Joint Filterbanks for Echo Cancellation and Audio Coding

Peter Eneroth

***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.**

***Index Terms*—Audio coding, echo cancellation, filter bank, MDCT.**

I. Introduction

ECHOES in telephone systems became severe with the introduction of long distance telephone services. In the 1960s, 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 nonwhite 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 nonwhite 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 order to reduce the complexity without introducing signal path delay, a class of algorithms denoted delayless subband

Manuscript received January 4, 2001; revised November 22, 2002. This work was performed at the Department of Electroscience, Lund University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hsiao-Chuan Wang.

The author is with Ericsson Mobile Platforms, SE-221 83 Lund, Sweden (e-mail: peter.eneroth@ericsson.com).

Digital Object Identifier 10.1109/TSA.2003.814798

adaptive filter has been introduced, e.g., [7]–[9]. These algorithms will not be studied in detail in this paper, however, a complexity example is given in Section V.

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 [10], [11]. 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 [12], 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 [13], [14], 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 [15], 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 II, 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 III, we will show possible modifications in order to combine the two types of filterbanks. Then, in Section IV, 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 V.Fig. 1.

II. 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

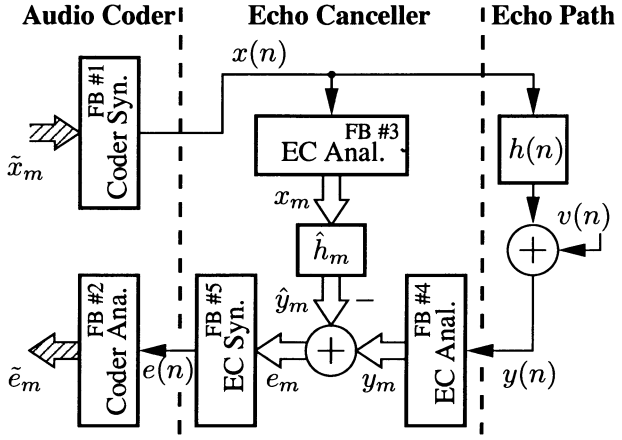

Fig. 1. Subband echo canceler and model of acoustic echo path. In each subband, an adaptive algorithm estimates the subband impulse response $\hat{h}_m$, and the subband residual echo signal $e_m(k)$ is generated.

$v(n)$. 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

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.

A. 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 filter updates are calculated as [4]

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

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.

B. Traditional Filterbank Design for Echo Cancelers

The purpose of the echo cancellation filterbank is to reduce the calculation complexity of the adaptive filters by multirate processing. 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 [16]. 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, $h^{\text{EC}}(n)$, [17]

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

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

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

where $*$ denotes convolution and $\downarrow$ downsampling. Because of symmetry between the subband filters, $h_m^{\text{EC}}(n)$, efficient fast Fourier transform (FFT) based implementations are possible, [17], [18]. 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

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

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

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

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

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

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 cancel aliasing. Instead, aliasing is appropriately suppressed by efficient stopband attenuation [17], [18]. 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 $(M^{\text{EC}}/2) + 1$ subbands, whereas the upper subband signals can be reconstructed from the lower subbands, due to the symmetry in (2).

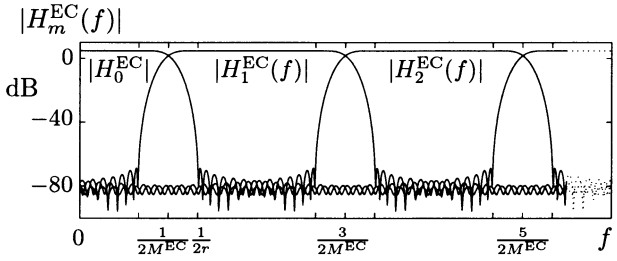

Fig. 2. Amplitude response of the echo cancellation filterbank, with noncritical downsampling. The filters suppress aliasing by sufficient stopband attenuation; see filter $|H_0^{\text{EC}}(f)|$ for $f > 1/2r$.

