robust behavior, and a structure that allows for simple implementation. The error signal and the

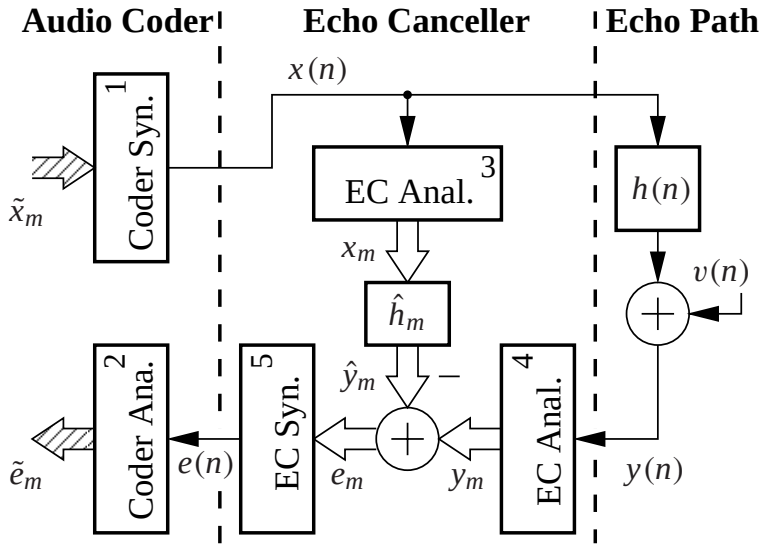

Figure 1: Subband echo canceler and model of acoustic echo path. In each subband, an adaptive algorithm estimates the subband impulse response $\hat{h}_m$, and the subband residual echo signal $e_m(k)$ is generated.

filter updates are calculated as [4],

$$\begin{aligned} e(n) &= y(n) - \underbrace{\mathbf{x}^H(n)\hat{\mathbf{h}}(n)}_{\hat{y}(n)}, \\ \hat{\mathbf{h}}(n+1) &= \hat{\mathbf{h}}(n) + \frac{\mu}{\mathbf{x}^H(n)\mathbf{x}(n) + e_{\text{reg}}}\mathbf{x}(n)e(n), \end{aligned} \tag{1}$$

where $\mathbf{x}(n)$ is the transmission room signal vector containing the latest L samples, $\hat{\mathbf{h}}(n)$ is the filter estimate vector, μ is the adaptive filter step size parameter, and e_{reg} is the regularization parameter. The Hermitian transpose is denoted ^{H}. The normalization factor $\mathbf{x}^H(n)\mathbf{x}(n)$ may be estimated with a low order recursive filter, and total calculation complexity is then of the order of $2L$ real-valued multiplications per sample for real-valued signals and $8L$ for complex valued signals. The complex valued version is usually needed in subband echo cancelers.

2.2 A Traditional Filterbank Design for Echo Cancelers

The purpose of the echo cancellation filterbank is to reduce the calculation complexity of the adaptive filters. The complexity reduction is proportional to the downsampling

factor of the subband signals, i.e. we desire as high a downsampling factor as possible. On the other hand, the convergence and tracking performance of the adaptive filters will be severely decreased if downsampling aliasing components are folded into the subband signals [13]. In order to minimize aliasing, the downsampling factor r is usually less than the number of subbands M^{EC}.

In an ordinary filterbank structure, all subband bandpass filters are modulated versions of a low pass prototype filter, $h^{\mathrm{EC}}(n)$, [14]

$$h_m^{\mathrm{EC}}(n) = h^{\mathrm{EC}}(n) e^{-j\frac{2\pi nm}{M^{\mathrm{EC}}}}, \tag{2}$$

where m denotes the subband number, and $^{\mathrm{EC}}$ is used as an acronym for echo canceler. Each subband signal can then be expressed as,

$$x_m(k) = [h_m^{\mathrm{EC}}(n) * x(n)] \downarrow r, \qquad n = kr, \tag{3}$$

where $*$ denotes convolution and $\downarrow$ downsampling. Because of symmetry between the subband filters, $h_m^{\mathrm{EC}}(n)$, efficient fast Fourier transform (FFT) based implementations are possible, [14], [15]. With a synthesis filterbank it is possible to reconstruct the fullband signal. The subband signals are first upsampled,

$$x_m(n) \uparrow r = \begin{cases} x_m(k), & n = rk, \\ 0, & n \neq rk. \end{cases} \tag{4}$$

Imaging, due to interpolation, is suppressed with bandpass filters, which are modulated versions of a prototype filter,

$$f_m^{\mathrm{EC}}(n) = f^{\mathrm{EC}}(n) e^{-j\frac{2\pi nm}{M^{\mathrm{EC}}}}. \tag{5}$$

In order to guarantee a linear phase response of the analysis/synthesis filterbank structure, the synthesis prototype filter, $f^{\mathrm{EC}}(n)$, is usually chosen as

$$f^{\mathrm{EC}}(n) = h^{\mathrm{EC}}(K^{\mathrm{EC}} - n - 1), \qquad 0 < n < K^{\mathrm{EC}} - 1, \tag{6}$$

where K^{EC} is the length of the prototype filters. The fullband signal can then be reconstructed with

$$x^{rec}(n) = \sum_{m=0}^{M^{\mathrm{EC}}-1} (x_m(n) \uparrow r) * f_m^{\mathrm{EC}}(n), \tag{7}$$

and as for the analysis filterbank, efficient implementations based on the FFT exist. In contrast to the pseudo quadrature mirror filter (QMF) and the perfect reconstruction filterbanks discussed in the next section, the echo canceler filterbank usually does not

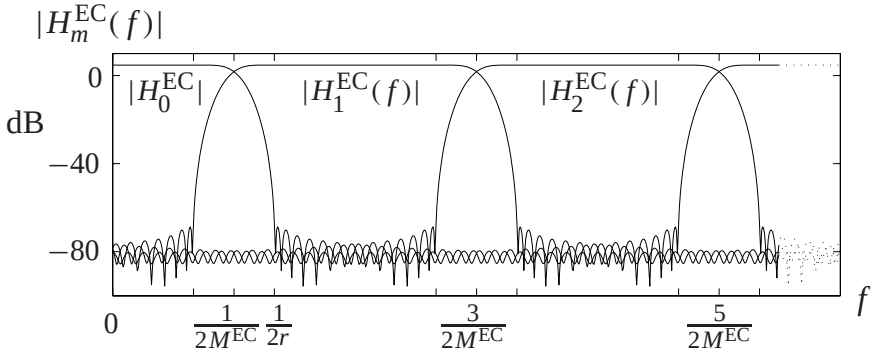

Figure 2: Amplitude response of the echo cancellation filterbank, with non-critical downsampling. The filters suppress aliasing by sufficient stopband attenuation; see filter $|H_0^{EC}(f)|$ for $f > \frac{1}{2r}$.

cancel aliasing. Instead, aliasing is appropriately suppressed by efficient stopband attenuation [14], [15]. In Fig. 2, the frequency responses of h_m^{EC}, $m = 0, 1, 2$, are shown, and it also illustrates the suppression of aliasing components. It should be noted that the subband signals, (3), are complex valued, i.e. a complex valued version of the adaptive filter is needed. Normally the fullband signals are real-valued, and it is therefore only necessary to cancel echoes in the lower $\frac{M^{EC}}{2} + 1$ subbands, whereas the upper subband signals can be reconstructed from the lower subbands, due to the symmetry in (2).

Efficient structures suitable for implementation, both of the analysis and synthesis filterbanks, are presented in [15]. In this structure, the filterbanks are composed of two parts, polyphase filtering and an FFT. For both the analysis and the synthesis filterbank, the number of real-valued multiplications needed for the filtering is equal to the prototype filter length K^{EC}. A Radix-2 implementation of an FFT needs $2M^{EC}\log_2 M^{EC} - 7M^{EC} + 12$ multiplications [16], where M^{EC} is block size. In [16] it is also shown how one FFT can be used to transform two real-valued signals, it only increases the complexity by $2M^{EC} - 4$ real-valued additions. The filterbanks need to be updated once per r input samples. Therefore the two analysis filterbanks and the synthesis filterbank in Fig. 1 can be realized using $\frac{1}{r}(2K^{EC} + 2M^{EC}\log_2 M^{EC} - 7M^{EC} + 12)$ and $\frac{1}{r}(K^{EC} + 2M^{EC}\log_2 M^{EC} - 7M^{EC} + 12)$ real-valued multiplications per fullband sample, respectively.

2.3 A Traditional Filterbank Design for Audio Coders

The purpose of an audio coder filterbank is to decompose the fullband signal into a set of low rate subband signals, which easily can be manipulated from a psycho-acoustic

point of view and allow for the coder to efficiently reduce redundancy in each subband signal. A non-critically downsampled filterbank would actually increase the information rate and thus reduce efficiency. Aliasing components exist in the subband signals, but these can be canceled in the synthesis filterbank. In the pseudo QMF filterbank used in the MPEG 1 and 2 audio coder [10], only aliasing from adjacent bands is cancelled, whereas in a perfect reconstruction filterbank [9], all aliasing components are canceled. The pseudo QMF filterbank can be designed to have a very high stopband attenuation, making reconstruction artifacts very small. If a lossy coder is used, such as the MPEG 1 audio coder, some of the properties necessary for perfect reconstruction is destroyed. That is, the advantage that a perfect reconstruction filterbank has over a pseudo QMF filterbank in audio coders is limited, the latter may even be better if it is properly designed [17].

