Deepen Sinha, et. Al. “The Perceptual Audio Coder (PAC).”
2000 CRC Press LLC. <http://www.engnetbase.com>.

The Perceptual Audio Coder (PAC)

Deepen Sinha
Bell Laboratories
Lucent Technologies

James D. Johnston
AT&T Research Labs

Sean Dorward
Bell Laboratories
Lucent Technologies

Schuyler R. Quackenbush
AT&T Research Labs

PAC is a perceptual audio coder that is flexible in format and bitrate, and provides high-quality audio compression over a variety of formats from 16 kb/s for a monophonic channel to 1024 kb/s for a 5.1 format with four or six auxiliary audio channels, and provisions for an ancillary (fixed rate) and auxiliary (variable rate) side data channel. In all of its forms it provides efficient compression of high-quality audio. For stereo audio signals, it provides near compact disk (CD) quality at about 56 to 64 kb/s, with transparent coding at bit rates approaching 128 kb/s.
PAC has been tested both internally and externally by various organizations. In the 1993 ISO-MPEG-2 5-channel test, PAC demonstrated the best decoded audio signal quality available from any algorithm at 320 kb/s, far outperforming all algorithms, including the layer II and layer III backward compatible algorithms. PAC is the audio coder in most of the submissions to the U.S. Digital Audio Radio (DAR) standardization project, at bit rates of 160 kb/s or 128 kb/s for two-channel audio compression. It has been adapted by various vendors for the delivery of high quality music over the Internet as well as ISDN links. Over the years PAC has evolved considerably. In this paper we present an overview for the PAC algorithm including some recently introduced features such as the use of a signal adaptive switched filterbank for efficient encoding of non-stationary signals.

42.1 Introduction

With the overwhelming success of the compact disc (CD) in the consumer audio marketplace, the public's notion of "high quality audio" has become synonymous with "compact disc quality". The CD represents stereo audio at a data rate of 1.4112 Mbps (mega bits per second). Despite continued

growth in the capacity of storage and transmission systems, many new audio and multi-media applications require a lower data rate.

In compression of audio material, human perception plays a key role. The reason for this is that source coding, a method used very successfully in speech signal compression, does not work nearly as well for music. Recent U.S. and international audio standards work (HDTV, DAB, MPEG-1, MPEG-2, CCIR) therefore has centered on a class of audio compression algorithms known as *perceptual coders*. Rather than minimizing analytic measures of distortion, such as signal-to-noise ratio, perceptual coders attempt to minimize perceived distortion. Implicit in this approach is the idea that signal fidelity perceived by humans is a better quality measure than "fidelity" computed by traditional distortion measures. Perceptual coders define "compact disc quality" to mean "listener indistinguishable from compact disc audio" rather than "two channel of 16-bit audio sampled at 44.1 kHz".

PAC, the Perceptual Audio Coder [10], employs source coding techniques to remove signal redundancy and perceptual coding techniques to remove signal irrelevancy. Combined, these methods yield a high compression ratio while ensuring maximal quality in the decoded signals. The result is a high quality, high compression ratio coding algorithm for audio signals. PAC provides a 20 Hz to 20 kHz signal bandwidth and codes monophonic, stereophonic, and multichannel audio. Even for the most difficult audio material it achieves approximately ten to one compression while rendering the compression effects inaudible. Significantly higher level of compression, e.g., 22 to 1, is achieved with only a little loss in quality.

The PAC algorithm has its roots in a study done by Johnston [7, 8] on the perceptual entropy (PE) vs. the statistical entropy of music. Exploiting the fact that the perceptual entropy (the entropy of that portion of the music signal above the masking threshold) was less than the statistical entropy resulted in the perceptual transform coder (PXFM) [8, 16]. This algorithm used a 2048 point real FFT with 1/16 overlap, which gave good frequency resolution (for redundancy removal) but had some coding loss due to the window overlap.

The next-generation algorithm was ASPEC [2], which used the modified discrete-cosine transform (MDCT) filterbank [15] instead of the FFT, and a more elaborate bit allocation and buffer control mechanism as a means of generating constant-rate output. The MDCT is a critically sampled filterbank, and so does not suffer the 1/16 overlap loss that the PXFM coder did. In addition, ASPEC employed an adaptive window size of 1024 or 256 to control noise spreading resulting from quantization. However, its frequency resolution was half that of PXFM's resulting in some loss in the coding efficiency (c.f., Section 42.3).

PAC as first proposed in [10] is a third-generation algorithm learning from ASPEC and PXFM-Stereo [9]. In its current form, it uses a long transform window size of 2048 for better redundancy removal together with window switching for noise spreading control. It adds composite stereo coding in a flexible and easily controlled form, and introduces improvements in noiseless compression and threshold calculation methods as well. Additional threshold calculations are made for stereo signals to eliminate the problem of binaural noise unmasking.

PAC supports encoders of varying complexity and quality. Broadly speaking, PAC consists of a core codec augmented by various enhancement. The full capability algorithm is sometimes also referred to as *Enhanced* PAC (or EPAC). EPAC is easily configurable to (de)activate some or all of the enhancements depending on the computational budget. It also provides a built-in scheduling mechanism so that some of the enhancements are automatically turned on or off based on averaged short term computational requirement.

One of the major enhancements in the EPAC codec is geared towards improving the quality at lower bit rates of signals with sharp attacks (e.g., castanets, triangles, drums, etc.). Distortion of attacks is a particularly noticeable artifact at lower bit rates. In EPAC, a signal adaptive switched filterbank which switches between a MDCT and a wavelet transform is employed for analysis and synthesis [18]. Wavelet transform offer natural advantages for the encoding of transient signals and

the switched filterbank scheme allows EPAC to merge this advantage with the advantages of MDCT for stationary audio segments.

Real-time PAC encoder and decoder hardware have been provided to standards bodies, as well as business partners. Software implementation of real time decoder algorithm is available on PCs and workstations, as well as low cost general-purpose DSPs, making it suitable for mass-market applications. The decoder typically consumes only a fraction of the CPU processing time (even on a 486-PC). Sophisticated encoders run on current workstations and RISC-PCs; simpler real-time encoders that provide moderate compression or quality are realizable on correspondingly less inexpensive hardware.

In the remainder of this paper we present a detailed overview of the various elements of PACs, its applications, audio quality, and complexity issues. The organization of the chapter is as follows. In Section 42.2, some of applications of PAC and its performance on formalized audio quality evaluation tests is discussed. In Section 42.3, we begin with a look at the defining blocks of a perceptual coding scheme followed by the description of the PAC structure and its key components (i.e., filterbank, perceptual model, stereo threshold, noise allocation, etc.). In this context we also describe the switched MDCT/wavelet filterbank scheme employed in the EPAC codec. Section 42.4 focuses on the multichannel version of PAC. Discussions on bitstream formation and decoder complexity are presented in Sections 42.5 and 42.6, respectively, followed by concluding remarks in Section 42.7.

42.2 Applications and Test Results

In the most recent test of audio quality [4] PAC was shown to be the best available audio quality choice [4] for audio compression applications concerning 5-channel audio. This test evaluated both backward compatible audio coders (MPEG Layer II, MPEG Layer III) and non-backward compatible coders, including PAC. The results of these tests showed that PAC's performance far exceeded that of the next best coder in the test.