Efficient structures suitable for implementation, both of the analysis and synthesis filterbanks, are presented in [18]. 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 $2M^{\text{EC}}\log_2 M^{\text{EC}} - 7M^{\text{EC}} + 12$ multiplications [19], where M^{EC} is block size. In [19] it is also shown how one FFT can be used to transform two real-valued signals, it only increases the complexity by $2M^{\text{EC}} - 4$ 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 $1/r(2K^{\text{EC}} + 2M^{\text{EC}}\log_2 M^{\text{EC}} - 7M^{\text{EC}} + 12)$ and $1/r(K^{\text{EC}} + 2M^{\text{EC}}\log_2 M^{\text{EC}} - 7M^{\text{EC}} + 12)$ real-valued multiplications per fullband sample, respectively.

C. 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 noncritically 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 [13], only aliasing from adjacent bands is cancelled, whereas in a perfect reconstruction filterbank [12], 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 [20].

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

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

$$= \tfrac{1}{2}h^{\mathrm{A}}(n)\left(e^{-j(\pi/M^{\mathrm{A}})(m+(1/2))(n-16)} + e^{j(\pi/M^{\mathrm{A}})(m+(1/2))(n-16)}\right), \quad 0 < m < M^{\mathrm{A}}-1 \quad (9)$$

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 1/2. The phase shift 16 can differ in different filterbanks, but only some phase values are valid for aliasing cancellation [12]. 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 $\pm(m+(1/2))$, and this is illustrated in Fig. 3. In the figure, the frequency response of $h^{\mathrm{A}}(n)e^{-j(\pi/M^{\mathrm{A}})(m+(1/2))(n-16)}$ and $h^{\mathrm{A}}(n)e^{j(\pi/M^{\mathrm{A}})(m+(1/2))(n-16)}$ 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 (7), 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 (3) 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

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

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.

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 [12]). 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 [16], but the adaptive filtering will also make alias cancellation to be performed by the synthesis filterbank impossible. One solution presented in [16] is to let the adaptive filter have special cross-band filters, but it is also concluded in [16] that this solution has a very high calculation complexity, and that the adaptive filters have a slow convergence rate.

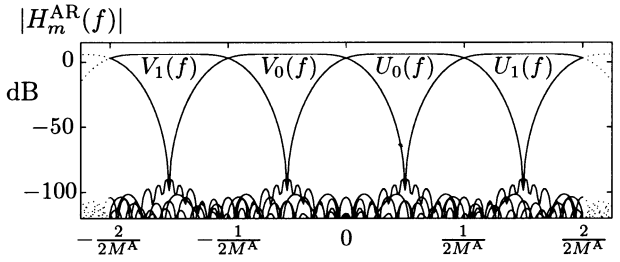

Fig. 3. Filterbank used in MPEG 1 audio, usually denoted pseudo QMF filterbank. It has 32 real-valued subbands, and it is critically downsampled. Each subband consist of a positive and a negative part, denoted $U_m(f)$ and $V_m(f)$, respectively.

An efficient structure for the pseudo QMF filterbank is given in [12]. 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 $(M^{\mathrm{A}}/2)\log_2 M^{\mathrm{A}}$ real-valued multiplications [21]. That is, both the analysis and synthesis filterbank need $(1/M^{\mathrm{A}})(K^{\mathrm{A}}+(M^{\mathrm{A}}/2)\log_2 M^{\mathrm{A}})$ real-valued multiplications per fullband sample each.

III. 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.

A. 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

$$-\frac{2m+3}{4M^{\mathrm{A}}} < f < -\frac{2m-1}{4M^{\mathrm{A}}} \quad \text{and} \quad \frac{2m-1}{4M^{\mathrm{A}}} < f < \frac{2m+3}{4M^{\mathrm{A}}}.$$

The frequency-domain formula for decimating a factor r can be expressed as [22]

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

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

$$-\frac{1}{2} < f < -\frac{2m-1}{2M^{\mathrm{A}}} \quad \text{and} \quad \frac{2m-1}{2M^{\mathrm{A}}} < f < \frac{1}{2}.$$

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.

