FILTER BANKS IN PERCEPTUAL AUDIO CODING

MARINA BOSI

Digital Theater Systems (DTS), Inc., Los Angeles, California, USA
mab@dtstech.com

This paper presents an overview of the filter-bank technologies used in the time to frequency mapping of perceptual audio coders. Filter banks allow for signal decorrelation and therefore provide a framework for removing redundancy in an audio signal. In addition, irrelevant components of the signal can be separated from relevant ones based on models of human perception. In this paper, filter-bank design and implementation issues are discussed. Time versus frequency resolution and filter banks in which the time-invariant constraint is relaxed will be described. Finally other methods for increasing redundancy extraction, such as prediction, will be examined.

INTRODUCTION

The first stage in perceptual audio coding schemes is usually represented by the time to frequency mapping of audio signals. The basic idea is to filter the signal into its components in various frequency bands. By subdividing the signal into its frequency components and representing the signal by its frequency component parameters, a great reduction in the amount of data needed to reproduce the audio signal can be achieved.

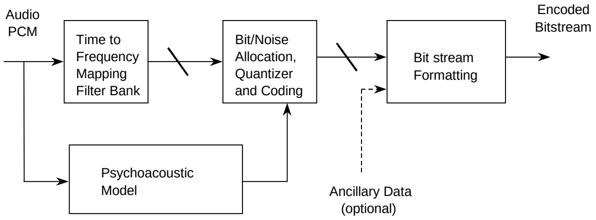

Audio PCM
Time to Frequency Mapping Filter Bank
Bit/Noise Allocation, Quantizer and Coding
Bit stream Formatting
Encoded Bitstream
Psychoacoustic Model
Ancillary Data (optional)

Figure1: Building blocks for a perceptual audio coder.

Consider for example a sine wave and its representation in the frequency domain. While only three parameters, namely frequency, phase, and amplitude, for each block of data fully describe the sine wave in the frequency domain, a large number of PCM data is needed to describe this simple signal in the time domain. This example clearly shows how, by filtering the signal, redundancies can be easily extracted from the audio signals.

In general, although audio signals will not exhibit strict periodicity as in the simple sine wave example, it can be shown that audio signals are quasi-stationary and that they can be modelled by using short-term spectrum analysis.

Once the signal is represented in the time-frequency domain, the number of bits used to encode each frequency component can be adjusted so that greater encoding accuracy can be placed in frequencies where it is most needed. For example, if we can break the signal into its energy per each critical band, we can apply masking models to separate irrelevant elements of the signal from relevant ones.

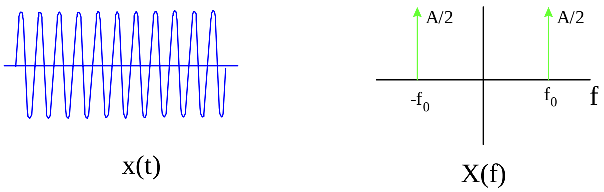


Figure 2: Time vs. frequency representation of a cosine.

A variety of time to frequency mapping algorithms, which differ by the degree to which they allow for source component separation and source redundancy extraction, are available. Just to mentioned only a few, discrete Fourier transform (DFT), discrete cosine transform (DCT), quadrature mirror filters (QMF), pseudo QMF (PQMF), modified DCT (MDCT), hybrid filter banks, wavelet, etc. are time to frequency mapping techniques found in literature for perceptual audio coding [1-9]. In the following Sections of this paper the most commonly used filter banks in perceptual audio coding will be described from the mathematical and application point of view. Issues and parameters related to redundancy extraction as well as design consideration will be examined. Finally other methods for increasing redundancy extraction, such as prediction, will be discussed.

1. WHY USE FILTER BANKS IN PERCEPTUAL AUDIO CODING?

Given the general assumption that audio signal spectra vary slowly with time, i.e. they are statistically stationary, and they can be described by short-term analysis, it is both practical and efficient to represent the signal in the frequency domain. In addition, it is meaningful from the human-perception point of view to be able to separately manipulate spectral components of the signal.

The filter bank framework provides the best medium for the removal of redundancy, i.e. information that is not necessary to uniquely identify the signal, and irrelevancies, i.e. information that is perceptually not important.

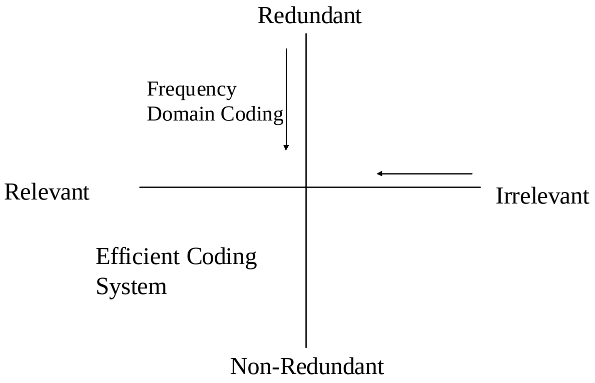


Figure 3: Characteristics of signal representation and coding efficiency [32].

1.1 Redundancy extraction

As mentioned in the Introduction, the basic idea in a data reduction scheme is to filter the signal into its components in various frequency bands (see Figure 4).

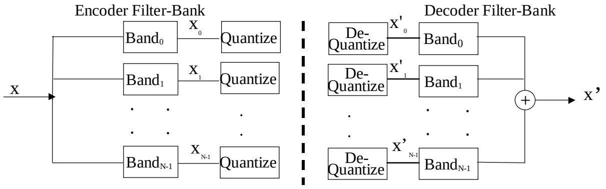


Figure 4: Basic idea in data rate reduction schemes.

The signal is then quantised in the frequency domain and the total bit pool is allocated dynamically depending on the energy of each spectrum component and its relevancy. Let us assume for a moment that the signal frequency components have equal energy and they populate the full spectrum. Let us also assume that we are not exploiting psychoacoustics models but we are concentrating on redundancies removal only. In this particular case there is really no gain in redistributing the bit pool throughout the spectrum because each component demands the same number of bits.

On the other hand, if we assume that the signal spectrum is coloured, e.g. the spectral components at low frequencies are stronger, then there is an increase in coding gain by redistributing the bit pool throughout the spectrum.