Among the emerging applications of PAC audio compression technology, the Internet offers one of the best opportunities. High quality audio on demand is increasingly popular and promises both to make existing Internet services more compelling as well as open avenues for new services. Since most Internet users connect to the network using as low bandwidth modem (14.4 to 28.8 kb/s) or at best an ISDN link, high quality low bit rate compression is essential to make audio streaming (i.e., real time playback) applications feasible. PAC is particularly suitable for such applications as it offers near CD quality stereo sound at the ISDN rates and the audio quality continues to be reasonably good for bit rates as low as 12 to 16 kb/s. PAC is therefore finding increasing acceptance in the Internet world.

Another application currently in the process of standardization is digital audio radio (DAR). In the U.S. this may have one of several realizations: a terrestrial broadcast in the existing FM band, with the digital audio available as an adjunct to the FM signal and transmitted either coincident with the analog FM, or in an adjacent transmission slot; alternatively, it can be a direct broadcast via satellite (DBS), providing a commercial music service in an entirely new transmission band. In each of the above potential services, AT&T and Lucent Technologies have entered or partnered with other companies or agencies, providing PAC audio compression at a stereo coding rate of 128 to 160 kb/s as the audio compression algorithm proposed for that service.

Some other applications where PAC has been shown to be the best audio compression quality choice is compression of the audio portion of television services, such as high-definition television (HDTV) or advanced television (ATV).

Still other potential applications of PAC that require compression but are broadcast over wired channels or dedicated networks are DAR, HDTV or ATV delivered via cable TV networks, public switched ISDN, or local area networks. In the last case, one might even envision an "entertainment

bus" for the home that broadcasts audio, video, and control information to all rooms in a home.

Another application that entails transmitting information from databases of compressed audio are network-based music servers using LAN or ISDN. This would permit anyone with a networked decoder to have a "virtual music catalog" equal to the size of the music server. Considering only compression, one could envision a "CD on a chip", in which an artist's CD is compressed and stored in a semiconductor ROM and the music is played back by inserting it into a robust, low-power palm-sized music player. Audio compression is also important for read-only applications such as multi-media (audio plus video/stills/text) on CD-ROM or on a PC's hard drive. In each case, video or image data compete with audio for the limited storage available and all signals must be compressed.

Finally, there are applications in which point-to-point transmission requires compression. One is radio station studio to transmitter links, in which the studio and the final transmitter amplifier and antenna may be some distance apart. The on-air audio signal might be compressed and carried to the transmitter via a small number of ISDN B-channels. Another application is the creation of a "virtual studio" for music production. In this case, collaborating artists and studio engineers may each be in different studio, perhaps very far apart, but seamlessly connected via audio compression links running over ISDN.

42.3 Perceptual Coding

PAC, as already mentioned, is a "Perceptual Coder" [6], as opposed to a source modelling coder. For typical examples of source, perceptual, and combined source and perceptual coding, see Figs. 42.1, 42.2, and 42.3. Figure 42.1 shows typical block diagrams of source coders, here exemplified by DPCM, ADPCM, LPC, and transform coding [5]. Figure 42.2 illustrates a basic perceptual coder. Figure 42.3 shows a combined source and perceptual coder.

"Source model" coding describes a method that eliminates redundancies in the source material in the process of reducing the bit rate of the coded signal. A source coder can be either lossless, providing perfect reconstruction of the input signal or lossy. Lossless source coders remove no information from the signal; they remove redundancy in the encoder and restore it in the decoder. Lossy coders remove information from (add noise to) the signal; however, they can maintain a constant compression ratio regardless of the information present in a signal. In practice, most source coders used for audio signals are quite lossy [3].

The particular blocks in source coders, e.g., Fig. 42.1, may vary substantially, as shown in [5], but generally include one or more of the following.

- Explicit source model, for example an LPC model.
- Implicit source model, for example DCPM with a fixed predictor.
- Filterbank, in other words a method of isolating the energy in the signal.
- Transform, which also isolates (or "diagonalizes") the energy in the signal.

All of these methods serve to identify and potentially remove redundancies in the source signal. In addition, some coders may use sophisticated quantizers and information-theoretic compression techniques to efficiently encode the data, and most if not all coders use a bitstream formatter in order to provide data organization. Typical compression methods do not rely on information-theoretic coding alone; explicit source models and filterbanks provide superior source modeling for audio signals.

All perceptual coders are lossy. Rather than exploit mathematical properties of the signal or attempt to understand the producer, perceptual coders model the listener, and attempt to remove irrelevant (undetectable) parts of the signal. In some sense, one could refer to it as a "destination" rather than "source" coder. Typically, a perceptual coder will have a lower SNR than an equivalent rate source coder, but will provide superior perceived quality to the listener.

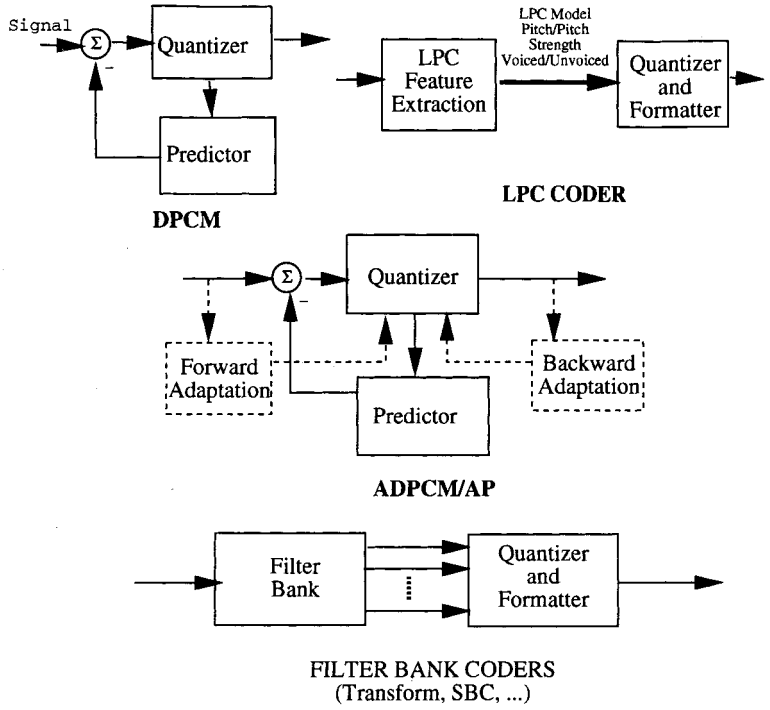

FIGURE 42.1: Block diagrams of selected source-coders.

The perceptual coder shown in Fig. 42.2 has the following functional blocks.

- Filterbank — Converts the input signal into a form suitable for perceptual processing.
- Perceptual model — Determines the irrelevancies in the signal, generating a perceptual threshold.
- Quantization — Applies the perceptual threshold to the output of the filterbank, thereby removing the irrelevancies discovered by the perceptual model.
- Bit stream former — Converts the quantized output and any necessary side information into a form suitable for transmission or storage.

The combined source and perceptual coder shown in Fig. 42.3 has the following functional blocks.