B. Oversampled Complex Valued Pseudo QMF Filterbank

If we once again return to (9), we can find a better modification, resulting in a noncritically 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

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

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 noncausal taps, usually needed in subband echo cancellation [2], [23]. 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 noncausal 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

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

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 $\pm(1/2M^{\mathrm{A}})$ 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 areas centered around $\pm(2/2M^{\mathrm{A}})$ 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:

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

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 III-A, 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 V. 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

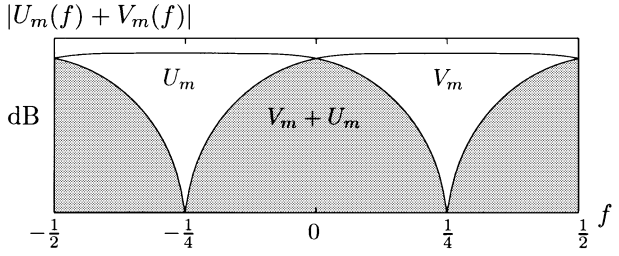

Fig. 4. Frequency response of subband m, m odd, after filtering and downsampling of the pseudo QMF filterbank used in MPEG 1. The gray areas represent aliasing. For even m, U_m and V_m will exchange position, see (11).

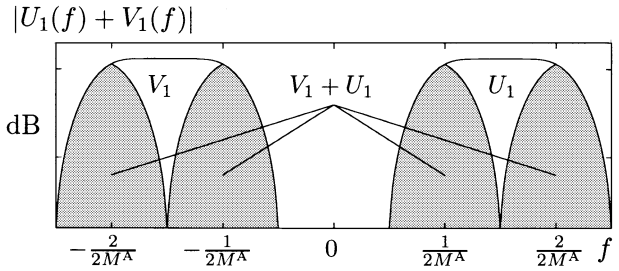

Fig. 5. Frequency response of subband 1 after upsampling and suppression of imaging, given by $\check{x}_1^{rec}(n)$ in (16).

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 based audio coder, will only increase the signal transmission delay by a small value corresponding to a few noncausal taps needed in each subband [2], [23]. In Section V, 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 II-B. This filterbank will be used to replace filterbanks 3 and 4 in Fig. 1, i.e., we will need two filterbanks. From Section II-B we find that, by replacing M^{EC} with $2M^{\mathrm{A}}$ and r with M^{A}, the two filterbanks will need $(1/M^{\mathrm{A}})(2K^{\mathrm{A}} + 4M^{\mathrm{A}}\log_2 2M^{\mathrm{A}} - 14M^{\mathrm{A}} + 12)$ 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.

C. 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 II-B and the audio coder the filterbank in Section II-C. We will use M^{EC} subbands with a downsampling factor of $r = 3M^{\mathrm{EC}}/4$ for the echo canceler filterbank and a $M^{\mathrm{A}} = M^{\mathrm{EC}}/2$ subbands critically downsampled audio coder filterbank. The frequency response of the lower subbands of the two filterbanks are shown in Fig. 6.

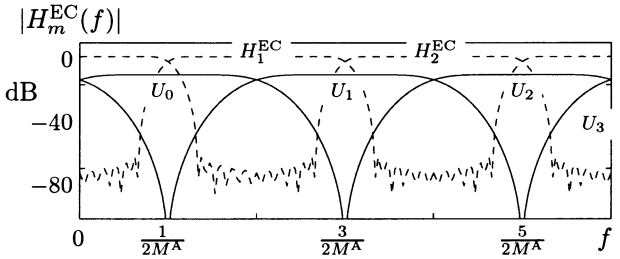

Fig. 6. Frequency response of the lower subbands of the two filterbanks. The audio coder's QMF filterbank is illustrated with a solid line, and the EC filterbank with a dashed line.