In contrast to the echo canceler filterbank presented in the previous section, the subband signals are generally real-valued. The cosine function is used as modulator. For the pseudo QMF filterbank used in the MPEG 1 audio coder, the subband filters can be expressed as,

$$h_m^{\mathrm{AR}}(n) = h^{\mathrm{A}}(n)\cos\left(\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)\right), \qquad 0 < m < M^{\mathrm{A}}-1, \tag{8}$$

$$= \frac{1}{2}h^{\mathrm{A}}(n)\left(e^{-j\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)} + e^{j\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)}\right), \tag{9}$$

where $h^{\mathrm{A}}(n)$ denotes the prototype filter for the audio coder filterbank, and $h_m^{\mathrm{AR}}(n)$ the real-valued filter corresponding to subband m. Note how the filters are shifted in the frequency one half band with the constant $\frac{1}{2}$. The phase shift 16 can differ in different filterbanks, but only some phase values are valid for aliasing cancellation [9]. In (9) it is seen how this filterbank can be constructed as a sum of two exponential function modulated filterbanks. The real-valued subband $h_m^{\mathrm{AR}}(n)$ is constructed as the sum of two frequency shifted prototype filters, one that is shifted to the right and one to the left as $\pm(m+\frac{1}{2})$, and this is illustrated in Fig. 3. In the figure, the frequency response of $h^{\mathrm{A}}(n)e^{-j\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)}$ and $h^{\mathrm{A}}(n)e^{j\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)}$ are denoted $V_m(f)$ and $U_m(f)$, respectively. $V_m(f)+U_m(f)$ constitute the real-valued subband m. The subband signals can be expressed as in (3), with $r = M^{\mathrm{A}}$ and $h_m^{\mathrm{AR}}(n)$ instead of $h_m^{\mathrm{EC}}(n)$. Reconstruction of the fullband signal can be performed as in (7) with $f_m^{\mathrm{EC}}(n)$ interchanged to $f_m^{\mathrm{AR}}(n)$. The bandpass filter $f_m^{\mathrm{AR}}(n)$ is also created as a modulated version of a prototype filter,

$$f_m^{\mathrm{AR}}(n) = f^{\mathrm{A}}(n)\cos\left(\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n+16)\right), \qquad 0 < m < M^{\mathrm{A}}-1, \tag{10}$$

where $f^{\mathrm{A}}(n)$ denotes the prototype low pass filter used in the synthesis filterbank. Again, (10) corresponds to the filters used in the MPEG 1 audio coder.

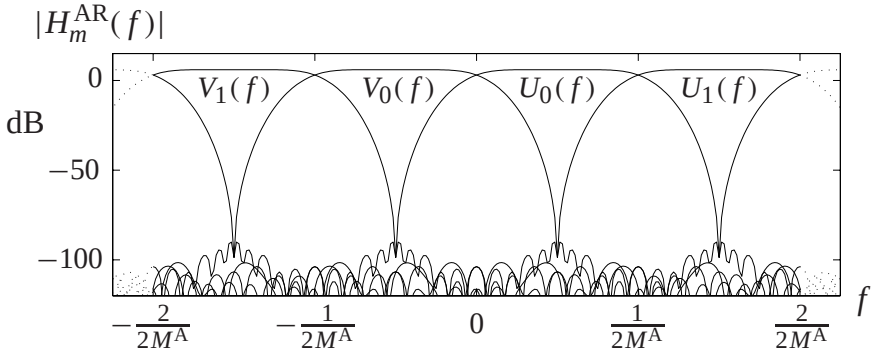

Figure 3: Filterbank used in MPEG 1 audio, usually denoted pseudo QMF filterbank. It has 32 real-valued subbands, and it is critically downsampled. Each subband consist of a positive and a negative part, denoted $U_m(f)$ and $V_m(f)$ respectively.

The pseudo QMF filterbank is critically downsampled and therefore significant aliasing exists in the subbands (see Fig. 3). Moreover, the pseudo QMF synthesis filterbank cancels alias components from adjacent subbands which will be shown in more detail in a later section (see also [9]). This makes the pseudo QMF filterbank unsuitable to use in a subband echo canceler. Not only will the aliasing drastically decrease the performance of the adaptive filter [13], but the adaptive filtering will also make alias cancellation to be performed by the synthesis filterbank impossible. One solution presented in [13] is to let the adaptive filter have special cross-band filters, but it is also concluded in [13] that this solution has a very high calculation complexity, and that the adaptive filters have a slow convergence rate.

An efficient structure for the pseudo QMF filterbank is given in [9]. As for the EC filterbank in the previous section, the implementation includes polyphase filtering and a transform, namely a discrete cosine transform (DCT). The polyphase filtering needs K^{A} real-valued multiplications, where K^{A} is the length of the prototype filter, and the DCT needs $\frac{M^{\mathrm{A}}}{2}\log_2 M^{\mathrm{A}}$ real-valued multiplications [18]. That is, both the analysis and synthesis filterbank need $\frac{1}{M^{\mathrm{A}}}(K^{\mathrm{A}} + \frac{M^{\mathrm{A}}}{2}\log_2 M^{\mathrm{A}})$ real-valued multiplications per fullband sample each.

3 Modified Audio Coder Filterbank Structures for Echo Cancelers

In this section we will investigate the possibilities of using a modified version of the pseudo QMF filterbank, described in the previous section, for echo cancellation. We

consider the coding filterbank to be pre-specified, i.e., we need to be able to reconstruct a final output, $\tilde{e}_m$ in Fig. 1, which is identical to the output of the audio coder filterbank. We will also describe how the subband signals can be converted between the traditional echo canceler filterbank and the pseudo QMF filterbank.

3.1 An Oversampled Real Valued Pseudo QMF Filterbank

One possible modification to the critically downsampled cosine modulated filterbank is reduction of the downsampling factor. The hypothesis would be to use a downsample factor of $r = M^{\mathrm{A}}/2$ before echo cancellation. The output of the echo canceler should then be decimated by 2. However, this will not work, since we have significant alias components in some subbands. Actually, significant alias components arise even when we decimate with the smallest possible decimation factor, $r = 2$. Let us study subband number $m = M^{\mathrm{A}}/2$ in a M^{A} band filterbank, where each subband is decimated by a factor 2. Remember that cosine modulated real-valued filterbanks have one positive and one negative frequency part, as is shown in (9) and in Fig. 3. For $m = M^{\mathrm{A}}/2$, the subband filter, (8), will have a passband in the frequency regions $-\frac{2m+3}{4M^{\mathrm{A}}} < f < -\frac{2m-1}{4M^{\mathrm{A}}}$ and $\frac{2m-1}{4M^{\mathrm{A}}} < f < \frac{2m+3}{4M^{\mathrm{A}}}$. The frequency-domain formula for decimating a factor r can be expressed as [19],

$$Y(f) = \frac{1}{r}\sum_{i=0}^{r-1} X\left(\frac{f-i}{r}\right). \tag{11}$$

If we decimate the signal in subband m by a factor two, using (11), we see that we will have alias components in the frequency regions $-\frac{1}{2} < f < -\frac{2m-1}{2M^{\mathrm{A}}}$ and $\frac{2m-1}{2M^{\mathrm{A}}} < f < \frac{1}{2}$. Therefore, all oversampled pseudo QMF filterbanks, except for a filterbank operating without downsampling, will have significant alias components and will perform poorly in a echo canceler scheme.

3.2 An Oversampled Complex Valued Pseudo QMF Filterbank

If we once again return to (9), we can find a better modification, resulting in a non-critically downsampled filterbank without significant aliasing components. Instead of adding the two complex conjugate terms together as in (9), we can construct a $2M^{\mathrm{A}}$ complex valued filterbank with the following modulation scheme,

$$h_m^{\mathrm{AC}}(n) = \frac{1}{2}h^{\mathrm{A}}(n)e^{-j\frac{\pi}{M^{\mathrm{A}}}(m+\frac{1}{2})(n-16)}, \qquad 0 < m < 2M^{\mathrm{A}} - 1, \tag{12}$$

where ^{AC} denotes a complex valued version of the audio coder filterbank, defined in (8). We will continue to use (3) with $r = M^{\mathrm{A}}$ and $h_m^{\mathrm{EC}}(n)$ exchanged for $h_m^{\mathrm{AC}}(n)$,

i.e., we have $2M^{\mathrm{A}}$ subbands with a downsampling factor of M^{A}. As can be seen in Fig. 3, if we apply this modulation scheme to the filterbank used in the MPEG 1 audio coder, all alias components will be suppressed by almost 100 dB. We would exchange the filterbanks 3 and 4 in Fig. 1 with this complex valued version of the audio coder filterbank.

The reconstruction of $\tilde{e}_m$ from e_m, performed by filterbanks 5 and 2 in Fig. 1, can be replaced by the trivial operation of adding 2 complex conjugate subband signals together, as

$$\tilde{e}_m(k) = e_m(k) + e_{2M^{\mathrm{A}}-m-1}(k), \qquad 0 < m < M^{\mathrm{A}} - 1, \tag{13}$$
$$= e_m(k) + e_m^*(k), \tag{14}$$
$$= 2\mathrm{Re}\{e_m(k)\}. \tag{15}$$

That is, we have reduced the $2M^{\mathrm{A}}$ subband complex valued filterbank used for echo cancellation to a M^{A} subband real-valued filterbank which is identical to the filterbank used in the audio coder, (8). For both filterbanks we use a downsampling factor of M^{A}. The total delay of the signal path, from $\tilde{x}_m$ to $\tilde{e}_m$ in Fig. 1, is now reduced to the delay of filterbank 1 and 4. The next questions are; can we combine filterbank 1 and 3 in Fig. 1, and if we can, what would we gain?