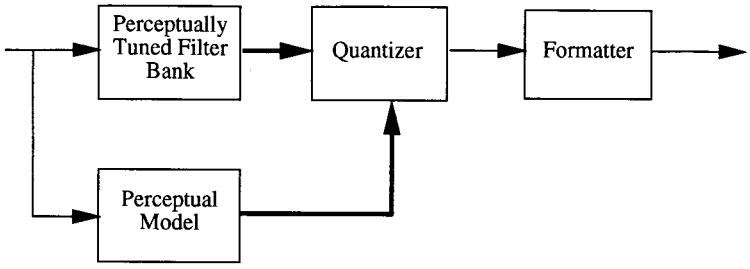

FIGURE 42.2: Block diagrams of a simple perceptual coder.

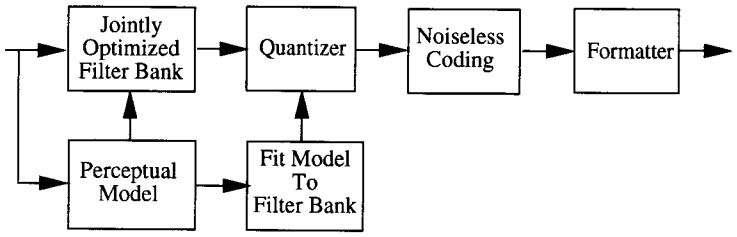

FIGURE 42.3: Block diagrams of an integrated source-perceptual coder.

- Filterbank — Converts the input signal into a form that extracts redundancies and is suitable for perceptual processing.
- Perceptual model — Determines the irrelevancies in the signal, generates a perceptual threshold, and relates the perceptual threshold to the filterbank structure.
- Fitting of perceptual model to filtering domain — Converts the outputs of the perceptual model into a form relevant to the filter bank.
- Quantization – Applies the perceptual threshold to the output of the filterbank, thereby removing the irrelevancies discovered by the perceptual model.
- Information-theoretic compression — Removes redundancy from the output of the quantizer.
- Bit stream former — Converts the compressed output and any necessary side information into a form suitable for transmission or storage.

Most coders referred to as perceptual coders are combined source and perceptual coders. Combining a filterbank with a perceptual model provides not only a means of removing perceptual irrelevancy, but also, by means of the filterbank, provides signal diagonalization, ergo source coding gain. A combined coder may have the same block diagram as a purely perceptual coder; however, the choice of filterbank and quantizer will be different. PAC is a combined coder, removing both irrelevancy and redundancy from audio signals to provide efficient compression.

42.3.1 PAC Structure

Figure 42.4 shows a more detailed block diagram of the monophonic PAC algorithm, and illustrates the flow of data between the algorithmic blocks. There are five basic parts.

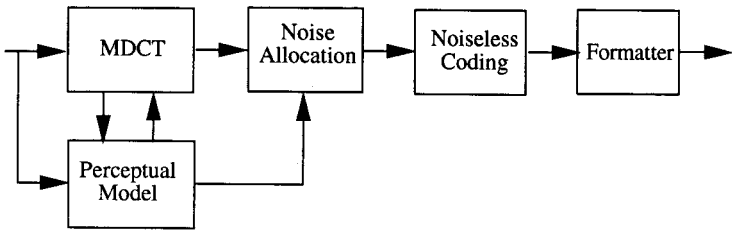

FIGURE 42.4: Block diagram of monophonic PAC encoder.

1. Analysis filterbank — The filterbank converts the time domain audio signal to the short-term frequency domain. Each block is selectably coded by 1024 or 128 uniformly spaced frequency bands, depending on the characteristics of the input signal. PAC's filterbank is used for source coding and cochlear modeling (i.e., perceptual coding).
2. Perceptual model — The perceptual model takes the time domain signal and the output of the filterbank and calculates a frequency domain threshold of masking. A threshold of masking is a frequency dependent calculation of the maximum noise that can be added to the audio material without perceptibly altering it. Threshold values are of the same time and frequency resolution as the filterbank.
3. Noise allocation — Noise is added to the signal in the process of quantizing the filter bank outputs. As mentioned above, the perceptual threshold is expressed as a noise level for each filterbank frequency; quantizers are adjusted such that the perceptual thresholds are met or exceeded in a perceptually gentle fashion. While it is always possible to meet the perceptual threshold in a unlimited rate coder, coding at high compression ratios requires both overcoding (adding less noise to the signal than the perceptual threshold requires) and undercoding (adding more noise to the signal than the perceptual threshold requires). PAC's noise allocation allows for some time buffering, smoothing local peaks and troughs in the bitrate demand.
4. Noiseless compression — Many of the quantized frequency coefficients produced by the noise allocator are zero; the rest have a non-uniform distribution. Information-theoretic methods are employed to provide an efficient representation of the quantized coefficients.
5. Bitstream former — Forms the bitstream, adds any transport layer, and encodes the entire set of information for transmission or storage.

As an example, Fig. 42.5 shows the perceptual threshold and spectrum for a typical (trumpet) signal. The staircase curve is the calculated perceptual threshold, and the varying curve is the short-term spectrum of the trumpet signal. Note that a great deal of the signal is below the perceptual threshold, and therefore redundant. This part of the signal is what we discard in the perceptual coder.

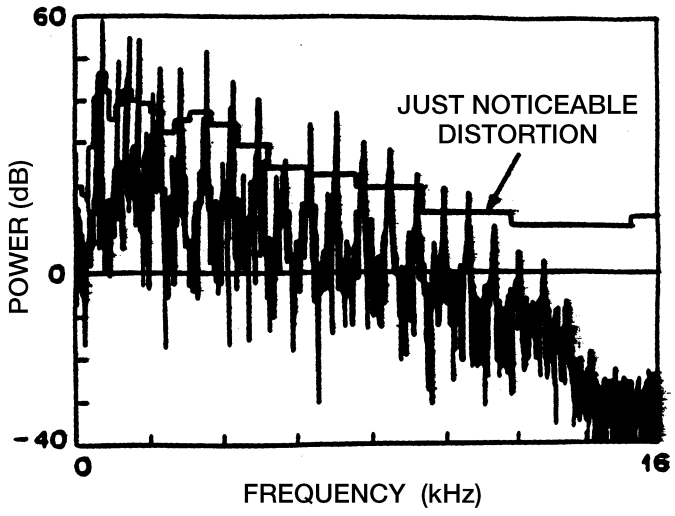

FIGURE 42.5: Example of masking threshold and signal spectrum.

42.3.2 The PAC Filterbank

The filterbank normally used in PAC is referred to as the modified discrete cosine transform (MDCT) [15]. It may be viewed as a modulated, maximally decimated perfect reconstruction filterbank. The subband filters in a MDCT filterbank are linear phase FIR filters with impulse responses twice as long as the number of subbands in the filterbank. Equivalently, MDCT is a lapped orthogonal transform with a 50% overlap between two consecutive transform blocks; i.e., the number of transform coefficients is equal to one half the block length. Various efficient forms of this algorithm are detailed in [11]. Previously, Ferreira [10] has created an alternate form of this filterbank where the decimation is done by dropping the imaginary part of an odd-frequency FFT, yielding and odd-frequency FFT and an MDCT from the same calculations.