First we study possibilities of combining filterbanks 1 and 3 in Fig. 1. In Section III-B, it is shown that the only delay reduction possible is a reduction of the delay that will compensate for a few noncausal taps needed in each subband [2], [23]. Also, Section III-B 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.,

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

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

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

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 nonzero 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

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

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

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

The total minimization problem can now be expressed as

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

where $\alpha_i \geq 0$ 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 II-B 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 fullband signal. Actually, it is enough to increase the sampling rate by a factor 3 because a signal with the bandwidth of $F_1^{\text{EC}}(f)$ plus $F_2^{\text{EC}}(f)$ 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 $H_q^{\text{EC}}(f)$ and $F_q^{\text{EC}}(f)$ 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 $h_1^{\text{AR}}(n)$ has to be used

$$\tilde{e}_1(k') = 2\cdot\text{Re}\left\{\sum_{q=1}^{2}(e_q(k)\uparrow 3) * \tilde{f}_q^{\text{EC}}(n') * \tilde{h}_1^{\text{AC}}(n')\right\}\downarrow 2 \tag{25}$$

where $\tilde{f}_q^{\text{EC}}(n')$ and $\tilde{h}_1^{\text{AC}}(n')$ denote $f_q^{\text{EC}}(n) \downarrow (r/3)$ and $h_1^{\text{AC}}(n) \downarrow (r/3)$, 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

$$\tilde{e}_m(k') = 2\cdot\text{Re}\left\{\sum_{q=1}^{2}(e_q(k)e^{j\kappa_{m,q}k}\uparrow 3) * \tilde{f}_q^{\text{EC}}(n') * \tilde{h}_1^{\text{AC}}(n')\right\}\downarrow 2 \tag{26}$$

where $\kappa_{m,q}$ is the modulation factor, individual for each subband. By creating one filter from of $\tilde{f}_m^{\text{EC}}(n')$ and $\tilde{h}_1^{\text{AC}}(n')$, like in (19), we can reduce the total length the same way as was done in (21)–(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.

IV. 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 II-C 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.

A. 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 nonwhite 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

$$\mathbf{X}(k) = \text{diag}\left\{\mathbf{F}[x(kN-N)\cdots x(kN+N-1)]^T\right\} \tag{27}$$

$$\mathbf{y}(k) = [y(kN-N)\cdots y(kN+N-1)]^T \tag{28}$$

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

$$[\mathbf{F}]_{(p,q)} = \frac{1}{\sqrt{2N}}\, e^{-j(2\pi pq/2N)}, \qquad p, q \in [0, \ldots, 2N-1] \tag{29}$$

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

$$\mathbf{e}(k) = \mathbf{y}(k) - \underbrace{[\,\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}\,]\mathbf{F}^{-1}\mathbf{X}(k)\hat{\mathbf{H}}(k)}_{\hat{\mathbf{y}}(k)} \tag{30}$$

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

$$\mathbf{E}(k) = \mathbf{F}[\,\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}\,]^T\mathbf{e}(k) \tag{31}$$

$$\hat{\mathbf{H}}(k+1) = \hat{\mathbf{H}}(k) + \mu\mathbf{X}^H(k)\mathbf{E}(k) \tag{32}$$

where μ is the step size parameter. A complete version of the FDAF algorithm is given in Table I. 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 multichannel FDAF can be found in [24], [25].

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 V. 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 I 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 $\alpha/N\{60N + 2(2N\log_2 N - 7N + 12)\}$ 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 $\alpha/N\{52NP + 8N + (2 + 2P)(2N\log_2 N - 7N + 12)\}$ real-valued multiplications per fullband input sample.