The gain is actually less obvious. We will of course still need filterbank 1 in order to reconstruct $x(n)$, needed for the receiving room, and we will need filterbank 4. The signal path delay is now determined by filterbank 1 and 4, plus a small extra delay needed in order for the adaptive filters to be able to estimate a few non-causal taps, usually needed in subband echo cancellation [2], [20]. If we, by combining filterbank 1 and 3, could reduce the delay of signal $x_m(k)$, the only signal delay reduction would be to compensate for a few non-causal taps in each subband. Another possible gain would be less calculation complexity.

One of the most important property of filterbank 1 in Fig. 1 is to cancel aliasing. If we are to combine filterbank 1 and 4, we need to be able to perform this cancellation. Therefore, we must study how alias cancellation is performed, i.e., how aliasing terms in adjacent subband cancel each other in the reconstruction. In Fig. 4, the frequency response of subband m is shown. The gray areas show how U_m and V_m overlap after downsampling which is the cause of aliasing. If we were to reconstruct subband 1, we could try to modify (7) to only include subband 1,

$$\check{x}_1^{rec}(n) = (\tilde{x}_1(n) \uparrow M^{\mathrm{A}}) * f_1^{\mathrm{AR}}(n), \tag{16}$$

where $\uparrow$ is defined in (4). The frequency response of $\check{x}_1^{rec}(n)$ is shown in Fig. 5. We will have significant aliasing components, illustrated by the gray areas in Fig. 5. The two gray areas centered around $\pm\frac{1}{2M^{\mathrm{A}}}$ in Fig. 5 cover a subinterval of the frequency range of subband 0, $U_0(f) + V_0(f)$, shown in Fig. 3. In the same way the two gray

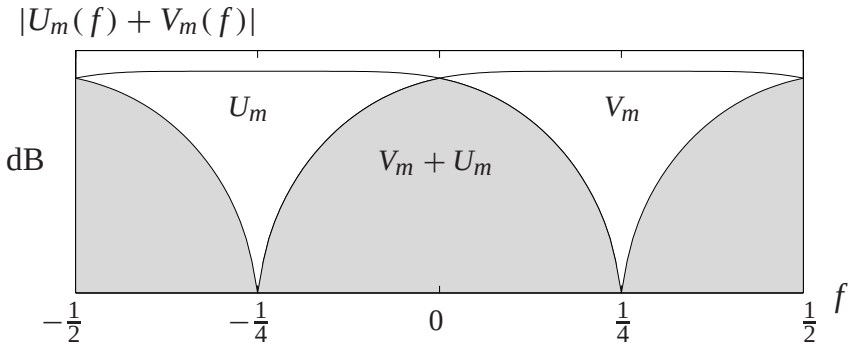

Figure 4: The frequency response of subband m, m odd, after filtering and downsampling of the pseudo QMF filterbank used in MPEG 1. The gray areas represent aliasing. For even m, U_m and V_m will exchange position, see (11).

areas centered around $\pm\frac{2}{2M^{\mathrm{A}}}$ covers a subinterval of the frequency range of subband 2, $U_2(f) + V_2(f)$. The pseudo QMF filterbank is designed in such way that, that the gray areas in Fig. 5 are cancelled when $\check{x}_0^{rec}(n)$ and $\check{x}_2^{rec}(n)$ are added to $\check{x}_1^{rec}(n)$. That is, in order to reconstruct the complex valued subband 1 we need to perform the following operation,

$$x_1^{rec}(n) = \left[\left(\sum_{m=0}^{2} (\tilde{x}_m(n) \uparrow M^{\mathrm{A}}) * f_m^{\mathrm{AR}}(n) \right) * h_1^{\mathrm{AC}}(n) \right] \downarrow M^{\mathrm{A}}. \tag{17}$$

The convolution with $f_m^{\mathrm{AR}}(n)$ is a necessary condition for alias cancellation and the convolution with $h_1^{\mathrm{AC}}(n)$ is necessary in order to suppress the frequency areas outside of subband 1. As is shown in Section 3.1, the upsample factor of $\tilde{x}_m(n)$ in (4) needs to be M^{A} in order to guarantee alias free subband signals. This shows that we cannot gain any reduction in neither signal path delay nor calculation complexity by combining filterbank 1 and 3 in Fig. 1.

The use of a complex valued version of a QMF filterbank, e.g., a modified MPEG 1, 2 layer 1, 2 and 3 audio coder filterbank, has two major disadvantages compared with a specially designed echo canceler filterbank. First of all, the passband region in the QMF filterbank is larger. This will increase the eigenvalue spread, and therefore decrease the convergence rate of an NLMS adaptive filter. Examples of this are given in Section 5. The second disadvantage concerns calculation complexity. Since we have $2M^{\mathrm{A}}$ subbands and the downsampling factor is only M^{A}, the adaptive filters will have a higher computational complexity than in a specifically designed echo canceler filter bank, where the downsampling factor typically could be $3M^{\mathrm{EC}}/4$ for a system with M^{EC} subbands. Nevertheless, it should be remembered that using a complex valued version of a pseudo QMF filterbank in combination with a pseudo QMF filterbank

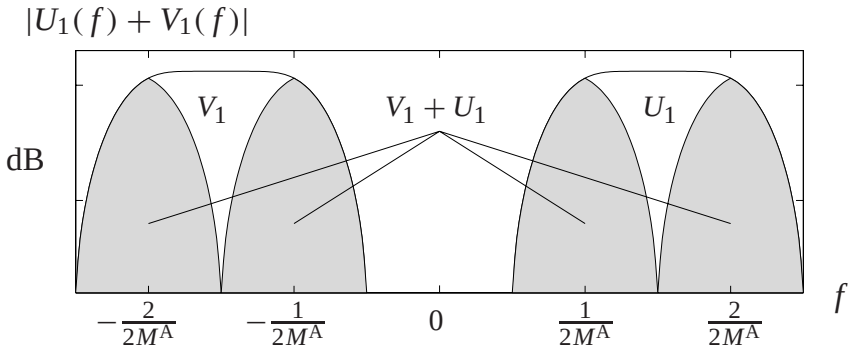

Figure 5: Frequency response of subband 1 after upsampling and suppression of imaging, given by $\check{x}_1^{rec}(n)$ in (16).

based audio coder, will only increase the signal transmission delay by a small value corresponding to a few non-causal taps needed in each subband [2], [20]. In Section 5, an example shows that this delay is 16 ms.

The filterbank given by (12) can be realized in the same way as the filter bank in Section 2.2. This filterbank will be used to replace filterbanks 3 and 4 in Fig. 1, i.e., we will need two filterbanks. From Section 2.2 we find that, by replacing M^{EC} with $2M^{\mathrm{A}}$ and r with M^{A}, the two filterbanks will need $\frac{1}{M^{\mathrm{A}}}(2K^{\mathrm{A}}+4M^{\mathrm{A}}\log_2 2M^{\mathrm{A}}-14M^{\mathrm{A}}+12)$ real-valued multiplications per fullband sample. Filterbanks 2 and 5 in Fig. 1 will then be replaced with (15), requiring no multiplications. It should also be remembered that we only need adaptive filters in the M^{A} lower of the $2M^{\mathrm{A}}$ subbands.

3.3 A Joint Filterbank Structure for Audio Coding and Echo Cancellation

As is discussed above, the use of the complex valued version of a pseudo QMF filterbank for echo cancellation has two disadvantages. The calculation complexity could be reduced in a filterbank with a larger downsampling factor, which also improves the convergence rate of the adaptive filter. Therefore, in this section we will examine the possibilities of a joint audio coder and echo canceler structure, where the echo canceler uses the filterbank described in Section 2.2 and the audio coder the filterbank in Section 2.3. We will use $M^{\mathrm{EC}} = 64$ subbands with a downsampling factor of $r = 48$ for the echo canceler filterbank and a $M^{\mathrm{A}} = 32$ subbands critically downsampled audio coder filterbank. The frequency response of the lower subbands of the two filterbanks are shown in Fig. 6.