Luckily, the latter case is the most common. In this case the signal contains redundancies. These redundancies can be more or less efficiently removed. The efficiency of the removal depends upon the characteristics of the filter bank.

A measure of the redundancies present in the signal representation is given by the spectral flatness measure (sfm): the flatter is the spectrum, the less redundant is the signal. The sfm is given by the ratio of the geometric mean to the average of the power spectral density of the signal. Low sfm implies potential high coding gains [1].

By comparing various bit allocations at a given level of average block distortion $<q^2>$, where $q_k=x_k-Q^{-1}(Q(x_k))$ is the quantization noise for each spectral component and k is the spectral component index, we can find a method that optimally allocates bits through the spectrum, where our ultimate goal is to localize the quantization noise below the masking thresholds.

We can increase the coding gain with respect to the PCM coding gain if we can find a set of R_k, where R_k represents the number of bits used to code the spectral line of index k, that minimizes the error[1]:

$$< q^2 >= \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{x_k^2}{3\cdot 2^{2R_k}}\right) \qquad (1)$$

such that

$$\frac{1}{N}\sum_{k=0}^{N-1}R_k = R \qquad (2)$$

where N is the number of spectral lines, and R is the average number of bits per spectral line available.

This is a problem of constrained minimisation that can be solved by using a Lagrange multiplier λ, to enforce

[1] We are assuming error-free transmission, non-overlapping equal-width sub-bands, and the use of PCM coding of individual sub-bands with a midrise quantizer with maximum non-overload value equal to x_k.

the average bit rate constraint as specified in (2)[2]. By taking the derivative with respect to each R_k and with respect to λ and by solving the resulting equations for R_k and then enforcing average bit rate constraint we obtain:

$$R_k = R + \tfrac{1}{2}\log_2(x_k^2) - \tfrac{1}{2}\log_2\left(\prod_{j=0}^{N-1} x_j^2\right)^{\frac{1}{N}} \qquad (3)$$

From (3) it is apparent that, for each block of samples, a bit allocation based upon the spectral energy distribution of the signal will introduce an improvement with the respect to uniform allocation, when the geometric mean of the signal power spectral density is much smaller than the average of the signal power spectral density.

The ratio of the geometric mean of the signal power spectral density to the average of the signal power spectral density is the sfm of the signal where:

$$sfm = \frac{\left(\prod_{k=0}^{N-1} x_k^2\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} x_k^2} \qquad (4).$$

Notice that the sfm varies between 0 and 1; sfm = 1 implies a signal with a flat spectrum, and no coding gain with respect to uniform distribution of bits throughout the block can be achieved since, by substituting sfm = 1 in (3) we obtain

$$R_k = R\,.$$

Notice also that the sfm depends not only on the spectral energy distribution of the signal but also on the resolution of the filter bank, i.e. the total number of the frequency lines, N, or block length. If N is >>2, for a given signal, then the sfm decreases by increasing the block size N.

1.2 Irrelevancy extraction

In perceptual audio coding, the goal is not just to extract redundancy from the source, but also to isolate the irrelevant parts of the signal spectrum. This translates in not just trying to minimize the average error power $<q^2>$ per block, but trying to get the quantization noise below the masking curves generated by the signal under examination.

For components above the masking curve, i.e. relevant signals, this means that we want to maximize the difference between the signal to noise ratio (SNR), and the signal to mask ratio (SMR), or equivalently, minimize SMR-SNR, where

$$SNR = 10\log\frac{\left\langle x^2\right\rangle}{\left\langle q^2\right\rangle}$$

and

$$SMR = 10\log\frac{\left\langle x^2\right\rangle}{\left\langle M^2\right\rangle}$$

with M_k corresponding to the masking threshold value for the k component of the block spectrum. This differs from the minimization problem described in (1) in that we need to minimize the error weighted by the masking factor, i.e.:

$$< {}^{q^2}\!\!\diagup_{M^2} >= \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{{}^{x_k^2}\!\!\diagup_{M_k^2}}{3\cdot 2^{2R_k}}\right) \qquad (5)$$

with the same constraint as described in (2). The resulting optimal bit allocation leads to:

$$R_k = R + \tfrac{1}{2}\log_2\left({}^{x_k^2}\!\!\diagup_{M_k^2}\right) - \tfrac{1}{2}\log_2\left(\prod_{j=0}^{N-1} {}^{x_j^2}\!\!\diagup_{M_j^2}\right)^{\frac{1}{N}} \qquad (6).$$

The "perceptual" sfm can then be described as:

$$psfm = \frac{\left(\prod_{k=0}^{N-1} {}^{x_k^2}\!\!\diagup_{M_k^2}\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} {}^{x_k^2}\!\!\diagup_{M_k^2}} \qquad (7)$$

Notice that the psfm depends on the spectral energy distribution of the signal weighted by the masking energy distribution. In this case, depending on the characteristics of the input signal, increasing the frequency resolution of the filter bank may or may not imply an increase in the coding gain. While for signals with steady state characteristics increasing the frequency resolution of the filter bank causes an increase in the coding gain, this is not true for transients. Work done by J. Johnston [10], showed that for tonal signals like the harpsichord, increasing the filter-bank block length is reflected in an increase in coding gain, while for transient-like signals, e.g. castanets, the coding gain tends to decrease by increasing the block size.

[2] In our derivations we assume that we would always get R_k>=0; the above algorithm, however, will sometimes give us negative values of R_k when x_k is much below its geometric mean. (We really should have included Kuhn-Tucker multipliers to keep all of the R_k non-negative.) In practice, one usually rounds those R_ks to zero and takes bits away from other parts of the spectrum. In this case we use an approximate solution allocating bits one by one locally (e.g. water filling algorithms, etc.).

2. FILTER-BANK DESIGN CONSIDERATIONS

A number of factors come into play in the design of the filter banks in perceptual audio coding. Firstly, as discussed in previous Sections, we would like to optimally separate the different spectral components so that the perceptual coding gain can be maximised. Since we will be performing short-time analysis/synthesis of the signals, we would like to minimise the audibility of blocking artefacts both in terms of boundary discontinuities and pre-echo effects.