TABLE I
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, $\mathbf{FW}_2\mathbf{F}^{-1}$ IS REPLACED BY THE IDENTITY MATRIX $\mathbf{I}$. THE FOURIER MATRIX $\mathbf{F}$ IS DEFINED IN (29)

Input signals	**Matrix sizes**
$\mathbf{X}(k) = \text{diag}\left\{\mathbf{F}\left[x(kN-N) \quad \cdots \quad x(kN+N-1)\right]^T\right\}$	$(2N\times 2N)$
$\mathbf{y}(k) = \left[y(kN) \quad \cdots \quad y(kN+N-1)\right]^T$	$(N\times 1)$
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	
$\mathbf{e}(k) = \mathbf{y}(k) - \mathbf{W}_1\mathbf{F}^{-1}\sum_{p=0}^{P-1}\mathbf{X}(k-p)\hat{\mathbf{H}}_p(k)$	$(N\times 1)$
$\mathbf{E}(k) = \mathbf{FW}_1^T\mathbf{e}(k),$	$(2N\times 1)$
$\hat{\mathbf{H}}_p(k+1) = \hat{\mathbf{H}}_p(k) + \mu\mathbf{FW}_2\mathbf{F}^{-1}\tilde{\mathbf{S}}^{-1}(k)\mathbf{X}^H(k-p)\mathbf{E}(k)$	$(2N\times 1)$
Definitions	
$\mathbf{e}(k) = \left[e(kN) \quad \cdots \quad e(kN+N-1)\right]^T$	$(N\times 1)$
$\hat{\mathbf{H}}_p(k) = \mathbf{F}\left[\hat{\mathbf{h}}_p^T(k) \quad \mathbf{0}_{1\times N}\right]^T$	$(2N\times 1)$
$\hat{\mathbf{h}}_p(k) = \left[\hat{h}_{pN}(k) \quad \cdots \quad \hat{h}_{pN+N-1}(k)\right]^T$	$(N\times 1)$
$\mathbf{W}_1 = \left[\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}\right],\ \mathbf{W}_2 = \text{diag}\left\{\left[\mathbf{1}_{1\times N} \quad \mathbf{0}_{1\times N}\right]\right\}$	

B. 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 [15], [26], [27]

$$X_m(k) = \sum_{n=0}^{2M-1} w(n)x(kM+n)\cos\left[\frac{\pi}{4M}(2n+1+M)\cdot(2m+1)\right], \qquad m = 0\cdots M-1 \quad (33)$$

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 $2M$

$$\begin{aligned}\tilde{X}_m(k) &= \tfrac{1}{2}e^{-j(\pi/4M)(1+M)(2m+1)}\cdot\sum_{n=0}^{2M-1}w(n)x(kM+n)\\ &\quad\cdot e^{-j(\pi n/2M)}e^{-j(2\pi nm/2M)},\\ X_m(k) &= \tilde{X}_m(k) + \tilde{X}_m^*(k), \qquad m = 0\cdots M-1 \end{aligned} \quad (34)$$

where $w(n)x(kM+n)e^{-j(\pi n/2M)}$ 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 overlap add method is used to reconstruct the fullband signal

$$\begin{aligned}\tilde{x}^{rec}(p,k) &= \frac{w(p)}{M}\sum_{m=0}^{M-1}X_m(k)\\ &\quad\cdot\cos\left[\frac{\pi}{4M}(2p+1+M)(2m+1)\right],\\ &\quad p = 0\cdots 2M-1,\\ x^{rec}(n) &= \tilde{x}^{rec}(n\%M+M,\ \lfloor n/M\rfloor)\\ &\quad+\tilde{x}^{rec}(n\%M,\ \lfloor n/M\rfloor+1)\end{aligned} \quad (35)$$

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)

$$w(n) = w(2M-n-1), \qquad n = 0\cdots 2M-1, \quad (36)$$

$$w^2(n) + w^2(n+M) = 2, \qquad n = 0\cdots 2M-1. \quad (37)$$

Two simple windows that satisfy (36) and (37) are

$$w_1(n) = 1, \qquad n = 0\cdots 2M-1 \quad (38)$$

$$w_2(n) = \sqrt{2}\sin\left(\frac{\pi}{2M}\left(n+\frac{1}{2}\right)\right), \qquad n = 0\cdots 2M-1. \quad (39)$$

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 [15].