First we study possibilities of combining filterbanks 1 and 3 in Fig. 1. In Section 3.2, it is shown that the only delay reduction possible is a reduction of the delay

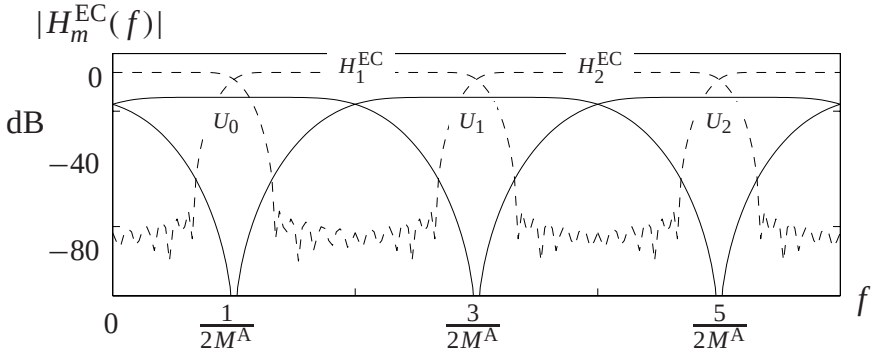

Figure 6: Frequency response of the lower subbands of the two filterbanks. The audio coder's QMF filterbank is illustrated with a solid line, and the EC filterbank with a dashed line.

that will compensate for a few non-causal taps needed in each subband [2], [20]. Also, Section 3.2 describes how alias cancellation is performed. Let us examine how to reconstruct echo canceler subband number 2, i.e. $x_2(k)$, directly from the audio coder subband signals, $\tilde{x}_m(k')$, $0 \le m \le M^{\mathrm{A}}$. In Fig. 6, it is apparent that we need the audio coder subbands 0 to 3 in order to reconstruct an alias cancelled signal that spans the same frequency regions as $H_2^{EC}(f)$, i.e.,

$$x_2(k) = \left[\left(\sum_{q=0}^{3} (\tilde{x}_q(k') \uparrow M^{\mathrm{A}}) * f_q^{\mathrm{AR}}(n) \right) * h_2^{\mathrm{EC}}(n) \right] \downarrow r. \tag{18}$$

If we neglect the small alias components from audio coder subbands $U_0(f)$ and $U_3(f)$ the sum only needs two terms. We can also move the filter $h_2^{\mathrm{EC}}(n)$ to be performed inside the sum,

$$x_2(k) = \left[\sum_{q=1}^{2} (\tilde{x}_q(k') \uparrow M^{\mathrm{A}}) * \underbrace{\left\{ f_q^{\mathrm{AR}}(n) * h_2^{\mathrm{EC}}(n) \right\}}_{g_{2,q}(n)} \right] \downarrow r. \tag{19}$$

If we for each subband create the new filters $g_{2,q}(n) = f_q^{\mathrm{AR}}(n) * h_2^{\mathrm{EC}}(n)$, the total signal transmission delay is given by $g_{2,q}(n)$. In order to decrease the delay, we must redesign this filter. This filter has a stopband attenuation equal to the sum of the attenuation of $f_q^{\mathrm{AR}}(n)$ and $h_2^{\mathrm{EC}}(n)$ which is more than required. We can therefore relax the design constraint on $g_{2,q}(n)$ by altering its passband region and reducing its

length. Let us form the new shorter filter,

$$\tilde{g}_{2,q}(n) = \begin{cases} g_{2,q}(n), & R \le n \le K - R - 1, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

where K is the length of $g_{2,q}(n)$, and $2R$ is the reduction in length we would like to have. Only the non-zero coefficients of $\tilde{g}_{2,q}(n)$ are then used in reconstruction of the echo canceler subbands, and thereby the filterbank delay is reduced. Now we need to find the filter coefficients in $\tilde{g}_{2,q}(n)$ such that the frequency response in the passband is nearly the same as for $g_{2,q}(n)$. We can achieve this by minimizing,

$$\Phi_1 = \int_{f \in \text{passband}} |G(f) - \tilde{G}(f)|^2 df, \tag{21}$$

where $G(f)$ and $\tilde{G}(f)$ are the Fourier transform of $g_{2,q}(n)$ and $\tilde{g}_{2,q}(n)$, respectively. We also need to make sure the stopband suppression is sufficient by minimizing

$$\Phi_2 = \int_{f \in \text{stopband}} |\tilde{G}(f)|^2 df. \tag{22}$$

The total minimization problem can now be expressed as

$$\min_{\tilde{g}(n)} \quad \alpha_1 \Phi_1 + \alpha_2 \Phi_2, \tag{23}$$

where $\alpha_i \ge 0$ are trade-off parameters.

In contrast to the combination of filterbanks 1 and 3 in Fig. 1, where the signal $x(n)$ is needed for the receiving room, we do not need to reconstruct the signal $e(n)$ if we are to combine filterbanks 2 and 5. That is, we could possibly decrease the delay significantly. We will start to study the reconstruction of $\tilde{e}_1(n)$, by studying the frequency region corresponding to the region of filter $V_1(f)$ and $U_1(f)$ in Fig. 3. From Fig. 6, it can be seen how $U_1(f)$ is covered by the filters $H_1^{\text{EC}}(f)$ and $H_2^{\text{EC}}(f)$. Similarly, $V_1(f)$ is covered by the filters $H_{M^{\text{EC}}-1}^{\text{EC}}(f)$ and $H_{M^{\text{EC}}-2}^{\text{EC}}(f)$. But since $e_m(k) = e^*_{M^{\text{EC}}-m-1}(k)$, we can reconstruct the frequency region that covers $V_1(f)$ by taking the real value of the signal. We also know from Section 2.2 that we do not need to cancel aliasing. The audio coder subband $\tilde{e}_1^{rec}(k')$ can therefore be reconstructed as,

$$\tilde{e}_1(k') = \left[2 \cdot \text{Re} \left\{ \sum_{q=1}^{2} (e_q(k) \uparrow r) * f_q^{\text{EC}}(n) \right\} * h_1^{\text{AR}}(n) \right] \downarrow M^{\text{A}}. \tag{24}$$

For subband 1 it is not necessary to increase the sampling rate to the rate of the full-band signal. Actually, it is enough to increase the sampling rate by a factor 3 because

a signal with the bandwidth of $F_1^{\mathrm{EC}}(f)$ plus $F_2^{\mathrm{EC}}(f)$ can be represented by a signal with downsampling factor $r/3 = 16$. This can be seen in Fig. 6, as we can use the fact that $H_q^{\mathrm{EC}}(f)$ and $F_q^{\mathrm{EC}}(f)$ have the same amplitude response, see (6). In order to avoid aliasing, the creation of the real-valued signal need to be performed last, and the complex valued version of the $h_1^{\mathrm{AR}}(n)$ has to be used,

$$\tilde{e}_1(k') = 2 \cdot \mathrm{Re}\left\{ \sum_{q=1}^{2} (e_q(k) \uparrow 3) * \tilde{f}_q^{\mathrm{EC}}(n') * \tilde{h}_1^{\mathrm{AC}}(n') \right\} \downarrow 2, \tag{25}$$

where $\tilde{f}_q^{\mathrm{EC}}(n')$ and $\tilde{h}_1^{\mathrm{AC}}(n')$ denote $f_q^{\mathrm{EC}}(n) \downarrow (r/3)$ and $h_1^{\mathrm{AC}}(n) \downarrow (r/3)$, respectively. By studying Fig. 4 we realize that (25) can be generalized for reconstruction of all subbands. We just need to modulate each input subband signal to the correct frequency region, according to

$$\tilde{e}_m(k') = 2 \cdot \mathrm{Re}\left\{ \sum_{q=1}^{2} (e_q(k) e^{j\kappa_{m,q}k} \uparrow 3) * \tilde{f}_q^{\mathrm{EC}}(n') * \tilde{h}_1^{\mathrm{AC}}(n') \right\} \downarrow 2, \tag{26}$$

where $\kappa_{m,q}$ is the modulation factor, individual for each subband. By creating one filter from of $\tilde{f}_m^{\mathrm{EC}}(n')$ and $\tilde{h}_1^{\mathrm{AC}}(n')$, like in (19), we can reduce the total length the same way as was done in equations (21) to (23).

It should be noted that even if it is possible reduce the delay, by using these reconstruction methods, the calculation complexity will increase. This since we will not be able to use the efficient structures available for the exponential modulated echo canceler filterbanks or the cosine modulated QMF filterbank.

4 FDAF and MDCT Based Audio Coders

In order to achieve higher compression ratio in audio coders, high frequency resolution is advantageous. This can be achieved by exchanging the filterbank presented in Section 2.3 for a modified discrete cosine transform (MDCT). As this transform has several similarities with the discrete Fourier transform, which is used in frequency-domain adaptive filtering, we will in this section investigate the possibilities of reducing signal transmission delay and calculation complexity by combining an audio coder based on the MDCT transform with an echo canceler using a frequency-domain adaptive filter.

4.1 Frequency-Domain Adaptive Filtering

In this section, the frequency-domain adaptive filter (FDAF) [5], [4] is explained. The basic aim of the FDAF is to reduce the calculation complexity and to increase the convergence rate (compared to the classical time domain NLMS adaptive filter) for non-white input signals by performing the adaptive filtering in the frequency-domain.