Secondly, given that the ultimate goal is to decrease the data rate while maintaining the quality of the audio signal, critically sampled systems are desirable. In these systems, the overall rate at the output of the analysis stage equals the overall rate at the input of the analysis stage.

Thirdly, while this is not a strict requirement, most of the perceptual audio coders currently in use employ perfect reconstruction, PR, or “nearly” PR filter banks.

Finally, depending on the type of application, time delay [11] and computational complexity play a role in the design of filter banks.

2.1 Signal components separation

One of the basic assumptions is that audio signals can be modelled by using short-term spectrum analysis. In order to reduce the distortion introduced by boundary discontinuities, the input to the filter bank is typically windowed and overlapped. The window shape plays an important role in the spectral separation of the signal.

In Figure 5 a comparison of the minimum masking threshold to the 1024-line MDCT frequency selectivity of the Kaiser Bessel derived [12], KBD, and sine windows at 48 kHz is shown [13]. Notice how, while the close selectivity of the sine window is better than the close selectivity of the KBD window (the alpha parameter selected for the KBD window is four in this case), the ultimate rejection of the sine window falls short of the requirement for the minimum masking threshold. The KBD window satisfies much better this requirement. Depending on the signal characteristics, either the sine or the KBD window may provide better resolution for the signal representation. If we consider, for example, a signal with a closely spaced picket-fence spectral structure, then close selectivity plays a more important role than ultimate rejection, given the superimposition of masking effects from different parts of the signal spectrum. On the other hand, in a signal that exhibits wide separation among its components, higher ultimate rejection allows for a better exploitation of the signal masking.

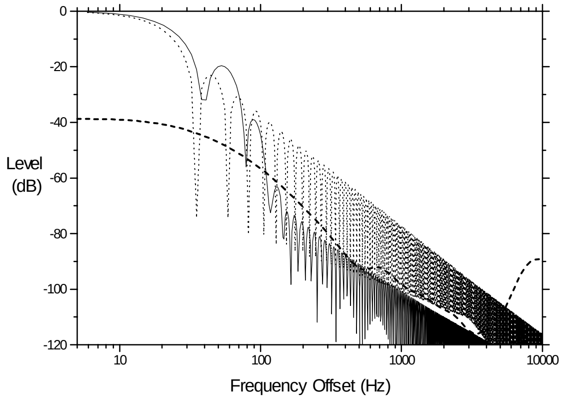


Figure 5: Comparison of the masking template to the 1024-line MDCT frequency selectivity of the KBD and sine windows at 48 kHz [13]. (The solid line represents the KBD window, the dotted line represents the sine window, and the dashed line represents the minimum-masking template).

No single window provides optimal resolution for all signals. The encoder should adaptively select the window shape of the filter bank based on the signal characteristics. An example of a window shape sequence where the KBD window is alternate with the sine window is shown in Figure 6 [13].

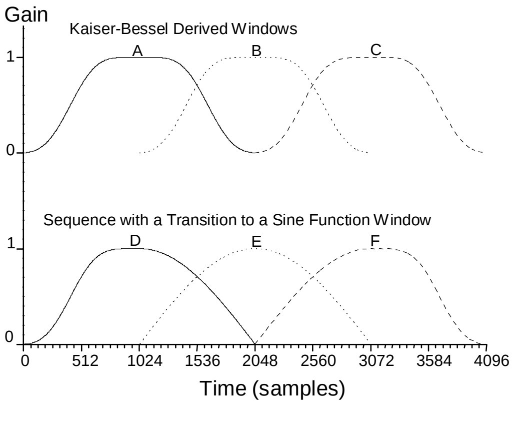


Figure 6: Example of different window shapes sequence [13].

From the discussion in Section 1, it is apparent that perhaps the most significant decision in the design of the filter bank is the window size, since this parameter is directly related to the time/frequency resolution and the coding gain of the system. If we examine (7) we notice that, at least for steady state signals, the longer is block size the higher is the coding gain. Work from Johnston [10] showed that, by comparing the average energy for a variety of audio signals sampled at 44.1 kHz from consecutive blocks and finding the block size that provides the most stationary energy values and the average coding gain, the optimal block size corresponds to around 1024 frequency samples. Above this number the very small increase in coding gain does not justify the increased delay and complexity introduced.

On the other hand, from [10] we also deducted that certain signals, e.g. castanets, require much shorter block lengths for optimal coding gains, typically around 64 frequency samples.

Once again, no single, fixed time/frequency resolution can optimally represent all type of audio signals. By relaxing the time-invariant constraint of filter banks we avoid to force a possibly unsatisfactory compromise between time versus frequency resolution. In Figure 7 an example of adaptive block size is shown for the MDCT [14, 13].

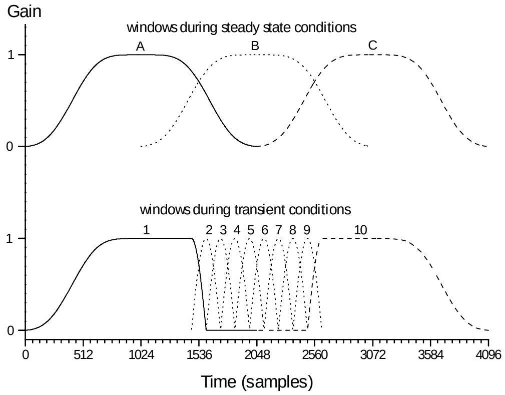
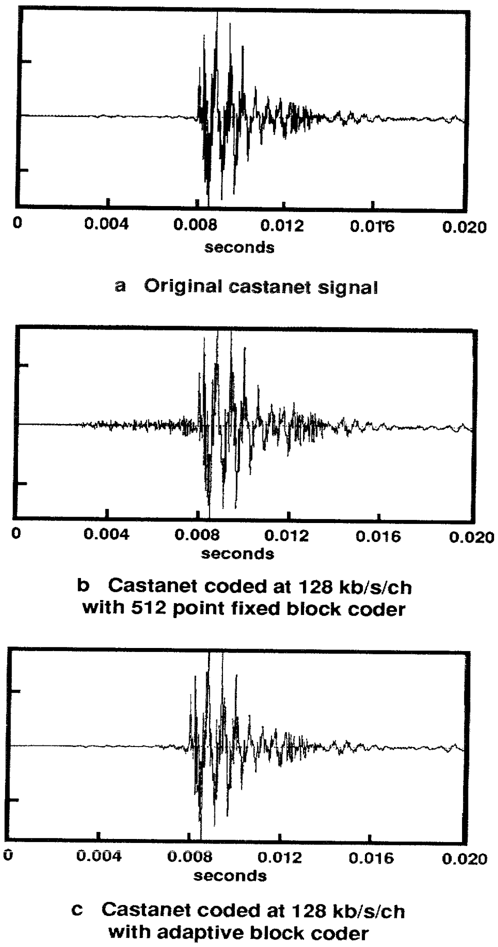


Figure 7: Adaptive block size for the MDCT [13].