In [28], 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.

C. Joint Transform Structure for Audio Coding and Echo Cancellation

In a standalone 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

$$\mathbf{E}(k) = \underbrace{\mathbf{F}[\mathbf{0}_{N\times N} \quad \mathbf{I}_{N\times N}]^T \mathbf{y}(k)}_{\mathbf{Y}(k)} - \underbrace{\mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}\mathbf{X}(k)\hat{\mathbf{H}}(k)}_{\hat{\mathbf{Y}}(k)}. \quad (40)$$

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 $\mathrm{diag}\{\mathbf{w}\}$, where $\mathbf{w} = [w(0)\cdots w(N-1)]$. Also in (30), we need to pre-multiply $\mathbf{y}$ with $\mathrm{diag}\{\mathbf{w}\}$. Equivalently, the two $\mathbf{I}_{N\times N}$ in (40) can be exchanged for $\mathrm{diag}\{\mathbf{w}\}$. 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 [15]. 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 I. 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 (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%.

V. 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[1] (NMSE) energy of the residual is used as performance index. The NMSE is given by

$$\mathrm{NMSE} = \frac{\mathrm{LPF}[e(n) - w(n)]^2}{\mathrm{LPF}[y(n) - w(n)]^2} \tag{41}$$

where w denotes the receiving room background noise signal and LPF denotes a lowpass filter; in this case it has a single real pole at 0.999.

In Table II, 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.

A. 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 four different systems considerations, listed below, and the NMSE of the first three 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 II-B 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 noncausal 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 III-B. 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 noncausal 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 noncausal filter taps, and this delay is as before 16 ms.

System 4: The delayless open-loop subband structure presented in [7]. The system has 64 subbands and a downsampling factor of 32. The fullband impulse response is 774 taps plus 250 noncausal taps. Complexity calculations are perfomed as in [7]. This system is given for complexity comparisons.

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

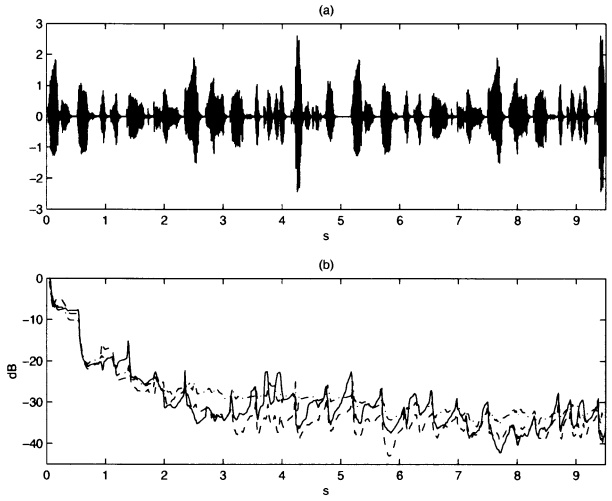

Fig. 7. (a) Echo signal, $y(n)$, used in all simulations. (b) The NMSE performance for three different systems: System 1a, fullband NLMS (solid line). System 2, subband system with noncritical EC filterbank (dashed line). System 3, subband system with modified filterbank from the MPEG 1 and 2 audio coder (dashed-dotted line).