The input signals are partitioned in blocks, and the transmission room signal $x(n)$ is transformed with a discrete Fourier transform into the frequency-domain,

$$\mathbf{X}(k) = \text{diag}\left\{\mathbf{F}\left[x(kN-N)\ldots x(kN+N-1)\right]^T\right\}, \tag{27}$$

$$\mathbf{y}(k) = \left[y(kN-N)\ldots y(kN+N-1)\right]^T, \tag{28}$$

where N is the block size of the FDAF, $\mathbf{X}(k)$ is a $(2N \times 2N)$ matrix and $\mathbf{y}(k)$ is a $(N \times 1)$ matrix. The Fourier matrix $\mathbf{F}$ can be expressed as,

$$[\mathbf{F}]_{(p,q)} = \frac{1}{\sqrt{2N}} e^{-j\frac{2\pi pq}{2N}}, \quad p,q \in [0,\ldots,2N-1], \tag{29}$$

where $[\mathbf{F}]_{(p,q)}$ denotes element (p,q) in the matrix $\mathbf{F}$. The actual residual error can be formed in the time domain as,

$$\mathbf{e}(k) = \mathbf{y}(k) - \underbrace{\left[\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}\right]\mathbf{F}^{-1}\mathbf{X}(k)\hat{\mathbf{H}}(k)}_{\hat{\mathbf{y}}(k)}, \tag{30}$$

where $\mathbf{I}_{N\times N}$ denotes the identity matrix of size N. In (30), the last N samples of the vector $\mathbf{F}^{-1}\mathbf{X}(k)\hat{\mathbf{H}}(k)$ are the estimated time domain echo signal, $\hat{\mathbf{y}}(n)$. Only the last N samples are used since multiplication of two discrete Fourier transformed variables corresponds to a circular convolution of the time domain variables. Finally, a zero padded and transformed version of the residual error signal, $\mathbf{e}(n)$, is used to update the transfer function estimate, $\hat{\mathbf{H}}(k)$, as,

$$\mathbf{E}(k) = \mathbf{F}\left[\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}\right]^T \mathbf{e}(k), \tag{31}$$

$$\hat{\mathbf{H}}(k+1) = \hat{\mathbf{H}}(k) + \mu\mathbf{X}^H(k)\mathbf{E}(k), \tag{32}$$

where μ is the step size parameter. A complete version of the FDAF algorithm is given in Table 1. In this version the step size parameter is individually normalized in each frequency bin, and it is possible to have several filter taps in each frequency bin. An analysis of the FDAF can be found in [5], [6], and a multi-channel FDAF can be found in [21], [22].

It should be noted that it is always possible to update the transfer function estimate $\hat{\mathbf{H}}(k)$ more frequently than once per block by letting input data overlap. This will

improve the convergence rate, and examples are given in Section 5. The number of updates per N samples is denoted α, e.g., without input data overlap $\alpha = 1$, and for 25% new data per block $\alpha = 4$. The calculation complexity of the algorithm is increased by a factor α.

The unconstrained version of the FDAF algorithm given in Table 1 is only usable for $P = 1$ filter tap per frequency bin. This version needs $20NP + 8N$ real-valued multiplications, $8NP$ real-valued divisions, 2 FFTs with real-valued input signals and 1 inverse FFT (IFFT) with a real-valued output signal. Like before, we assume that we use a Radix 2 FFT, and that two real-valued signals can be transformed with one FFT. We also assume that the cost of one real-valued division equals 4 multiplications. The number of real-valued multiplications is then $\frac{\alpha}{N}\{60N + 2(2N\log_2 N - 7N + 12)\}$ per fullband input sample. For the constrained version, an additional P complex valued FFTs and P complex valued IFFTs are needed. The calculation complexity is therefore $\frac{\alpha}{N}\{52NP + 8N + (2 + 2P)(2N\log_2 N - 7N + 12)\}$ real-valued multiplications per fullband input sample.

4.2 MDCT

The MDCT is an overlapped transform. In a transform with M output components we need $2M$ input samples, out of which M samples overlap with the previous frame. In contrast to the discrete Fourier transform, the MDCT is a real-valued transform based on the cosine function. The MDCT is defined as [23], [24], [12],

$$X_m(k) = \sum_{n=0}^{2M-1} w(n)x(kM+n)\cos\left[\frac{\pi}{4M}(2n+1+M)(2m+1)\right], \quad m = 0\ldots M-1, \tag{33}$$

where m denotes the frequency component and k the time frame. The window function $w(n)$ can be used to improve frequency selectivity. Equation (33) can also be written as the sum of two components of discrete Fourier transforms of size $2M$,

$$\begin{aligned} \tilde{X}_m(k) &= \frac{1}{2}e^{-j\frac{\pi}{4M}(1+M)(2m+1)} \sum_{n=0}^{2M-1} w(n)x(kM+n)e^{-j\frac{\pi n}{2M}}e^{-j\frac{2\pi nm}{2M}}, \\ X_m(k) &= \tilde{X}_m(k) + \tilde{X}_m^*(k), \qquad m = 0\ldots M-1, \end{aligned} \tag{34}$$

where $w(n)x(kM+n)e^{-j\frac{\pi n}{2M}}$ is the windowed and modulated input signal to the discrete Fourier transforms. With an inverse MDCT (IMDCT) it possible to perfectly reconstruct the original signal. Each IMDCT results in $2M$ output samples, and an

Table 1: Frequency-domain adaptive filter. The block size is denoted N, and the number of adaptive coefficients per frequency bin P, resulting in a total filter length of PN. The power spectrum estimation forgetting factor is denoted β and the adaptive step size μ. In the unconstrained version, $\mathbf{F}\mathbf{W}_2\mathbf{F}^{-1}$ is replaced by the identity matrix $\mathbf{I}$. The Fourier matrix $\mathbf{F}$ is defined in (29).

Input signals	**Matrix sizes**
$\mathbf{X}(k) = \text{diag}\left\{\mathbf{F}\left[x(kN-N) \;\; \cdots \;\; x(kN+N-1)\right]^T\right\}$	$(2N \times 2N)$
$\mathbf{y}(k) = \left[y(kN) \;\; \cdots \;\; y(kN+N-1)\right]^T$	$(N \times 1)$
Power spectrum estimation with regularization	
$\mathbf{S}(k) = \beta\mathbf{S}(k-1) + (1-\beta)\mathbf{X}^H(k)\mathbf{X}(k)$	$(2N \times 2N)$
$\tilde{\mathbf{S}}(k) = \mathbf{S}(k) + \text{diag}\{\mathbf{e}_{\text{reg}}\}$	$(2N \times 2N)$
Filtering	
$\mathbf{e}(k) = \mathbf{y}(k) - \mathbf{W}_1\mathbf{F}^{-1}\sum_{p=0}^{P-1}\mathbf{X}(k-p)\hat{\mathbf{H}}_p(k)$	$(N \times 1)$
$\mathbf{E}(k) = \mathbf{F}\mathbf{W}_1^T\mathbf{e}(k),$	$(2N \times 1)$
$\hat{\mathbf{H}}_p(k+1) = \hat{\mathbf{H}}_p(k) + \mu\mathbf{F}\mathbf{W}_2\mathbf{F}^{-1}\tilde{\mathbf{S}}^{-1}(k)\mathbf{X}^H(k-p)\mathbf{E}(k)$	$(2N \times 1)$
Definitions	
$\mathbf{e}(k) = \left[e(kN) \;\; \cdots \;\; e(kN+N-1)\right]^T$	$(N \times 1)$
$\hat{\mathbf{H}}_p(k) = \mathbf{F}\left[\hat{\mathbf{h}}_p^T(k) \;\; \mathbf{0}_{1\times N}\right]^T$	$(2N \times 1)$
$\hat{\mathbf{h}}_p(k) = \left[\hat{h}_{pN}(k) \;\; \cdots \;\; \hat{h}_{pN+N-1}(k)\right]^T$	$(N \times 1)$
$\mathbf{W}_1 = \left[\mathbf{0}_{N\times N} \;\; \mathbf{I}_{N\times N}\right],\ \mathbf{W}_2 = \text{diag}\left\{\left[\mathbf{1}_{1\times N} \;\; \mathbf{0}_{1\times N}\right]\right\}$	

overlap add method is used to reconstruct the fullband signal,

$$\begin{aligned} \tilde{x}^{rec}(p,k) &= \frac{w(p)}{M}\sum_{m=0}^{M-1} X_m(k)\cos\left[\frac{\pi}{4M}(2p+1+M)(2m+1)\right], \\ & \qquad\qquad p = 0\ldots 2M-1, \\ x^{rec}(n) &= \tilde{x}^{rec}(n\%M+M, \lfloor n/M\rfloor) + \tilde{x}^{rec}(n\%M, \lfloor n/M\rfloor+1), \end{aligned} \tag{35}$$

where $\lfloor \cdot \rfloor$ denotes the nearest smaller integer and % the modulus operator. Usually the $2M$ long window is a symmetric function and in order to achieve perfect reconstruction, the window function needs to satisfy the property (37),

$$w(n) = w(2M - n - 1), \qquad n = 0 \ldots 2M - 1, \tag{36}$$

$$w^2(n) + w^2(n + M) = 2, \qquad n = 0 \ldots 2M - 1. \tag{37}$$

Two simple windows that satisfy (36) and (37) are

$$w_1(n) = 1, \qquad n = 0 \ldots 2M - 1, \tag{38}$$

$$w_2(n) = \sqrt{2} \sin\left(\frac{\pi}{2M}\left(n + \frac{1}{2}\right)\right), \quad n = 0 \ldots 2M - 1. \tag{39}$$

It is possible to construct windows with better frequency selectivity by using numerical design methods. This is done in, e.g., the MPEG 2 AAC coder [12].

In [25], it is shown how the MDCT can be repartitioned, making it possible calculate it with an FFT of size M. Using this method, and a Radix 2 FFT, the MDCT can be calculated with only $2M\log_2 M + 3M + 12$ real-valued multiplications, including the multiplications needed to repartition the data before and after the FFT.