It should be mentioned that, by employing a shorter block size for transient-like signals, blocking artefacts like pre-echo are greatly attenuated. In Figure 8 a comparison of the quantization noise for different block sizes for a castanets hit at a fixed data rate is shown. Notice how the spreading in time of the quantization error before the onset of the attack (pre-echo) is decreased and closer localised in time to the signal in the case of the adapted shorter block (N= 64) in Figure 8c than in the case of the longer one (N = 256) in Figure 8b. Based on temporal masking models [15], only a few milliseconds of quantization noise that precedes the onset of a signal can be effectively masked (pre-masking). Typically, the decision to increase the time resolution in the signal representation is made at the encoder stage and it is based upon the characteristics of the input signal and temporal masking models. This information is then conveyed to the analysis filter bank and to the decoder as a control parameter for the synthesis filter bank.



a Original castanet signal

b Castanet coded at 128 kb/s/ch with 512 point fixed block coder

c Castanet coded at 128 kb/s/ch with adaptive block coder

Figure 8: Comparison between time-variant MDCT- and time-invariant MDCT coded castanets signal.

2.2 Critically sampled N-channel filter banks

A simple example of a critically sampled N-channel filter bank is shown in Figure 9. The input signal is filtered by N band-pass filters, H_k[3]. Each band-pass filter output is then sub-sampled by a factor of N, i.e. it is critically sampled at a rate that is twice the nominal bandwidth of each band-pass filter. In the synthesis filter, the signal is up-sampled and filtered by the set of the N G_k filters. If perfect reconstruction filters can be applied, in absence of quantization, the sum of the

[3] One could apply different structures, e.g. tree-structures cascades of two-band filters, etc.

output of the G_k filters equals the delayed original signal.

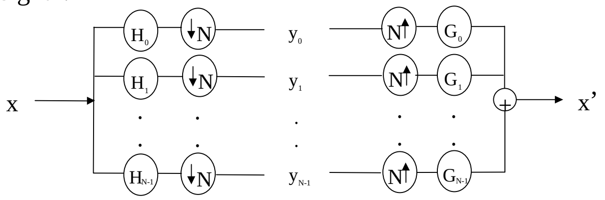


Figure 9: Critically sampled N-channel filter bank.

The down-sampling operation can introduce aliasing in the signal spectrum if there is overlap between adjacent band-pass filters, while the up-sampling operation can introduce imaging. With an appropriate choice of analysis/synthesis filters these distortions in the spectrum cancel each other in the synthesis stage after all components are added together.

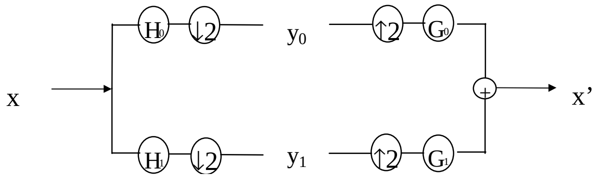


Figure 10: Critically sampled two-channel filter bank.

In the simple case of N = 2 (Figure 10), we can write the z-Transform of the output of the analysis filters y_k, $Y_k(z)$, and of the output of the synthesis stage $X'(z)$ as

$$Y_k(z) = 1/2\ [H_k(z^{1/2})X(z^{1/2}) + H_k(-z^{1/2})X(-z^{1/2})] \quad k = 0,1$$

$$\begin{aligned} X'(z) &= Y_0(z^2)\ G_0(z) + Y_1(z^2)\ G_1(z) \\ &= 1/2\ [H_0(z)\ G_0(z) + H_1(z)\ G_1(z)]\ X(z) + \\ &\quad 1/2\ [H_0(-z)\ G_0(z) + H_1(-z)\ G_1(z)]\ X(-z). \end{aligned}$$

The aliasing component X(-z) can be cancelled when:

$G_0(z) = -H_1(-z)$ and $G_1(z) = H_0(-z)$.

Or, equivalently

$g_0(n) = -\ h_0(L-1-n)$ and $g_1(n)=(-1)^n h_0(n)$

where L is the even length of the filter.

This leaves us with

$$X'(z)= 1/2\ [-H_0(z)\ H_1(-z) + H_1(z)\ H_0(-z)]\ X(z).$$

If we also require the quadrature mirror filter bank, QMF, condition that

$$H_1(z) = -H_0(-z)$$

or equivalently

$$h_1(n) = (-1)^n h_0(n)$$

we find that

$$X'(z)= 1/2\ [H_0(z)^2 - H_0(-z)^2\ ]\ X(z).$$

When the analysis filters satisfy the QMF conditions and we require that

$$H_0(z)^2 - H_0(-z)^2 = 2z^{-D} \tag{8}$$

where D is an appropriate delay, we have a perfect reconstruction filter bank. The implication of the QMF condition is that

$$H_1(e^{jw}) = -H_0(e^{j(w+\pi)})$$

Thus, if H_0 is a low pass filter, then H_1 is a high pass filter with symmetrical response. In general, no filter $H_0(z)$ with finite order greater than one can exactly satisfy the PR requirement (8) (the only exception is the Haar filter). However, we can find FIR filters that reasonably well approximate the QMF PR requirement.