TABLE II
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 V. The Contributions From All Components in Fig. 1 are Included

Filterbank based audio coder and EC

	Sys. 1a	Sys. 2	Sys. 3	Sys. 4
nbr. of subb., M^{EC}	–	64	64	64
downsamp. fac., r	–	48	32	32
complex. (mult.)	2085	327	406	461
delay (ms)	32	104	48	48

MDCT based audio coder and FDAF based EC

	Sys. 1b	Sys. 5
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 EC

	Sys. 1c	Sys. 6	Sys. 7
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†

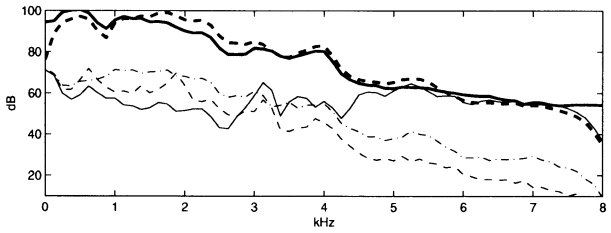

Fig. 8. Power spectra of the residual echo signals used to derive the NMSE plots in Fig. 7, averaged over the time interval 1.8 to 3.1 s. The thick solid lines depict the power spectrum of the transmission signal, $x(n)$, and the thick dashed line of the receiving room signal, $y(n)$. Other conditions same as in Fig. 7(b).

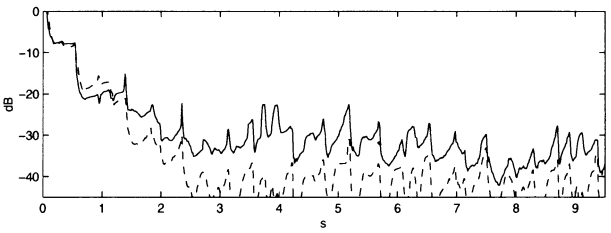

Fig. 9. NMSE performance: Fullband NLMS (System 1b, solid line), and System 5, an unconstrained FDAF (dashed line). The FDAF block size $N = 1024$ was used.

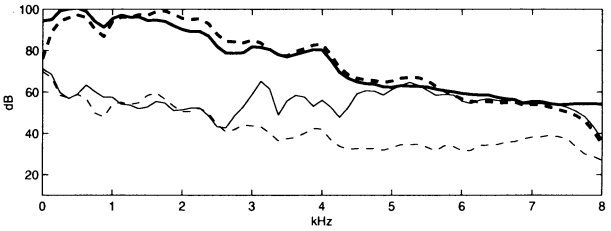

Fig. 10. Power spectra of the residual echo signals used to derive the NMSE in Fig. 9, other conditions same as in Fig. 8.

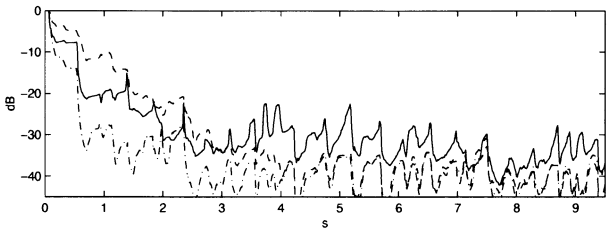

Fig. 11. NMSE performance: Fullband NLMS (System 1c, solid line), constrained FDAF with block overlap parameter $\alpha = 1$ (System 6, dashed line) and constrained FDAF with block overlap parameter $\alpha = 4$ (System 7, dashed-dotted line). The FDAF block size $N = 256$ and $P = 4$ adaptive taps per frequency bin were used.

Systems 1a to 3 appear to have similar NMSE performance. As was predicted in Section III-B, 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 II-B, see Fig. 7(b) at approximately the time instance 3 s.

In Section I 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 NMSE 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 II, the calculation complexity and the transmission delay for the considered systems are summarized.

B. 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 I. 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 nontransient 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 and calculation complexity caused by the coder will make this system differ from System 1a.

System 5: 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 I, 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 I is used.

In Fig. 9, the NMSEs for System 1b and 5 are depicted. The power spectra estimates are presented in Fig. 10.

Finally we have an audio coder with an MDCT size of $M = 256$, 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 [29]. 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 I 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 $M = 256$.

System 6: 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 7: Like System 6, 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, 6 and 7 are shown in Figs. 11 and 12.