In an audio coder it is quite important to appropriately choose the frequency resolution of the filterbank. During the development of the PAC algorithm, a detailed study of the effect of filterbank resolution for a variety of signals was examined. Two important considerations in perceptual coding, i..e, coding gain and non-stationarity within a block, were examined as a function of block length. In general the coding gain increases with the block length indicating a better signal representation for redundancy removal. However, increasing non-stationarity within a block forces the use of more conservative perceptual masking thresholds to ensure the masking of quantization noise at all times. This reduces the realizable or net coding gain. It was found that for a vast majority of music samples the realizable coding gain peaks at the frequency resolution of about 1024 lines or subbands, i.e., a window of 2048 points (this is true for sampling rates in the range of 32 to 48 kHz). PAC therefore employs a 1024 line MDCT as the normal "long" block representation for the audio signal.

In general, some variation in the time frequency resolution of the filterbank is necessary to adapt to the changes in the statistics of the signal. Using a high frequency resolution filterbank to encode a signal segment with a sharp attack leads to significant coding inefficiencies or *pre-echo* conditions. Pre-echos occur when quantization errors are spread over the block by the reconstruction filter. Since pre-masking by an attack in the audio signal lasts for only about 1 msec (or even less for stereo signals), these reconstruction errors are potentially audible as pre-echos unless significant readjustments in the perceptual thresholds are made resulting in coding inefficiencies.

PAC offers two strategies for matching the filterbank resolution to the signal appropriately. A lower computational complexity version is offered in the form of *window switching* approach whereby the MDCT filterbank is switched to a lower 128 line spectral resolution in the presence of attacks. This approach is quite adequate for the encoding of attacks at moderate to higher bit rates (96 kbps or higher for a stereo pair). Another strategy offered as an enhancement in the EPAC codec is the switched MDCT/wavelet filterbank scheme mentioned earlier. The advantages of using such a scheme as well as its functional details are presented below.

42.3.3 The EPAC Filterbank and Structure

The disadvantage of the window switching approach is that the resulting time resolution is uniformly higher for all frequencies. In other words, one is forced to increase the time resolution at the lower frequencies to increase it to the necessary extent at higher frequencies. The inefficient coding of lower frequencies becomes increasingly burdensome at lower bit rates, i.e., 64 kbps and lower. An ideal filterbank for sharp attacks is a non-uniform structure whose subband matches the critical band scale. Moreover, it is desirable that the high frequency filters in the bank be proportionately shorter. This is achieved in EPAC by employing a high spectral resolution *MDCT* for stationary portions of the signal and switching to a non-uniform (tree structured) wavelet filterbank (*WFB*) during non-stationarities.

WFBs are quite attractive for the encoding of attacks [17]. Besides the fact that wavelet representation of such signals is more compact than the representation derived from a high resolution *MDCT*,

wavelet filters have desirable temporal characteristics. In a WFB, the high frequency filters (with a suitable moment condition as discussed below) typically have a compact impulse response. This prevents excessive time spreading of quantization errors during synthesis.

The overview of an encoder based on the switched filterbank idea is illustrated in Fig. 42.6. This structure entails the design of a suitable WFB which is discussed next.

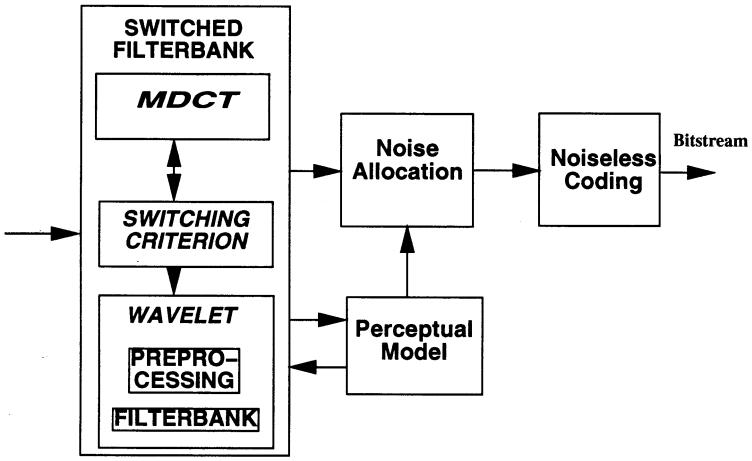

FIGURE 42.6: Block diagram of the switched filterbank audio encoder.

The WFB in EPAC consists of a tree structured wavelet filterbank which approximates the critical band scale. The tree structure has the natural advantage that the effective support (in time) of the subband filters is progressively smaller with increasing center frequency. This is because the critical bands are wider at higher frequency so fewer cascading stages are required in the tree to achieve the desired frequency resolution. Additionally, proper design of the prototype filters used in the tree decomposition ensures (see below) that the high frequency filters in particular are compactly localized in time.

The decomposition tree is based on sets of prototype filterbanks. These provide two or more bands of split and are chosen to provide enough flexibility to design a tree structure that approximates the critical band partition closely. The three filterbanks were designed by optimizing parametrized para-unitary filterbanks using standard optimization tools and an optimization criterion based on weighted stopband energy [20]. In this design, the *moment* condition plays an important role in achieving desirable temporal characteristics for the high frequency filters. An M band para-unitary filterbank with subband filters $\{H_i\}_{i=1}^{i=M}$ is said to satisfy a Pth order moment condition if $H_i(e^{jw})$ for $i = 2, 3, \ldots M$ has a Pth order zero at $\omega = 0$ [20]. For a given support for the filters, K, requiring $P > 1$ in the design yields filters for which the "effective" support decreases with increasing P. In the other words, most of the energy is concentrated in an interval $K' < K$ and K' is smaller for higher P (for a similar stopband error criterion). The improvement in the temporal response of the filters occurs at the cost of an increased transition band in the magnitude response. However, requiring at least a few vanishing moments yields filters with attractive characteristics.

The impulse response of a high frequency *wavelet* filter (in a 4-band split) is illustrated in Fig. 42.7. For comparison, the impulse response of a filter from a modulated filterbank with similar frequency characteristics is also shown. It is obvious that the *wavelet* filter offers superior localization in time.