In general, in perceptual audio coding we are interested in filter banks with a number of frequency channels N>>2. A generalization of the QMF properties for filters with N>2 channels was first introduced by Nussbaumer as pseudo-QMF, PQMF [16]. The basic idea is that the filters are designed so that aliasing from adjacent bands is exactly cancelled but aliasing from next-neighbor bands is ignored. Nussbaumer suggested that the band-pass filters, $h_k(n)$, where n is the time index and k is the frequency index, be a modulated version of a single low pass filter with bandwidth fs/N.

$$h_k(n) = h(n)\cos[\pi/N(k+1/2)(n-(L-1)/2)+\varphi_k)] \tag{9}$$

$k= 0, 1, \ldots N-1$ $\qquad$ $n= 0, 1, \ldots L-1$

where N is the number of frequency channels and L is the length of the filters h_k. The reconstruction filters can be derived from the analysis filter as follows:

$$h_k(n) = g_k(L-1-n).$$

The low pass filter prototype should also satisfy (as much as possible) the PR condition from (8):

$$|H\,(e^{j\omega})|^2 + |H(e^{j(\pi/K-\ \omega)})|^2 = 2 \qquad 0<|\omega|< \pi/2N$$

$$|H(e^{j\omega})|^2 = 0 \qquad\qquad |\omega| > \pi/N$$

These filters are computationally very efficient since they can be realized via an FFT and are of moderate complexity and low delay. A polyphase QMF with length L = 512, number of channels N =32, and φ_k= -N/2, is used in the MPEG Audio coding schemes [4] (see Section 3.1).

Another example of critically sampled filter bank is the MDCT [3]. This transform is based on time domain aliasing cancellation (TDAC) and was first introduced by Princen and Bradley [16]. The time-invariant TDAC transform provides a critically sampled system with 50% overlap between adjacent windows.

In the analysis stage, N new input time samples are buffered and windowed with a window of length 2N (see Figure 11). The signal is then mapped from time to frequency domain by using the MDCT (oddly stacked TDAC) or alternating an MDCT with a modified discrete sine transform, MDST (evenly stacked TDAC). The inverse-transformed signal contains time aliasing distortion, which, in absence of quantization, is cancelled during the window and overlap-add stage.

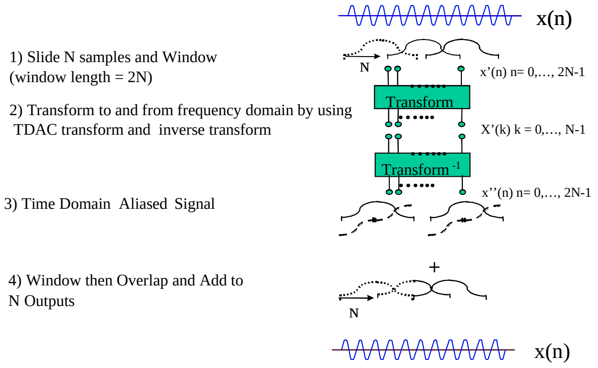

x(n)
1) Slide N samples and Window
(window length = 2N)
N
x'(n) n= 0,…, 2N-1
Transform
2) Transform to and from frequency domain by using
TDAC transform and inverse transform
X'(k) k = 0,…, N-1
Transform -1
x''(n) n= 0,…, 2N-1
3) Time Domain Aliased Signal
+
4) Window then Overlap and Add to
N Outputs
N
x(n)

Figure 11: TDAC Transform [3, 17].

The forward TDAC transform can be generalised as follows [12]:

$$X_i(k) = \sum_{n=0}^{2N-1} x_i(n) e^{-j(2\pi/2N)(k+k_0)(n+n_0)} \quad k = 0,1,\ldots,2N-1 \quad (10)$$

where

$x_i(n)$ is the windowed input sequence of 2N samples coefficients for the i-th block;

$X_i(k)$ is the sequence of 2N frequency coefficients for the i-th block;

k_0 is a frequency offset in the transform kernel;

$$k_0 = \begin{matrix} 1/2 & \textit{for the OTDAC} \\ 0 & \textit{for the ETDAC} \end{matrix}$$

n_0 is a time offset that allows for the cancellation of the time aliasing introduced in the signal; in general n_0 depends on the length of the overlapping region with the next block of samples; in the case of the time-invariant TDAC transforms we have:

$$n_0 = \frac{(N+1)}{2}$$

Accordingly, the inverse TDAC transform can be generalised as follows:

$$x_i'(n) = \frac{1}{N} \sum_{n=0}^{2N-1} X_i(k) e^{-j(2\pi/2N)(k+k_0)(n+n_0)} \quad k = 0,1,\ldots,2N-1 \quad (11)$$

where x'_i equals the delayed, time-aliased input sequence.

The ETDAC and OTDAC MDCT kernel can be obtained from (10) and (11) by taking the real part; the ETDAC MDST kernel can be obtained from (10) and (11) by taking the imaginary part. If x(n) is real, then the MDCT is odd-symmetric and the MDST is even-symmetric, therefore only N independent frequency coefficients are generated for each transform block. In absence of quantization, after the window and overlap-add stage of the time-invariant TDAC, the output signal becomes an exact delayed replica of the input signal provided that analysis and synthesis windows satisfy the following requirement:

$$W^a(n)W^s(n)+W^a(N+n)W^s(N+n)=1 \quad n=0,\ 1,\ \ldots,\ N-1 \quad (12)$$[4]

The TDAC transforms can be efficiently implemented via an FFT kernel; fast implementations of the MDCT exist in literature see for example [18]. For power of two block lengths[5], the number of complex multiplies /additions is $N/2 + N/2 \log_2(N/2)$, where N is the number of frequency channels. The ETDAC is used in coding schemes like AC-2 [12]; the OTDAC is used in MPEG Audio [4, 13, 19-22] , AC-3 [12], PAC [23], Twin VQ [30], etc..

2.3 A unified approach

While historically the PQMF and the MDCT were developed independently, Malvar [18] showed how these approaches can be unified in the frame of the lapped orthogonal transforms, LOT. Given a number of frequency channels N, by appropriately selecting the length, L, and phase, φ_k, of the filters $h_k(n)$ in (9) and

[4] Typically the same window is employed for both the analysis and synthesis stage. Notice that both the sine window and the KDB window satisfy condition (12).