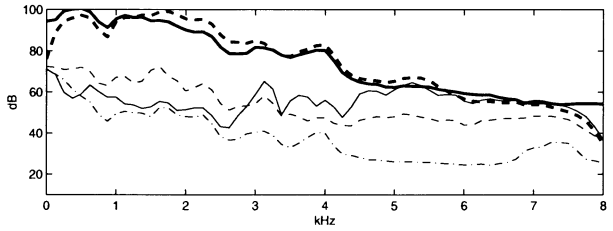

Fig. 12. Power spectra of the residual echo signals used to derive the NMSE in Fig. 11, other conditions same as in Fig. 8.

VI. 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 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 III-B 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.

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 IV-C 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 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 V.

Acknowledgment

The author would like to thank Dr. T. Gänsler, Bell Labs, for constructive discussions.

References

[1] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. XLVI, pp. 497–510, Mar. 1967.

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

[3] D. R. Morgan, "Slow asymtotic convergence of LMS acoustic echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 126–136, Mar. 1995.

[4] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[5] D. Mansour and A. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 726–734, Oct. 1982.

[6] P. Sommen, P. V. Gerwen, H. Kotmans, and A. Janssen, "Convergence analysis of a frequency-domain adaptive filter with exponetial power averaging and generalized window function," *IEEE Trans. Circuits Syst.*, vol. 34, pp. 788–798, July 1987.

[7] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Trans. Signal Processing*, vol. 43, pp. 1819–1830, Aug. 1995.

[8] R. Merched *et al.*, "A new delayless subband adaptive filter structure," *IEEE Trans. Signal Processing*, vol. 47, pp. 1580–1591, June 1999.

[9] N. Hirayama *et al.*, "Delayless subband adaptive filtering using the hadamard transform," *IEEE Trans. Signal Processing*, vol. 47, pp. 1731–1734, June 1999.

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

[11] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, vol. 14, pp. 59–81, Sept. 1997.

[12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[13] ISO/IEC 11 172–3, "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.

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

[15] M. Bosi *et al.*, "ISO/ICE MPEG-2 advanced audio coding," *J. Audio Eng. Soc.*, vol. 45, pp. 789–814, Oct. 1997.

[16] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Processing*, vol. 40, pp. 1862–1875, Aug. 1992.

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

[18] 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*, vol. 49, pp. 513–523, July 2001.

[19] H. Sorensen, D. Jones, M. Heideman, and S. Burrus, "Real-values fast Fourier transform algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 849–863, June 1987.

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

[21] M. Vetterli and H. Nussbaumer, "Simple FFT and DCT algorithms with reduced number of operations," *Signal Process.*, vol. 6, pp. 267–278, Aug. 1984.

[22] J. Proakis and D. Manolakis, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[23] P. Eneroth, "Stereophonic Acoustic Echo Cancellation: Theory and Implementation," Ph.D. dissertation, Lund Univ., Lund, Sweden, 2001.

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

[25] 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*, vol. 2, 2000, pp. 789–792.

[26] J. Princen and A. Bradley, "Analysis/synthesis filterbank designed based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1153–1161, Oct. 1986.

[27] 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.

[28] M. Iwadare *et al.*, "A 128 kb/s Hi-Fi audio CODEC based on adaptive transform coding with adaptive block size MDCT," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 138–144, Jan. 1992.

[29] S. Vernon, "Design and implementation of AC–3 coders," *IEEE Trans. Consumer Electron.*, vol. 41, p. , Aug. 1995.

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

During the Fall of 1998 and 1999 he was with Bell Labs, Lucent Technologies, Murray Hill, NJ, and from 2001 until July 2002 he was with Telia Research. Since August 2002, he has been the Technical Manager for the Speech and Audio Group at Ericsson Mobile Platforms, Lund. His research interests include adaptive filtering, subband signal processing, audio processing and real-time implementations.