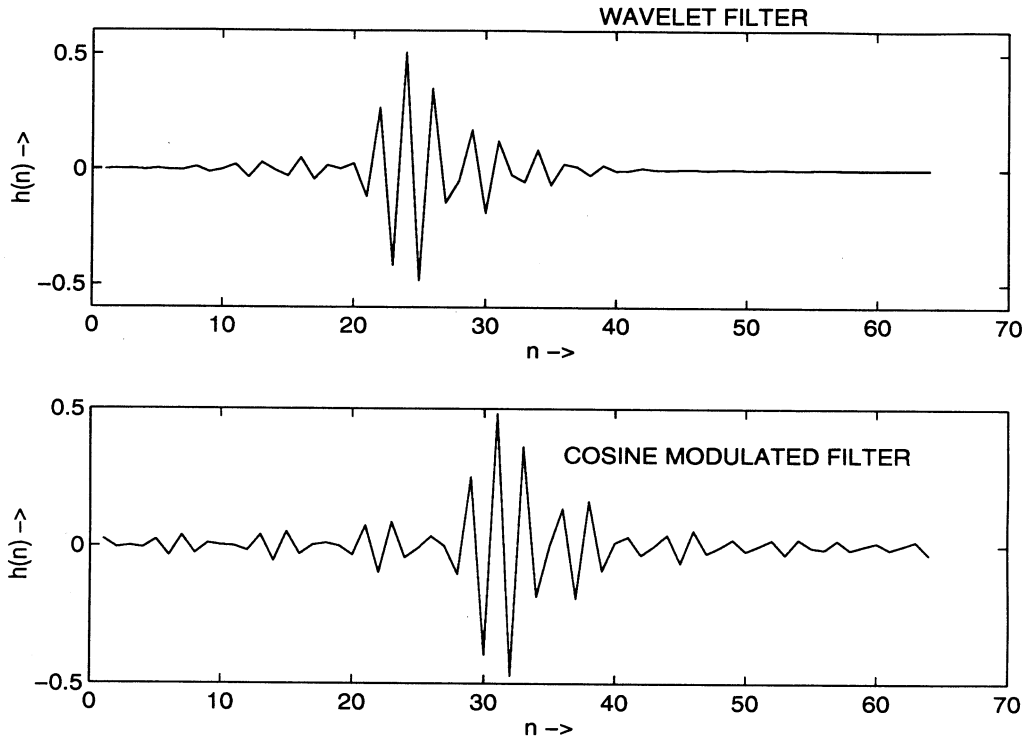

FIGURE 42.7: High frequency wavelet and cosine-modulated filters.

Switching Mechanism

The MDCT is a lapped orthogonal transform. Therefore, switching to a wavelet filterbank requires orthogonalization in the overlap region. While it is straightforward to set up a general orthogonalization problem, the resulting transform matrix is inefficient computationally. The orthogonalization algorithm can be simplified by noting that a MDCT operation over a block of $2 * N$ samples is equivalent to a symmetry operation on the windowed data (i.e., outer $N/2$ samples from either end of the window are folded into the inner $N/2$ samples) followed by an N point orthogonal block transform Q over these N samples. Perfect reconstruction is ensured irrespective of the choice of a particular block orthogonal transform Q. Therefore, Q may be chosen to be a DCT for one block and a wavelet transform matrix for the subsequent or any other block. The problem with this approach is that the symmetry operation extends the wavelet filter (or its translates) in time and also introduces discontinuities in these filters. Thus, it impairs the temporal as well as frequency characteristics of the wavelet filters. In the present encoder, this impairment is mitigated by the following two steps: (1) start and stop windows are employed to switch between $MDCT$ and WFB (this is similar to the window switching scheme in PAC), and (2) the effective overlap between the transition and wavelet windows is reduced by the application of a new family of *smooth* windows [19]. The resulting switching sequence is illustrated in Fig. 42.8.

The next design issue in the switched filterbank scheme is the design of a $N \times N$ orthogonal matrix Q^{WFB} based on the prototype filters and the chosen tree structure. To avoid circular convolutions, we employ transition filters at the edge of the blocks. Given a subband filter, c_k, of length K a total of $K_1 = (K/M) - 1$ transition filters are needed at the two ends of the block. The number at a particular end is determined by the rank of a $K \times (K_1 + 1)$ matrix formed by the translations

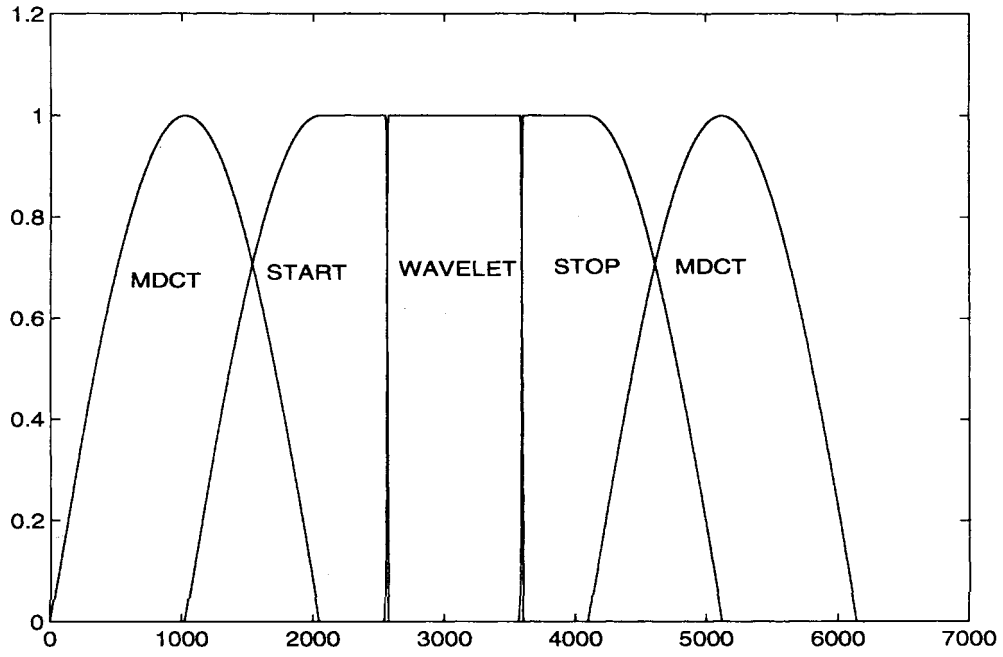

FIGURE 42.8: A filterbank switching sequence.

of c_k. The transition filters are designed through optimization in a subspace constrained by the pre-determined rows of Q^{WFB}.

42.3.4 Perceptual Modeling

Current versions of PAC utilize several perceptual models. Simplest is the monophonic model which calculates an estimated JND in frequency for a single channel. Others add MS (i.e., sum and difference) thresholds and noise-imaging protected thresholds for pairs of channels as well, as "global thresholds" for multiple channels. In this section we discuss the calculation of monophonic thresholds, MS thresholds, and noise-imaging protected thresholds.

Monophonic Perceptual Model

The perceptual model in PAC is similar in method to the model shown as "Psychoacoustic Model II" in the MPEG-1 audio standard annexes [14]. The following steps are used to calculate the masking threshold of a signal.

- Calculate the power spectrum of the signal in 1/3 critical band partitions.
- Calculate the tonal or noiselike nature of the signal in the same partitions, called the tonality measure.
- Calculate the spread of masking energy, based on the tonality measure and the power spectrum.
- Calculate the time domain effects on the masking energy in each partition.
- Relate the masking energy to the filterbank outputs.

Application of Masking to the Filterbank

Since PAC uses the same filterbank for perceptual modeling and source coding, converting masking energy into terms meaningful to the filterbank is straightforward. However, the noise allocator quantizes filterbank coefficients in fixed blocks, called *coder bands*, which differ from the 1/3 critical band partitions used in perceptual modeling. Specifically, 49 coder bands are used for the 1024-line filterbank, and 14 for the 128-line filterbank. Perceptual thresholds are mapped to coder bands by using the minimum threshold that overlaps the band.