[5] In the case of non-power of two block lengths, usually the FFT kernel is factorised into smaller, power of two length FFT, see for example the MPEG Layer III implementation [4], thus requiring a slightly higher number of multiplies/additions.

imposing the following conditions on the prototype filter:

$h(2N-1-n) = h(n)$ and

$h(n)^2 + h(n+N)^2 = 2$

not only perfect reconstruction is achieved, but (9) becomes equivalent to the OTDAC MDCT expression in (10) where the analysis window is identical to the synthesis window. By setting the filter length L = 2N, and $\varphi_k=(k+1/2)(2+1)\pi/2$, we obtain

$$h_k(n) = h(n)\cos[\pi/N(k+1/2)(n+(N+1)/2)] \quad (13)$$

which is equivalent to real part of (10) when $k_0 = ½$, i.e. equivalent to the OTDAC MDCT expression.

3. PERCEPTUAL AUDIO CODING FILTER-BANK IMPLEMENTATIONS

In this Section examples of different types of filter banks commonly employed in perceptual audio coding are briefly discussed. Starting with MPEG Audio and then examining other schemes in the marketplace, we will compare the different characteristics of the design, performance and implementation.

3.1 MPEG Audio

The Moving Picture Expert Group (MPEG) was established in '88 in the framework of the joint ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) Technical Committee, JTC 1, on information technology with the mandate to develop standards for coded representation of moving pictures, associated audio, and their combination. Three phases of MPEG development resulted in the following standards:

- MPEG-1, coding up to 1.5 Mb/s (ISO/IEC 11172)
- MPEG-2, coding up to 10 Mb/s (ISO/IEC 13818)
- MPEG-4, low bit rate coding of audio visual (ISO/IEC 14496)

MPEG Audio systems evolved from one/two-channel audio signal systems compatible with the CD format (i.e. 16 bits per sample equivalent) and sampling frequencies of 32, 44.1, and 48 kHz to systems with a configuration of up to 48 audio channels, sample resolution of 24 bits and sampling rates between 8 and 96 kHz (MPEG-2 Advanced Audio Coding, AAC). In its first phase of work, MPEG-1 Audio Layers I, II, and III (ISO/IEC 11172-3) were finalised in 1992, including the syntax of the coded bitstream, the decoding process, and compliance test vectors. Later, in 1994, MPEG-2 specified the multichannel extensions (up to 5.1 audio channels) of Layers I, II, and III and lower sampling frequencies than 32, 44.1, and 48 kHz, namely 16, 22.05, 24 kHz (ISO/IEC 13818-3). In 1997, the MPEG-2 AAC specifications were finalised (ISO/IEC 13818-7); AAC constitutes also the kernel of the MPEG-4 time to frequency mapping audio coding schemes (ISO/IEC 14496-3).

3.1.1 MPEG-1, 2 Layers I and II

The basic structure of the MPEG-1 and MPEG-2 perceptual audio coders is practically identical. Specifically, the same filter-bank structure per audio channel is employed. The kernel of the filter bank is a PQMF structure as per (9), with L = 512, N =32, and φ_k= -N/2. This kernel structure is the same for all Layers (including Layer III, although increased frequency resolution is achieved by cascading the kernel with an MDCT, see the Section 3.1.2). At a sampling rate of 48 kHz, each band has a width of 750 Hz and a time resolution of 0.66 ms. At this sampling rate, the prototype filter impulse response is 10.66 ms. The prototype filter was designed to achieve a stop-band attenuation of about 120 dB [24]. Without quantization, the composite response of the analysis and synthesis filter bank has a ripple of less than 0.07 dB [24].

While the design of this PQMF provides very good time resolution with a relatively simple structure, there are some shortcomings. Firstly, adjacent bands have significant frequency overlap, i.e. a signal at a single frequency can affect two adjacent frequency bands. Secondly, the width of the frequency band is much larger than critical bandwidth values for frequencies below 2000 Hz, therefore the signal representation does not provide sufficient separation for optimal bit distribution as described in (6). The implementation described in the ISO/IEC 11171-3 and 13818-3 specifications is very similar to the one described in [25] and requires about 80 multiplies and additions per sample. Recent publications showed a reduction in computation of a factor larger than 10 [26].

3.1.2 MPEG-1, 2 Layer III

In Layer III the filter bank is a hybrid consisting of the 32-channel PQMF as described in the previous Section followed by a time-variant MDCT [19]. The MDCT filter bank consists of 18-frequency lines MDCT for steady state signals or 6-frequency lines MDCT for transient-like signals. At a sampling rate of 48 kHz, the frequency resolution of the Layer III filter bank is 41.66 Hz and the time resolution is 4 ms. The impulse response is 34.66 ms for long blocks and 18.66 ms for short blocks.

While the increased frequency resolution allows for an increase in the coding gain of Layer III, there are still shortcomings in this filter bank. As mentioned in the previous Section, in the kernel PQMF there is overlapping of adjacent frequency bands, which in turn causes potential aliasing after the sub-sampling of the analysis, filters output. This aliased signal is then processed by the TDAC transform. While in Layer III there is a mechanism to attenuate this distortion [4], it cannot, however, be completely eliminated. Another short coming of the Layer III filter bank is the relatively long block size which can cause spreading of the quantization noise in the time domain region where it is not masked (pre-echo); the dynamically adaptive MDCT tends to mitigate this effect.

3.1.3 MPEG-2, 4 AAC

The MPEG AAC[6] filter bank utilises a 1024-point MDCT for steady state signals or a 128-point MDCT for transient-like signals. In addition, the MPEG AAC filter bank allows for dynamically switching between different window shapes, namely between the sine and the KBD windows [13, 21]. At a sampling rate of 48 kHz, the frequency resolution of the AAC filter bank is 23.4 Hz and the time resolution is 2.66 ms. The impulse response is 42.66 ms for long blocks and 5.33 ms for short blocks.

The AAC filter bank provides excellent frequency selectivity while the time selectivity could be higher, ideally about 1.3 ms (see Sections 1 and 2). In general the AAC filter bank, with its high degree of adaptability to the characteristics of the input signal, performs very well.