4.3 A Joint Transform Structure for Audio Coding and Echo Cancellation

In a stand alone FDAF echo canceler, the output signal is the time domain signal $\mathbf{e}(n)$. If we were to combine the FDAF based echo canceler with an audio coder that is based on the MDCT transform, it would be attractive to transform the error signal, (30), to the frequency-domain. Thus, we need to exchange (30) and (31) for the frequency-domain counterpart,

$$\mathbf{E}(k) = \underbrace{\mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}^T \mathbf{y}(k)}_{\mathbf{Y}(k)} - \underbrace{\mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}\mathbf{X}(k)\hat{\mathbf{H}}(k)}_{\hat{\mathbf{Y}}(k)}. \tag{40}$$

The constraint used in the calculation of $\hat{\mathbf{Y}}(k)$ is needed, since $\mathbf{X}(k)\hat{\mathbf{H}}(k)$ corresponds to circular convolution in the time domain. By using (34), we can construct the output of the MDCT directly from the frequency-domain output of the echo canceler, $\mathbf{E}(k)$ in (40). This requires the FDAF block size N to be equal to the number of input samples in each MDCT, i.e., $N = 2M$. It also requires that the MDCT window, $w(n)$, is incorporated in $\mathbf{E}(k)$. This can be done by exchanging $\mathbf{I}_{N\times N}$ in (30) for diag{$\mathbf{w}$}, where $\mathbf{w} = [w(0) \ldots w(N-1)]$. Also in (30), we need to pre-multiply $\mathbf{y}$ with diag{$\mathbf{w}$}.

Equivalently, the two $\mathbf{I}_{N\times N}$ in (40) can be exchanged for diag{$\mathbf{w}$}. Like for the MDCT, see (33), the input data will need to overlap by 50%. Simulations show a somewhat slower convergence rate for the windowed FDAF, but a proper convergence analysis is not available.

By exchanging (30) and (31) for (40) the calculation complexity for the FDAF will actually increase by one discrete Fourier transform of size $2N$. However, one transform of size $2M$ in the audio coder can be saved, since output of the MDCT in the coder can be constructed directly from $\mathbf{E}(k)$ in (40) by using the MDCT transform in (34). As mentioned above, the FDAF must have a block size of $N = 2M$. The frequency-domain residual echo vector, $\mathbf{E}(k)$, will be of size $(2N \times 1)$, due to the zero padding in (31) or equivalently in (40). Therefore, only half the components in $\mathbf{E}(k)$ are needed in the reconstruction of the output of the MDCT.

We have now shown that it is possible to combine the transforms in an MDCT based audio coder and an FDAF based echo canceler. However, there are a couple of things that limit the usefulness of combining the two transforms.

FDAF Calculation complexity: As we have seen above, due to the constraint needed in (40), the number of discrete Fourier transforms cannot be reduced by combining the transforms of the echo canceler and the audio coder, when standard FDAF based echo cancelers are used. This is a fundamental problem caused by the fact that the time domain filtering, $y(n) = x(n) * h(n)$, cannot be written as a simple matrix product in the discrete Fourier domain, i.e. $\mathbf{Y}(k) \neq \mathbf{X}(k)\mathbf{H}(k)$.

Adaptive MDCT size: The use of large MDCT transforms improve frequency resolution, whereas time resolution is decreased. Therefore, large transforms can create problems for transient signals. Encoder quantization errors, extending more than a few milliseconds *before* a transient event are not effectively masked by the transient itself. This leads to a phenomenon called preecho, in which quantization error from one transform block is spread in time and becomes audible. It is common that advanced audio coders use shorter MDCT transforms during transient signals and longer MDCT transforms otherwise [12]. Such a switch would of course affect the construction of joint transforms for echo cancellation and audio coding. One solution is to re-map the estimated transfer function, $\hat{\mathbf{H}}_p(k)$, from, e.g., an estimate with $L = 1024$ frequency bins and $P = 1$ taps per frequency bin to a $L = 256$ and $P = 4$ transfer function estimate. This can be done by using the definition of $\hat{\mathbf{H}}_p(k)$ in Table 1. In order to reduce calculation complexity, we could also decide to update $\hat{\mathbf{H}}_p(k)$ only for blocks where $L = 1024$.

Coder pre-processing: Some audio encoders process the signals before the MDCT transform is performed. For example, the MPEG-2 AAC coder offers three profiles, each one tuned for different needs. In one of these profiles, the scalable

sampling-rate profile, the input signal is divided into four frequency bands with a critically downsampled filterbank. Then, in each frequency band, the gain is adjusted before the bands are processed by four independent MDCT transforms. In this situation, it is impossible to combine the transforms of the coder and the echo canceler due to the downsampling alias caused by the 4 band filterbank.

Combining the discrete Fourier transform used in the FDAF based echo canceler with the MDCT transform used in the audio coder may not always be advantageous, as has been shown above. In those situations, we should at least use the same signal sample buffers for the audio coder and a frequency-domain based echo canceler. This way, the signal delay imposed by the echo canceler could be zero in theory. In a practical situation, the delay caused by the echo canceler would be limited to the time needed to process equations (27) and (30). This solution would also be very flexible, in that the suitable values of the FDAF block size N could be such that $M = kN$, k positive integer. For $N > M$ we have two options. One is to wait for all N samples, and thereby introduce a delay to the transmission signal path. The second way would be to let the data in the buffers represented by (27) and (28) overlap. If we use only M new samples for each FDAF cycle, no delay needs to be introduced to the transmission signal path. Using overlapped buffers will increase the calculation complexity, but the convergence rate of the adaptive filter will also increase. This should be compared to the value $N = 2M$ necessary for joint echo canceler and coder transforms. In this situation, the input data need to overlap by 50%.

5 Simulations

This section exemplifies the performance of the echo canceler in a couple of different situations. In all simulations, the same source signal is used. The transmission room signal, $x(n)$ (Fig. 1), is recorded at 16 kHz sampling rate in a quiet office-like room. The receiving room impulse response, $h(n)$, is measured in a quiet office-like room. The response is 2048 taps long, corresponding to 128 ms. Then the receiving room signal, $y(n)$, is given by filtering $x(n)$ with $h(n)$, and adding a recorded background noise signal. The average SNR, measured at the microphone, is 38 dB. The echo signal, $y(n)$, is shown in Fig. 7(a).

The normalized mean square error[2] (MSE) energy of the residual is used as performance index. The MSE is given by,

$$\mathrm{MSE} = \frac{\mathrm{LPF}[e(n) - w(n)]^2}{\mathrm{LPF}[y(n) - w(n)]^2}, \tag{41}$$

[2] Since we normalize with the power of the echo. We can regard this as the inverse of the echo return loss enhancement (ERLE).

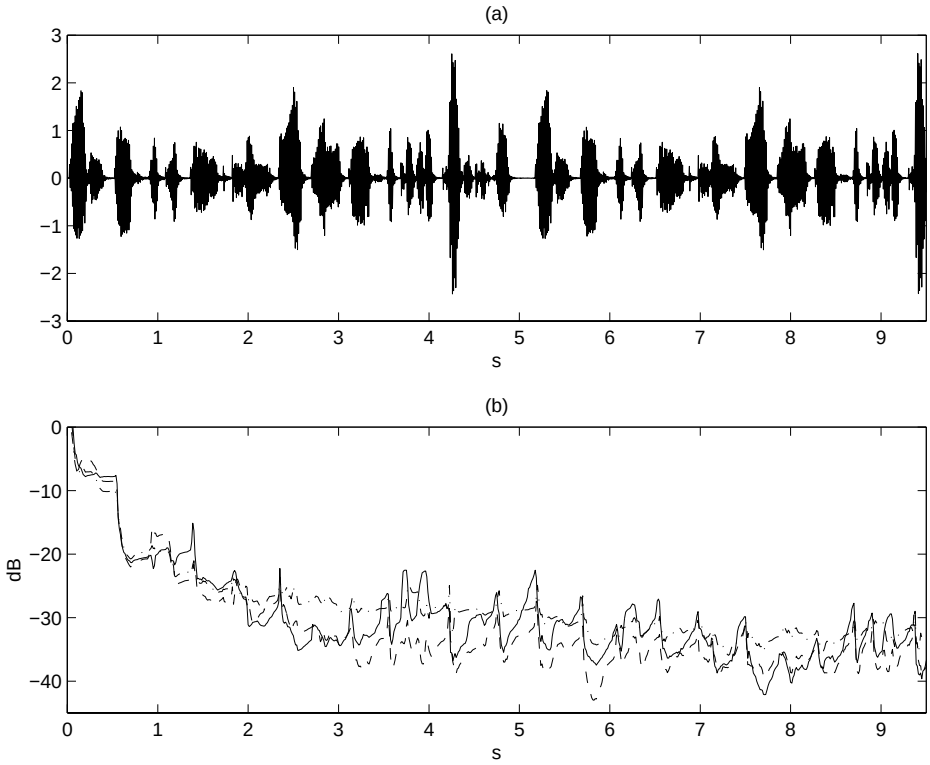

Figure 7: (a) The echo signal, $y(n)$, used in all simulations. (b) The MSE performance for three different systems: System 1a, fullband NLMS (solid line). System 2, subband system with non-critical EC filterbank (dashed line). System 3, subband system with modified filterbank from the MPEG 1 and 2 audio coder (dashed-dotted line).

where w denotes the receiving room background noise signal and LPF denotes a low-pass filter; in this case it has a single real pole at 0.999.

In Table 2, the number of real-valued multiplications per fullband input sample and the signal transmission delay are summarized for all systems considered in this section. The numbers include all components shown in Fig. 1, i.e., also the filterbank or the MDCT transform used in the audio coder. It should be noted that the delay figures are valid for 16 kHz sampling rate, i.e., the delay would be reduced by 50% for 32 kHz sampling rate.