In EPAC additional processing is necessary to apply the threshold to the WFB. The thresholds for the quantization of wavelet coefficients are based on an estimate of time-varying *spread* energy in each of the subbands and a tonality measure as estimated above. The spread energy is computed by considering the spread of masking across frequency as well as time. In other words, an inter-frequency as well as a temporal spreading function is employed. The shape of these spreading functions may be derived from the cochlear filters [1]. The temporal spread of masking is frequency dependent and is roughly determined by the (inverse of) bandwidth of the cochlear filter at that frequency. A fixed temporal spreading function for a range of frequencies (wavelet subbands) is employed. The coefficients in a subband are grouped in a *coder* band as above and one threshold value per coderband is used in quantization. The coderband span ranges from 10 msec in the lowest frequency subband to about 2.5 msec in the highest frequency subband.

Stereo Threshold Calculation

Experiments have demonstrated that the monaural perceptual model does not extend trivially to the binaural case. Specifically, even if one signal is masked by both the L (left) and R (right) signals individually, it may not be masked when the L and R signals are presented binaurally. For further details, see the discussion of Binary Masking Level Difference (BLMD) in [12].

In stereo PAC Fig. 42.9, we used a model of BLMD in several ways, all based on the calculation of the M (mono, $L + R$) and S (stereo, $L - R$) thresholds in addition to the independent L and R thresholds. To compute the M and S thresholds, the following steps are added after the computation of the masking energy.

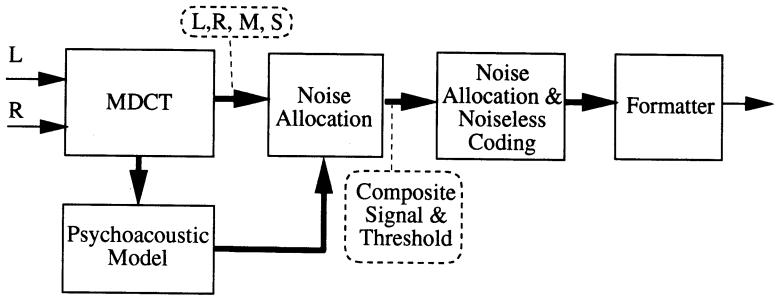

FIGURE 42.9: Stereo PAC block diagram.

- Calculate the spread of masking energy for the other channel, assuming a tonal signal and adding BMLD protection.
- Choose the more restrictive, or smaller, masking energy.

For the L and R thresholds, the following step is added after the computation of the masking energy.

- Calculation of the spread of masking energy for the other channel. If the two masking energies are similar, add BMLD protection to both.

These four thresholds are used for the calculation of quantization, rate, and so on. An example set of spectra and thresholds for a vocal signal are shown in Fig. 42.10. In this figure, compare the threshold values and energy values in the *S* (or "Difference") signal. As is clear, even with the BMLD protection, most of the *S* signal can be coded as zero, resulting in substantial coding gain. Because the signal is more efficiently coded as MS even at low frequencies where the BLMD protection is in effect, that protection can be greatly reduced for the more energetic *M* channel because the noise will image in the same location as the signal, and not create an unmasking condition for the *M* signal, even at low frequencies. This provides increases in both audio quality and compression rate.

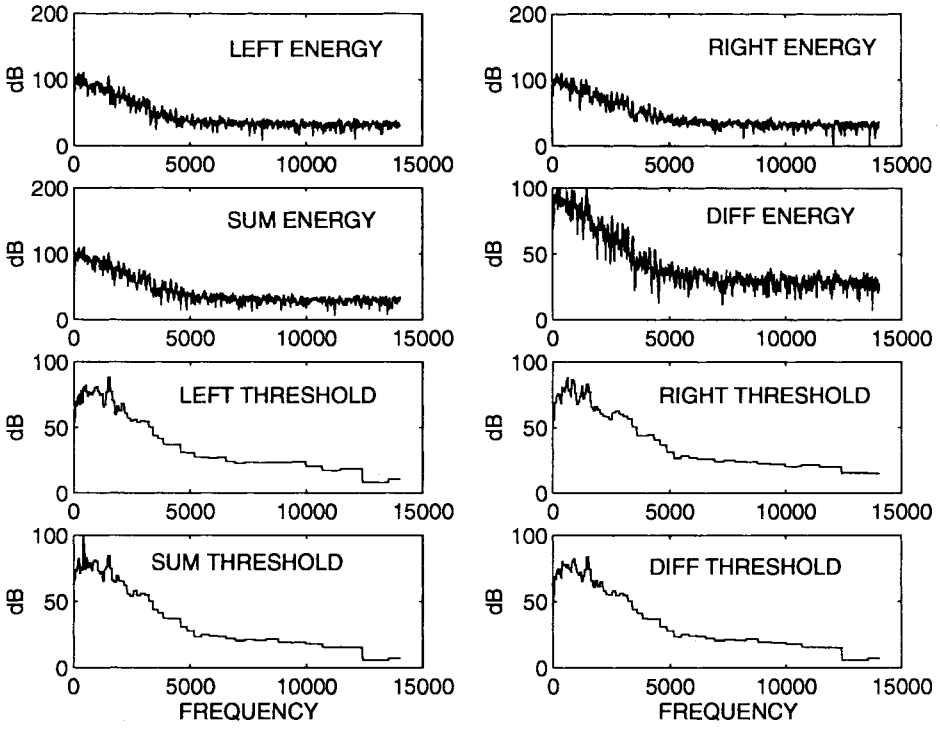

FIGURE 42.10: Examples of stereo PAC thresholds.

42.3.5 MS vs. LR Switching

In PAC, unlike the MPEG Layer III codec [13] MS decisions are made independently for each group of frequencies. For instance, the coder may alternate coding each group as MS or LR, if that proves most efficient. Each of the L, R, M, and S filterbank coefficients are quantized using the appropriate thresholds, and the number of bits required to transmit coefficients is computed. For each group of frequencies, the more efficient of LR or MS is chosen; this information is encoded with a Huffman codebook and transmitted as part of the bitstream.

42.3.6 Noise Allocation

Compression is achieved by quantizing the filter bank outputs into small integers. Each coder band's threshold is mapped onto 1 of 128 exponentially distributed quantizer step sizes, which is used to quantize the filter bank outputs for that coder band.

PAC controls the instantaneous rate of transmission by adjusting the thresholds according to an equal-loudness calculation. Thresholds are adjusted so that the compression ratio is met, plus or minus a small amount to allow for short term irregularities in demand. This noise allocation system is iterative, using a single estimator that represents the absolute loudness of the noise relative to the perceptual threshold. Noise allocation is made across all frequencies for all channels, regardless of stereo coding decision: ergo the bits are allocated in a perceptually effective sense between L, R, M, and S, without regard to any measure of how many bits are assigned to L, R, M, and S.

42.3.7 Noiseless Compression

After the quantizers and quantized coefficients for a block are determined, information-theoretic methods are employed to yield an efficient representation.