3.1.4 MPEG-4 Twin VQ

The MPEG-4 Twin VQ scheme constitutes an alternative quantization/coding scheme to the AAC kernel for scalable audio coding down to 6 kb/s [22]. MPEG-4 Twin VQ has been harmonised with the MPEG-4 AAC audio coder to use the same filter bank and other coding tools [33]. The MPEG-4 Twin VQ filter bank utilises a 1024-point MDCT for steady state signals or a 128-point MDCT for transient-like signals. At a sampling rate of 48 kHz, the frequency resolution of the Twin VQ filter bank is 23.4 Hz and the time resolution is 2.66 ms. The impulse response is 42.66 ms for long blocks and 5.33 ms for short blocks.

In addition, Twin VQ normalises the spectral coefficients prior to the quantization stage. The “flattening” of the spectral coefficients includes a linear predictive coding spectral envelope estimation (see also Section 4).

3.2 Examples of other coders in the marketplace

3.2.1 AC-3

The AC-3 filter bank utilises a 256-point MDCT for steady state signals or 128-frequency lines MDCT for transient-like signals. The block switching mechanism is different from the one shown in Figure 7, in that the time offset of the MDCT basis, n_0, is modified along with the block length and the window shape in the transient adapting transform [12, 5]. While the block switching mechanism in AC-3 is computationally relatively simple, it is somewhat sub-optimal in terms of window shape selectivity. At a sampling rate of 48 kHz, the frequency resolution of the AC-3 filter bank is 93.75 Hz and the time resolution is 2.66 ms. The impulse response is 10.66 ms for long blocks and 5.33 ms for short blocks.

3.2.2 ATRAC

The ATRAC system is optimised for use in MiniDisc recording. The ATRAC filter bank consists of a gain control stage for pre-echo control cascaded with a PQF and an MDCT stage [6]. In addition, window switching is employed to adapt the filter bank resolution [27]. At a sampling rate of 44.1 kHz, each band has a width of 43.07 Hz and a time resolution of down to 1.45 ms at high frequencies.

3.2.3 DTS

The DTS filter bank consists of a 512-tap PQMF with 32-frequency channels [28]. At a sampling rate of 48 kHz, each band has a width of 750 Hz and a time resolution of 0.66 ms. In order to increase the redundancy extraction, the PQMF stage is cascaded with a linear prediction stage (see Section 4). For signals with non-flat spectrum, by cascading a prediction stage to the 32-channels PQMF stage, one can increase the coding gain by further removing redundancies in the signal.

[6] The MPEG AAC filter bank description in Section 3.1.3 refers to the most commonly used configuration, i.e. the low complexity profile configuration. In the main profile configuration, linear prediction is also employed to further extract redundancies in the signal [29], while in the scalable sample rate profile configuration, a gain control and PQF stage is added to the low complexity profile configuration.

3.2.4 PAC

In PAC the steady state condition filter bank is a 1024-point MDCT, while for transients at lower sampling rates a 128-point MDCT is used. A different approach altogether is embraced in order to adapt to transient conditions at high sampling rates (at or above 48 kHz). In this case the filter bank switches between the MDCT-based filter bank and a wavelet-based filter bank with increased time resolution [23]. At a sampling rate of 48 kHz, each band has a width of 23.4 Hz or higher and a time resolution of down to 1.33 ms at high frequencies.

4. PREDICTION AND REDUNDANCY EXTRACTION

Linear prediction can be utilised to remove redundancy in the input time-domain signal with without explicitly mapping the signal in the frequency domain. In the encoder, the predictor subtracts an estimated value from the input signal. This leaves a residual value, which is the error or difference between the predicted and actual input values. The difference signal is then quantised, and sent to the decoder. At the decoder stage, the predicted signal, that was originally removed at the encoder, is regenerated and added back to the residual signal, thereby recreating the equivalent to the original input signal to the predictor. If the input signal is highly correlated a good prediction can be made of the value of each sample. On subtraction of the prediction from the input, small errors are generated which are significantly smaller that the input signal, and hence can be more efficiently quantised. Conversely with a random, noisy signal, only a poor prediction of the input can be made, leading to a large error signal.

For linear predictors, the predictor gain is upper-bounded by the inverse of the sfm, where in this case N is the order of the predictor. Examples of applications in audio coding can be found in [29, 30].

4.1 Predictive coding of spectral coefficients

While most of the time when we think of prediction we think of technologies like ADPCM, etc., prediction of audio signals spectral components lead to significant improvements in temporal shaping of quantization noise in perceptual audio coding [31]. Temporal noise shaping (TNS) was introduced in the development of the MPEG AAC coder. A “difficult” speech excerpt prompted the quest for an improvement of the pre-echo control. While the filter-bank adaptation was helpful, for this particular signal, where a very dynamic variation both in energy level and spectral distribution was present, was not enough to prevent pre-echo effects at very low data rates. By convolving the output of the filter bank and then quantising the frequency coefficients, frequency resolution is traded adaptively in favour of time resolution, allowing for a better control of the spreading in time the quantization noise.

5. CONCLUSIONS

In this paper an overview of the filter-bank technologies used in the time to frequency mapping of audio signals in perceptual coders was presented. Strengths and weaknesses of different approaches for filter-bank design were examined. Redundancy and irrelevancy removal was sought in order to increase the coding efficiency of the perceptual audio coding system at hand. Different approaches in standard audio coding schemes as well as examples of coding schemes in the marketplace were examined. Finally other methods for increasing redundancy extraction, such as prediction were discussed.

ACKNOWLEDGEMENTS

The author wishes to thank all her colleagues for the fruitful discussions that led to the conception this paper. Special thanks to Jim Johnston for a number of inspiring suggestions for this publication and Juergen Herre for proof reading it.

	Filter Bank Type	**Freq. Res. (Hz)**	**Time Res. (ms)**	**Impulse Response LW (ms)**	**Impulse Response SW (ms)**
Layer I	PQMF	750	0.66	10.67	-
Layer II	PQMF	750	0.66	10.67	-
Layer III	PQMF/ MDCT	41.66	4	34.67	18.67
AAC	MDCT	23.44	2.66	42.67	5.33
Twin VQ	MDCT	23.44	2.66	42.67	5.33
AC-3	MDCT	93.75	2.66	10.67	5.33
ATRAC[7]	GC/PQF/ MDCT	43.07	1.45	24.31	3.99
DTS	PQMF/ LPC	750	0.66	10.67	-
PAC	MDCT/ Wavelet	23.44	1.33	42.67	-

Table 1: Comparison Between Different Filter-Bank Implementations at 48 kHz.