Table 2: Calculation complexity as the number of real-valued multiplication per sample and the signal transmission delay in ms, for the algorithms used in the Section 5. The contributions from all components in Fig. 1 are included.

Filterbank based audio coder and echo canceler

	System 1a	**System 2**	**System 3**
nbr. of subbands, M^{EC}	–	64	64
downsamp. factor, r	–	48	32
complex. (mult.)	2085	327	406
delay (ms)	32	104	48

MDCT based audio coder and FDAF based echo canceler

	System 1b	**System 4**
MDCT size, M	1024	1024
FDAF size, N	–	1024
FDAF overlap, α	–	4
complex. (mult.)	2094	390
delay (ms)	128	128†

MDCT based audio coder and FDAF based echo canceler

	System 1c	**System 5**	**System 6**
MDCT size, M	256	256	256
FDAF size, N	–	256	256
FDAF overlap, α	–	1	4
complex. (mult.)	2086	345	1264
delay (ms)	32	32†	32†

†*The delay for the systems using a FDAF based echo canceler, does not include the delay needed for equations (27) and (30), explained in Section 4.3.*

Filterbank Based Echo Canceler/Audio coding Systems

In this section we consider systems where the audio coder is based on a real-valued QMF filterbank with 32 subbands. The filterbank used in the MPEG 1 coder is used for complexity and delay calculations. We will consider three different systems considerations, listed below, and the MSE of these systems are shown in Fig. 7(b).

System 1a A real-valued fullband NLMS EC which is considered as the reference system with 1024 adaptive filter taps and a step size parameter $\mu = 0.5$.

System 2 A subband EC based on the filterbank design in Section 2.2 is the second

system under consideration. This system has 64 subbands and a downsampling factor of 48. The length of the filterbank prototype filter is 895 and it introduces a delay of 56 ms. Each subband adaptive filter has 27 coefficients which corresponds to 1046 fullband taps. To these we have added 250 non-causal taps, which increase the delay by 16 ms. The step size parameter of NLMS adaptive filters is also 0.5.

System 3 The third system uses the complex version of the MPEG 1 audio coder filterbank described in Section 3.2. This filterbank, like the previous, has 64 subbands but the downsampling factor is only 32. Each adaptive filter has 40 adaptive filter taps, corresponding to 1030 fullband taps plus 250 non-causal taps. Since the filterbank can be combined between the audio coder and the echo canceler the transmission delay caused by the echo canceler is reduced. There is still a small delay caused by the need of the non-causal filter taps, and this delay is as before 16 ms.

The three systems appear to have similar MSE performance. As was predicted in Section 3.2, the system based on the MPEG 1 audio coder filterbank has a somewhat slower convergence rate than the system based on the filterbank design in Section 2.2, see Fig. 7(b) at approximately the time instance 3 s.

In Section 1 it was claimed that a subband NLMS based echo canceler has a faster convergence rate than a fullband NLMS based echo cancelers in frequency regions with small eigenvalues. In order to show this, the power spectrum of the residual echo signals of the three systems have been computed. The power spectra was averaged over the time interval 1.8 to 3.1 s of the residual echo signals used to derive the MSE plots in Fig. 7. These spectra are presented in Fig. 8 with the same line types as in Fig. 7. Additionally, the spectra of the transmission room signal, $x(n)$, and the receiving room signal, $y(n)$, are shown with a thick solid line and a dashed line, respectively. The subband systems have good suppression of the echo signal in all frequency regions, whereas the fullband NLMS systems only perform well in regions with large signal energy. In Table 2, the calculation complexity and the transmission delay for the considered systems are summarized.

Transform Based Echo Canceler/Audio coding Systems

In the following, the same set of simulations as above are performed with the frequency-domain algorithm presented in Table. 1. In the first scenario we apply an MDCT that is used in, e.g., the MPEG 2 AAC coder. For this MDCT, $M = 1024$ for non-transient signals. That is, 1024 new samples are needed for each block.

System 1b Like system 1a, however, instead of a filterbank based audio coder, an MDCT based audio coder, with block size $M = 1024$, is now used. The delay

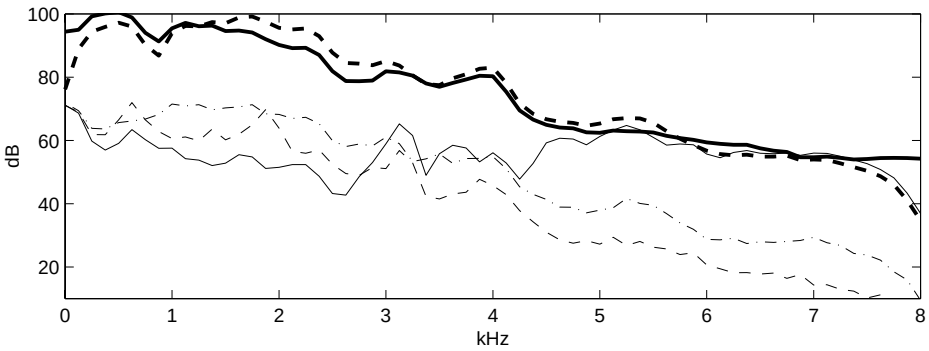

Figure 8: The power spectra of the residual echo signals used to derive the MSE plots in Fig. 7, averaged over the time interval 1.8 to 3.1 s. The thick solid lines depict the power spectrum of the transmission signal, $x(n)$, and the thick dashed line of the receiving room signal, $y(n)$. Other conditions same as in Fig. 7(b).

and calculation complexity caused by the coder will make this system differ from System 1a.

System 4 An MDCT with block size of $M = 1024$ could efficiently be combined with a frequency-domain EC with the same block size, $N = 1024$. We will use one filter tap per frequency bin ($P = 1$), the step size parameter $\mu = 0.08$ and the power spectrum estimation forgetting factor $\beta = 0.96$. In the algorithm in Table. 1, the input data blocks have no overlap, $\alpha = 1$. By overlapping the input data, it is possible to increase the convergence rate, and in this simulation, we update 4 times per block ($\alpha = 4$), i.e., only 25% new input-data is used for each iteration. The unconstrained version in Table. 1 is used.

In Fig. 9, the MSEs for System 1b and 4 are depicted. The power spectra estimates are presented in Fig. 10.

Finally we have an audio coder with an MDCT size of $M = 256$ for non-transient signals, and we will combine this with an FDAF based echo canceler also with the same block size, $N = 256$. This block size is used in, e.g., an AC-3 coder from Dolby Laboratories [26]. The shorter block size is compensated in the FDAF based echo canceler by having $P = 4$ filter taps per frequency bin. For $P > 1$, the constrained version of Table. 1 is needed. The filter parameters μ and β have the same values as in the previous simulations.

System 1c Like system 1b, however, the bock size of the audio coder is now $M = 256$.

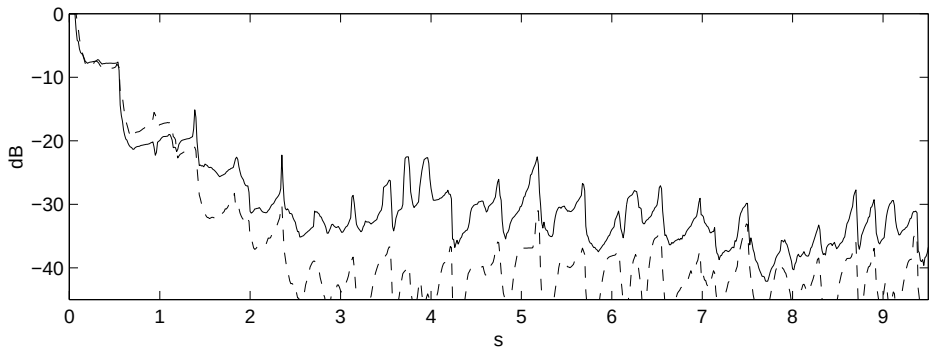

Figure 9: The MSE performance: Fullband NLMS (System 1b, solid line), and System 4, an unconstrained FDAF (dashed line). The FDAF block size $N = 1024$ was used.

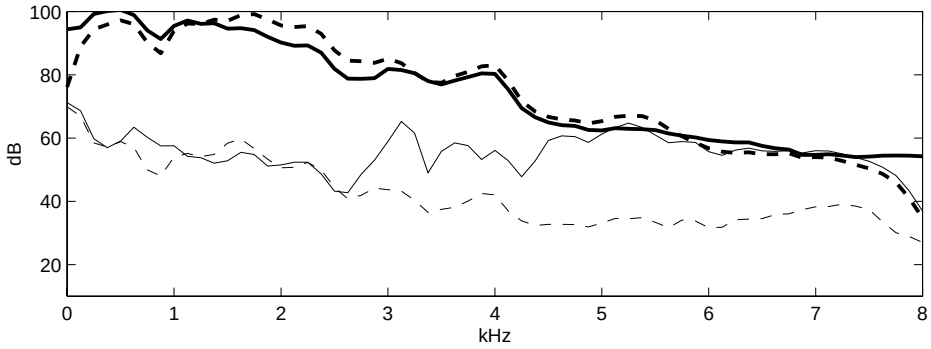

Figure 10: Power spectra of the residual echo signals used to derive the MSE in Fig. 9, other conditions same as in Fig. 8.