Coefficients for each coder band are encoded using one of eight Huffman codebooks. One of the tables encodes only zeros; the rest encode coefficients with increasing absolute value. Each codebook encodes groups of two or four coefficients, with the exception of the zero codebook which encodes all of the coefficients in the band. See Table 42.1 for details. In this table, LAV refers to the largest absolute value in a given codebook, and dimension refers to the number of quantized outputs that are coded together in one codeword. Two codebooks are special, and require further mention. The zero codebook is of indeterminate size, it indicates that all quantized values that the zero codebook applies to are in fact zero, no further information is transmitted about those values. Codebook seven is also a special codebook. It is of size -16:16 by -16:16, but the entry of absolute value 16 is not a data value, it is, rather, an escape indicator. For each escape indicator sent in codebook seven (there can be zero, one, or two per codeword), there is an additional escape word sent immediately after the Huffman codeword. This additional codeword, which is generated by rule, transmits the value of the escaped codeword. This generation by rule is a process that has no bounds; therefore, any quantized value can be transmitted by the use of an escape sequence.

TABLE 42.1 PAC Huffman Codebooks

Codebook	LAV	Dimension
0	0	*
1	1	4
2	1	4
3	2	4
4	4	2
5	7	2
6	12	2
7	ESC	2

Communicating the codebook used for each band constitutes a significant overhead; therefore, similar codebooks are grouped together in *sections*, with only one codebook transmitted and used for encoding each section.

Since the possible quantizers are precomputed, the indices of the quantizers are encoded rather than the quantizer values. Quantizer indices for coder bands which have only zero coefficients are discarded; the rest are differentially encoded, and the differences are Huffman encoded.

42.4 Multichannel PAC

The multichannel perceptual audio coder (MPAC) extends the stereo PAC algorithm to the coding of multiple audio channels. In general, the MPAC algorithm is software configurable to operate in 2, 4, 5, and 5.1 channel mode. In this document we will describe the MPAC algorithm as it is applied to a 5-channel system consisting of the five full bandwidth channels: Left (L), Right (R), Center (C), Left Surround (Ls), and Right Surround (Rs).

The MPAC 5-channel audio coding algorithm is illustrated in Fig. 42.11. Below we describe the various modules, concentrating in particular on the ones that are different from the stereo algorithm.

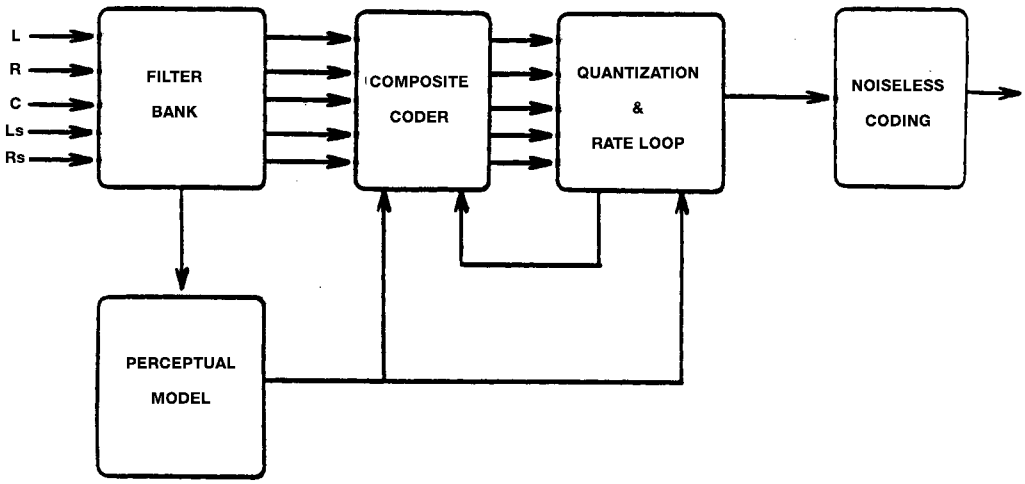

FIGURE 42.11: Block diagram of MPAC.

42.4.1 Filterbank and Psychoacoustic Model

Like the stereo coder, MPAC employs a MDCT filterbank with two possible resolutions, i.e., the usual long block which has 1024 uniformly spaced frequency outputs and a short bank which has 128 uniformly spaced frequency bins. A window switching algorithm, as described above, is used to switch to a short block in the presence of strong non-stationarities in the signal. In the 5-channel setup it desirable to be able to switch the resolution independently for various subsets of channels. For example, one possible, scenario is to apply the window switching algorithm to the front channels (L, R, and C) independently of the surround channels (Ls and Rs). However, this somewhat inhibits the possibilities for composite coding (see below) among the channels. Therefore, one needs to examine the relative gain of independent window switching vs. the gain from a higher level of composite coding. In the present implementation different filterbank resolutions for the front and surround channels are allowed.

The individual masking threshold for the five channels are computed using the PAC psychoacoustic model described above. In addition, the front pair LR and the surround pair Ls/Rs are used to generate two pairs of MS thresholds (c.f., Section “Stereo Threshold Calculation”). The five channels are coded with their individual thresholds excepting in the case where joint stereo coding is being used (either for the front or the surround pair), in which case the appropriate MS thresholds

are used. In addition to the five individual and four stereo thresholds, a joint (or "global") threshold based on all channels is also computed. The computation and role of the global threshold will be discussed later in this section.

42.4.2 The Composite Coding Methods

The MPAC algorithm extends the MS coding of the stereo algorithm to a more elaborate composite coding scheme. Like the MS coding algorithm, the MPAC algorithm uses adaptive composite coding in both time and frequency: the composite coding mode is chosen separately for each of the coder bands at every analysis instance. This selection is based on a "perceptual entropy" criterion and attempts to minimize the bit rate requirement as well as exercise some control over noise localization. The coding scheme uses two complementary sets of inter-channel combinations as described below:

- MS coding for the front and surround pair
- Inter-channel prediction

MS coding is a basis transformation operation and is therefore performed with the uncoded samples of the corresponding pair of channels. The resulting M or S channel is then coded using its own threshold (which is computed separately from the individual channel threshold). Inter-channel prediction, on the other hand, is performed using the quantized samples of the predicting channel.This is done to prevent the propagation of quantization errors (or "cross-talk"). The predicted value for each channel is subtracted from the channel samples and the resulting difference is encoded using the original channel threshold. It may be noted that the two sets of channel combinations are nested so that either, both, or none may be employed for a particular coder band. The coder currently employs the following possibilities for inter-channel prediction.

For the Front Channels (L, R & C): Front L and R channels are coded as LR or MS. In addition, one of the following two possibilities for inter-channel prediction may be used.

1. Center predicts LR (or M if MS coding mode is on).
2. Front M channel predicts the center.

For the Surround Channels (Ls and Rs): Ls and Rs channels are coded as Ls/Rs or Ms/Ss (where Ms and Ss are, respectively, the surround M and surround S). In addition, one or both of following two modes of interchannel prediction may be employed:

1. Front L, R, M channels predict Ls/Rs or Ms.
2. Center channel predicts Ls/Rs or Ms.

In the present implementation, the predictor coefficients in all of the above inter-channel prediction equations are all fixed to either zero or one.

Note that the possibility of completely independent coding is implicit in the above description, i.e., the possibility of turning off any possible prediction is always included. Furthermore, any of these conditions may be independently used in any of the 49 coder bands (long filter band length) or in the 14 coder bands (short filter band length), for each block of filterbank output. Also note that for the short filterbank where the outputs are grouped into 8 groups of 128 (each group of 128 has 14 bands), each of these 8 groups has independently calculated composite coding.