[7] ATRAC operates at a sampling rate of 44.1 kHz.

REFERENCES

[1] N. Jayant, P. Noll "Digital Coding of Waveforms: Principles and Applications to Speech and Video," Prentice-Hall, Englewood Cliffs, 1982.

[2] M. A. Krasner, "Digital Encoding of Speech and Audio Signals Based on the Perceptual Requirements of the Auditory System", Technical Report 535, MIT, Lincoln Laboratory, Lexington, 1979.

[3] J. Princen, A. Johnson, A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", Proc. of the ICASSP 1987, pp. 2161-2164.

[4] ISO/IEC 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 3: Audio", 1992.

[5] ATSC, United States Advanced Television Systems Committee Digital Audio Compression (AC-3) Standard, Doc. A/52/10, December 1995.

[6] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, R.M. Heddle, "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc", preprint 3456, 93rd AES-Convention, 1992.

[7] M. Vetterli and J. Kovacevic, "Wavelets and Subband Coding", Prentice Hall, Englewood Cliffs, 1995.

[8] D. Sinha and A. H. Tewfik, "Low Bit-Rate Transparent Audio Compression Using Adapted Wavelets", IEEE Trans. Acoust., Speech, and Signal Processing, 41(12):3463 – 3479, 1993.

[9] J. Princen, J. D. Johnston, "Audio Coding with Signal Adaptive Filterbanks," IEEE Proc. of ICASSP 1995, pp. 3071 - 3074.

[10] J. D. Johnston "Audio Coding with Filter Banks", pages 287-307 in: "Subband and Wavelet Transforms" by A. N. Akansu and M. J. T. Smith (editors), Kluwer Academic Publishers, Norwell 1996.

[11] G. Schuller, "A Low Delay Filter Bank for Audio Coding with Reduced Pre-Echos," preprint 4088, 99th AES-Convention, 1995.

[12] L. D. Fielder, M. Bosi, G. A. Davidson, M. Davis, C. Todd, and S. Vernon" AC-2 and AC-3: Low Complexity Transform-Based Audio Coding," in N. Gielchrist and C. Grewin (ed.), Collected Papers on Digital Audio Bit-Rate Reduction, AES 1996, pp. 54-72.

[13] M. Bosi, K. Brandenburg, S. Quackenbush, K. Akagiri, H. Fuchs, J.Herre, L. Fielder, M.Dietz, Y. Oikawa, G. Davidson, "ISO/IEC MPEG-2 Advanced Audio Coding", JAES, 51, 780 - 792, October 1997.

[14] B. Edler, "Coding of Audio Signals with Overlapping Transform and Adaptive Window Shape" (in German), Frequenz, Vol. 43, No. 9, pp. 252-256, Sept. 1989.

[15] E. Zwicker and H. Fastl, "Psychoacoustics, Facts and Models", Springer 1990.

[16] H. J. Nussbaumer, "Pseudo-QMF Filter Bank", IBM Tech. Disclosure Bull., vol. 24, Nov. 1981, pp. 3081-3087.

[17] J. Princen, A. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, Oct. 1986, pp. 1153-1161.

[18] H. S. Malvar, "Signal Processing with Lapped Transforms," Artech House, Norwood, MA, 1992.

[19] K. Brandenburg and G. Stoll, "The ISO/MPEG-1 Audio Codec: A Generic Standard for Coding of High Quality Digital Audio", JAES, 42, 780 - 792, October 1994.

[20] ISO/IEC 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio", 1994-1997.

[21] ISO/IEC 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 7: Advanced Audio Coding, AAC", 1997.

[22] ISO/IEC 14496-3 "Coding of Audio-Visual Objects, Part 3: Audio", 1998.

[23] D. Sinha, J. D. Johnston, "Audio Compression at Low Data Rates Using Signal Adaptive Switched Filter banks", IEEE Proc. of ICASSP 1996, pp. 1053 - 1056.

[24] P. Noll and D. Pan, "ISO/MPEG Audio Coding" in N. Jayant (ed.), Signal Compression- Coding of Speech, Audio, Text, Image and Video, World Scientific 1997, pp. 69-118.

[25] J. H. Rothweiler, "Polyphase Quadrature Filters - A new Subband Coding Technique", International Conference IEEE ASSP 1983, Boston, pp. 1280-1283.

[26] K. Kostantinides, "Fast Sub-Band Filtering in MPEG Audio Coding ", IEEE Signal Processing Lett., 1994, pp. 26-28.

[27] A. Sugiyama, F. Hazu, M. Iwadare, "Adaptive Transform Coding with an Adaptive Block Size (ATCABS)", Proc. of the ICASSP 1990, Albuquerque, pp. 1093 - 1096.

[28] S. Smyth, P. Smith, M. Smyth, M. Yan and T. Jung, "DTS Coherent Acoustics Delivering High-Quality Multichannel Sound to the Consumer," presented at the 100th AES Convention, Copenhagen, May 1996, pre-print 4293.

[29] H. Fuchs, "Improving MPEG Audio Coding by Backward Adaptive Linear Stereo Prediction", preprint 4086, 99th AES-Convention, 1995.

[30] N. Iwakami, T. Moriya, S. Miki, " High Quality Audio Coding at Less Than 64 kb/s by Using Transform-Domain Interleaved Vector Quantization (Twin-VQ) ", IEEE Proc. of ICASSP 1995, pp. 3095 - 3098.

[31] J. Herre, J. D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", 101st AES Convention, 1996, Preprint 4384.

[32] G. Davidson, " Filter banks Tutorial ", presented at the 105th AES Convention, San Francisco1998.

[33] J. Herre, E. Allamanche, K. Brandenburg, M. Dietz, B. Teichmann, B. Grill, A. Jin, T. Moriya, N. Iwakami, T. Norimatsu, M. Tsushima, T. Ishikawa: "The Integrated Filterbank Based Scalable MPEG-4 Audio Coder", 105th AES Convention, San Francisco 1998, Preprint #4810.