System 5 A FDAF based echo canceler with block size $N = 256$ and $P = 4$ filter taps per frequency bin. No input signal overlap, i.e. $\alpha = 1$.

System 6 Like System 5, however we now have only 25% new data per block, i.e., $\alpha = 4$, in order to increase the convergence rate.

The results of the simulations with System 1c, 5 and 6 are shown in Fig. 11 and Fig. 12.

6 Conclusions

In this paper, joint structures for audio coding and echo cancellation are considered. The paper focuses on two types of audio coders; coders based on cosine modulated

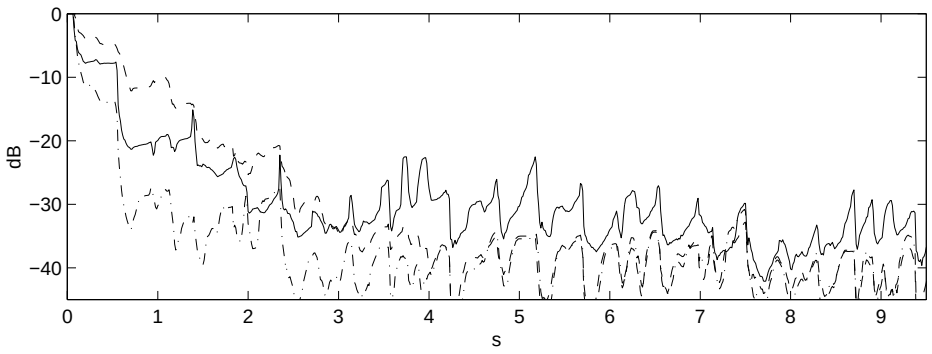

Figure 11: The MSE performance: Fullband NLMS (System 1c, solid line), constrained FDAF with block overlap parameter $\alpha = 1$ (System 5, dashed line) and constrained FDAF with block overlap parameter $\alpha = 4$ (System 6, dashed-dotted line). The FDAF block size $N = 256$ and $P = 4$ adaptive taps per frequency bin were used.

filterbanks and audio coders based on the modified discrete cosine transform (MDCT).

For audio coders based on filterbanks, the preferred configuration is highly dependent on the desired performance of the system. If we are to construct a system aimed for a platform where low calculation complexity is of importance, we would probably select the modified audio coder filterbank presented in Section 3.2 and denoted System 3 in the Simulation section. This choice is a good compromise between low calculation complexity and low signal transmission delay. In situations where long signal transmission delay is acceptable, we can gain even lower calculation complexity by using separate filterbanks for the coder and the echo canceler, as depicted in Fig. 1. This system will also have a better convergence rate for the echo canceler, as illustrated with System 2 in the Simulation section. For systems with two audio channels, we need a stereophonic echo canceler. In these systems, the demands on the adaptive filter in the echo canceler is higher, and usually more complex adaptive filters are needed [27]. If the two-channel fast recursive least squares (FRLS) algorithm [28], [15], [27] is chosen, the convergence performance should not be significantly affected by the filterbank choice. However, the calculation complexity for the two-channel FRLS algorithm is significantly higher than for the NLMS algorithm. Therefore the calculation complexity difference between systems using the two types of filterbanks will increase, to the advantage of the specially designed echo canceler filterbank. In this situation, the joint structures presented in Section 3.3 could be the best solution.

Audio coders based on the MDCT process blocks of data, and therefore a block based echo canceler, e.g. a frequency domain based echo canceler, is preferable. In Section 4.3 it is shown how an MDCT based audio coder can share transforms with a frequency-domain based echo canceler. However, it is also shown that the advantages

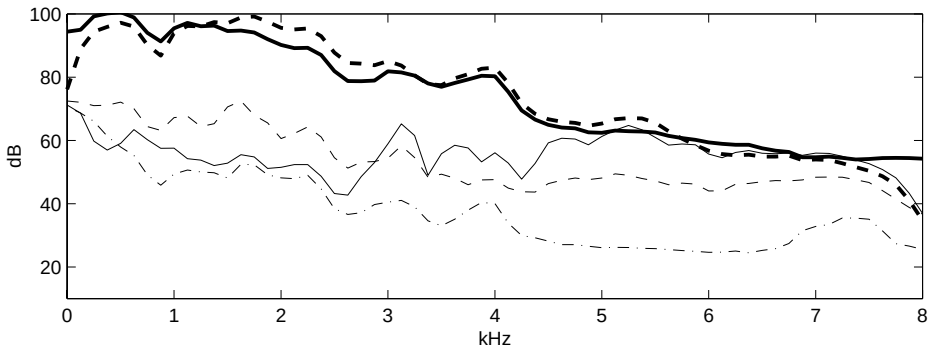

Figure 12: Power spectra of the residual echo signals used to derive the MSE in Fig. 11, other conditions same as in Fig. 8.

of joint transforms are limited. In most situations it may be preferable to let the audio coder and the echo canceler share the signal sample buffers. The advantages over a design where the audio coder and the echo canceler are two completely separated units include negligible signal transmission delay caused be the echo canceler. Compared to the situation with joint transforms, the shared buffers system has smaller restrictions on the design of the frequency-domain based echo canceler. The only restriction is that the maximum block size of the echo canceler is the block size of the audio coder. Larger blocks will increase the signal transmission delay. Frequency-domain adaptive algorithms usually have good convergence and properties, as is exemplified in Section 5. Another advantage of a frequency-domain based echo canceler is the existence of efficient two-channel algorithms, suitable for communication systems with stereophonic sound [21], [22], [27].

Acknowledgments

The author would like to thank Dr Tomas Gänsler, Bell Labs, for constructive discussions.

References

[1] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. XLVI, no. 3, pp. 497–510, Mar. 1967.

[2] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. of ICASSP*, 1988, pp. 2570–2573.

[3] D. R. Morgan, "Slow asymtotic convergence of LMS acoustic echo cancelers," *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 2, pp. 126–136, Mar. 1995.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice Hall International, 1996.

[5] D. Mansour and A. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 30, no. 5, pp. 726–734, Oct. 1982.

[6] P. Sommen, P. Van Gerwen, H. Kotmans, and A. Janssen, "Convergence analysis of a frequency-domain adaptive filter with exponetial power averaging and generalized window function," *IEEE Trans. on Circuits and Systems*, vol. 34, no. 7, pp. 788–798, July 1987.

[7] P. Noll, "Wideband speech and audio coding," *IEEE Communication Mag.*, pp. 34–44, Nov. 1993.

[8] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, vol. 14, no. 5, pp. 59–81, Sept. 1997.

[9] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall PTR, 1993.

[10] ISO/IEC 11172–3, "Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio," ISO/IEC JTC 1/SC 29, Case Postale 56, CH1211 Genève 20, Switzerland, 1993.

[11] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to* MPEG-*2*, chapter 4, pp. 55–79, Digital Multimedia Standards Series. Chapman & Hall, 1997.

[12] M. Bosi et. al., "ISO/ICE MPEG-2 advanced audio coding," *J. Audio Eng. Soc. (AES)*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[13] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[14] G. Wackersreuther, "On the design of filters for ideal QMF and polyphase filter banks," *AEÜ*, vol. 39, no. 2, pp. 123–130, 1985.

[15] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A real-time implementation of a stereophonic acoustic echo canceler," *IEEE Trans. Speech Audio Processing*, accepted for publication.

[16] H. Sorensen, D. Jones, M. Heideman, and S. Burrus, "Real-values fast Fourier transform algorithms," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, pp. 849–863, June 1987.

[17] T. Saramäki and J. Yli-Kaakinen, "Design of digital filters and filter banks by optimization: Applications," in *Proc. of EUSIPCO*, Sept. 2000.

[18] M. Vetterli and H. Nussbaumer, "Simple FFT and DCT algorithms with reduced number of operations," *Signal Processing*, vol. 6, no. 4, pp. 267–278, Aug. 1984.

[19] J. Proakis and D. Manolakis, *Digital Signal Processing*, Prentice Hall Inc., 1996.

[20] P. Eneroth and T. Gänsler, "Analysis of subband impulse responses in subband echo cancelers," *IEEE Trans. Signal Processing*, submitted.

[21] J. Benesty, A. Gilloire, and Y. Grenier, "A frequency domain stereophonic acoustic echo canceler exploiting the coherence between the channels," *J. Acoust. Soc. Am.*, vol. 106, pp. L30–L35, Sept. 1999.

[22] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 789–792.

[23] J. Princen and A. Bradley, "Analysis/synthesis filterbank designed based on time domain aliasing cancellation," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 34, no. 5, pp. 1153–1161, Oct. 1986.

[24] J. Princen, A. Johnson, and A. Bradley, "Suband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE ICASSP*, Apr. 1987, pp. 50.1.1–50.1.4.

[25] M. Iwadare et. al., "A 128 kb/s Hi-Fi audio CODEC based on adaptive transform coding with adaptive block size MDCT," *IEEE J. on Selected areas in Communications*, vol. 10, no. 1, pp. 138–144, Jan. 1992.

[26] Steven Vernon, "Design and implementation of AC–3 coders," *IEEE Trans. on Consumer Electronics*, vol. 41, no. 3, Aug. 1995.

[27] P. Eneroth, J. Benesty, T. Gänsler, and S. L. Gay, "Comparison of different adaptive algorithms for stereophonic acoustic echo cancellation," in *Proc. of EUSIPCO*, 2000.

[28] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.