The decisions for composite coding are based primarily on the "perceptual entropy" criterion; i.e., the composite coding mode is chosen to minimize the bit requirement for the perceptual coding of the filterbank outputs from the five channels. The decision for MS coding (for the front and surround pair) is also governed in part by noise localization considerations. As a consequence, the MPAC coding algorithm ensures that signal and noise images are localized at the same place in the

front and rear planes. The advantage of this coding scheme is that the quantization noise usually remains masked not only in a listening room environment but also during headphone reproduction of a stereo downmix of the five coded channels (i.e., when two downmixed channels of the form $Lc = L + \alpha C + \beta Ls$, and $Rc = R + \alpha C + \beta Rs$ are produced and fed to a headphone).

The method used for composite coding is still in the experimental phase and subject to refinements/modifications in future.

42.4.3 Use of a Global Masking Threshold

In addition to the five individual thresholds and the four MS thresholds, the MPAC coder also makes use of a global threshold to take advantage of masking across the various channels. This is done when the bit demand is consistently high so that the bit reservoir is close to depletion. The global threshold is taken to be the maximum of five individual thresholds minus a "safety margin". This global threshold is phased in gradually when the bit reservoir is really low (e.g., less than 20%) and in that case it is used as a lower limit for the individual thresholds.

The reason that global threshold is useful is because results in [12] indicate that if the listener is more than a "critical distance" away from the speakers, then the spectrum at either of listener's ear may be well approximated by the sum of power spectrums due to individual speakers.

The computation of a global threshold also involves a safety margin. This safety margin is frequency dependent and is larger for the lower frequencies and smaller for higher frequencies. The safety margin changes with the bit reservoir state.

42.5 Bitstream Formatter

PAC is a block processing algorithm; each block corresponds to 1024 input samples from each channel, regardless of the number of channels. The encoded filter bank outputs, codebook sections, quantizers, and channel combination information for one 1024-sample chunk or eight 128-sample chunks are packed into one *frame*.

Depending on the application, various extra information is added to first frame or to every frame. When storing information on a reliable media, such as a hard disk, one header indicating version, sample rate, number of channels, and encoded rate is placed at the beginning of the compressed music. For extremely unreliable transmission channels, like DAR, a header is added to each frame. This header contains synchronization, error recovery, sample rate, number of channels, an the transmission bit rate.

42.6 Decoder Complexity

The PAC decoder is of approximately equal complexity to other decoders currently known in the art. Its memory requirements are approximately

- 1100 words each for MDCT and WFB workspace
- 512 words per channel for MDCT memory
- (optional) 1024 words per channel for error mitigation
- 1024 samples per channel for output buffer
- 12 Kbytes ROM for codebooks

The calculation requirements for the PAC decoder are slightly more than doing a 512-point complex FFT per 1024 samples per channel. On an Intel 486 based platform, the decoder executes in real time using up approximately 30 to 40.

42.7 Conclusions

PAC has been tested both internally and externally by various organizations. In the 1993 ISO-MPEG-2 5-channel test, PAC demonstrated the best decoded audio signal quality available from any algorithm at 320 kb/s, far outperforming all algorithms, including the backward compatible algorithms. PAC is the audio coder in three of the submissions to the U.S. DAR project, at bit rates of 160 kb/s or 128 kb/s for two-channel audio compression.

PAC presents innovations in the stereo switching algorithm, the psychoacoustic model, filterbank, the noise-allocation method, and the noiseless compression technique. The combination provides either better quality or lower bit rates than techniques currently on the market.

In summary, PAC offers a single encoding solution that efficiently codes signals from AM bandwidth (5 to 10 kHz) to full CD bandwidth, over dynamic ranges that match the best available analog to digital convertors, from one monophonic channel to a maximum of 16 front, 7 back, 7 auxiliary, and at least 1 effects channel. It operates from 16 kb/s up to a maximum of more than 1000 kb/s for the multiple-channel case. It is currently implemented in 2-channel hardware encoder and decoder, and 5-channel software encoder and hardware decoder. Versions of the bitstream that include an explicit transport layer provide very good robustness in the face of burst-error channels, and methods of mitigating the effects of lost audio data.

In the future, we will continue to improve PAC. Some specific improvements that are already in motion are the improvement of the psychoacoustic threshold for unusual signals, reduction of the overhead in the bitstream at low bit rates, improvements of the filterbanks for higher coding efficiency, and the application of vector quantization techniques.

References

[1] Allen, J.B., Ed., *The ASA Edition of Speech Hearing in Communication*, Acoustical Society of America, Woodbury, New York, 1995.

[2] Brandenburg, K. and Johnston, J.D., ASPEC: Adaptive spectral entropy coding of high quality music signals, *AES 90th Convention*, 1991.

[3] G722. *The G722 CCITT Standard for Audio Transmission.*

[4] ISO-II, *Report on the MPEG/Audio Multichannel Formal Subjective Listening Tests*, ISO/MPEG document MPEG94/063. ISO/MPEG-II Audio Committee, 1994.

[5] Jayant, N.S. and Noll, P., *Digital Coding of Waveforms, Principles and Applications to Speech and Video*, Prentice-Hall, Englewoods Cliffs, NJ, 1984.

[6] Jayant, N.S., Johnston, J., and Safranek, R.J., Signal compression based on models of human perception, *Proc. IEEE*, 81(10), 1993.

[7] Johnston, J.D., Estimation of perceptual entropy using noise masking criteria, *ICASSP-88 Conf. Record*, 1988.

[8] Johnston, J.D., Transform coding of audio signals using perceptual noise criteria, *IEEE J. Sepected Areas in Commun.*, Feb. 1988.

[9] Johnston, J.D., Perceptual coding of wideband stereo signals, *ICASSP-89 Conf. Record*, 1989.

[10] Johnston, J.D. and Ferreira, A. J., Sum-difference stereo transform coding, *ICASSP-92 Conf. Record*, II-569 – II-572, 1992.

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

[12] Moore, B.C.J., *An Introduction to the Psychology of Hearing*, Academic Press, New York, 1989.

[13] MPEG, *ISO-MPEG-1/Audio Standard.*

[14] Mussmann, H.G., The ISO audio coding standard, *Proc. IEEE-Globecom.*, 1990.

[15] Princen, J.P. and Bradlen, A.B., Analysis/synthesis filter bank design based on time domain aliasing cancellation, *IEEE Trans. ASSP*, 34(5), 1986.

[16] Quackenbush, S.R., Ordentlich, E., and Snyder, J.H., Hardware implementation of a 128-kbps monophonic audio coder, in *1989 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1989.
[17] Sinha, D. and Tewfik, A. H., Low bit rate transparent audio compression using adapted wavelets, *IEEE Trans. Signal Processing*, 41(12), 3463-3479, Dec. 1993.
[18] Sinha, D. and Johnston, J.D., Audio compression at low bit rates using a signal adaptive switched filterbank, in *Proc. IEEE Intl. Conf. on Acoust. Speech and Signal Proc.*, II-1053, May 1996.
[19] Sinha, D., *A New Family of Smooth Windows*, in preparation.
[20] Vaidyanathan, P.P., Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial, *Proc. IEEE*, 78(1), 56-92, Jan. 1990.