AUDIO CODING STANDARDS

Chi-Min Liu

Department of Computer Science and Information Engineering

National Chiao Tung University, Taiwan

Wen-Whei Chang

Department of Communication Engineering

National Chiao Tung University, Taiwan

1. INTRODUCTION 1

2. ISO/MPEG AUDIO CODING STANDARDS 3

3. OTHER AUDIO CODING STANDARDS 10

4. ARCHITECTURAL OVERVIEW 10

5. CONCLUSIONS 18

1. INTRODUCTION

With the introduction of compact disc (CD) in 1982, the digital audio media has quickly replaced the analog audio media. However, a significant amount of uncompressed data (1.41 million bits per second) required for the digital audio has led to a large transmission and storage burden. The advances of audio coding techniques and the resultant standards have greatly eased the burden. Ten years ago, nearly nobody believed that 90% of the audio data

could be deleted without affecting audio fidelity. Nowadays, the fantasy becomes reality and the on-going coding technologies are inspiring new dreams. This chapter reviews some international and commercial product audio coding standards, including ISO/MPEG family [ISO, 1992][ISO, 1994][ISO, 1997][ISO, 1999], the Philips PASC [Lokhoff, 1992], the Sony ATRAC [Tsutsui, 1992], and the Dolby AC-3 [Todd, 1994] algorithm.

"Audio Coding Standards," A chapter for the book "Handbook of Multimedia Communication," to appear in a book by Academic Press, 2000

2. ISO/MPEG AUDIO CODING STANDARDS

The Moving Pictures Experts Group (MPEG) within the International Organization for Standardization (ISO) has developed a series of audio coding standards for storage and transmission of various digital media. The ISO standard specifies a syntax for only the coded bit-streams and the decoding process; sufficient flexibility is allowed for encoder implementation. The MPEG first-phase (MPEG-1) audio coder operates in single-channel or two-channel stereo mode at sampling rates of 32, 44.1, and 48 kHz. In the second phase of development, particular emphasis is placed on the multichannel audio support and on an extension of the MPEG-1 to lower sampling rates and lower bit rates. MPEG-2 audio consists of mainly two coding standards: MPEG-2 BC [ISO, 1994] and MPEG-2 AAC [ISO, 1997]. Unlike MPEG-2 BC, which is constrained by its backward compatibility (BC) with MPEG-1 format, MPEG-2 AAC (Advanced Audio Coding) is unconstrained and can therefore provide better coding efficiency. The most recent development is the adoption of MPEG-4 [ISO, 1999] for very-low-bit-rate channels, such as those found in Internet and mobile applications. Table 1 lists the configuration used in MPEG audio coding standards.

Standards	Audio sampling rate (kHz)	Compressed bit-rate (kbits/sec)	Channels	Standard Approved
MPEG-1 Layer I	32, 44.1, 48	32 – 448	1-2 channels	1992
MPEG-1 Layer II	32, 44.1, 48	32 – 384	1-2 channels	1992
MPEG-1 Layer III	32, 44.1, 48	32 – 320	1-2 channels	1993
MPEG-2 Layer I	32, 44.1, 48	32 – 448 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	32 – 256 for two BC channels		
MPEG-2 Layer II	32, 44.1, 48	32 – 384 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	8 – 160 for two BC channels		
MPEG-2 Layer III	32, 44.1, 48	32 – 384 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	8 – 160 for two BC channels		
MPEG-2 AAC	8, 11.025, 12, 16, 22.05, 24, 32, 44.1, 48, 64, 88.2, 96	Indicated by a 23-bit unsigned integer	1-48 channels	1997
MPEG-4 T/F coding	8, 11.025, 12, 16, 22.05, 24, 32, 44.1, 48, 64, 88.2, 96	Indicated by a 23-bit unsigned integer	1-48 channels	1999

Table 1 Comparison of ISO/MPEG audio coding standards

2.1 MPEG-1

The MPEG-1 standard consists of three layers of audio coding schemes with increasing complexity and subjective performance. These layers were developed in collaboration mainly with AT&T, CCETT, FhG/University of Erlangen, Philips, IRT, and Thomson Consumer Electronics. MPEG-1 operates in one of four possible modes: mono, stereo, dual channel, and

joint stereo. With a joint stereo mode, further compression can be realized through some intelligent exploitation of either the correlation between the left and right channels or the irrelevancy of the phase difference between them.

2.1.1 MPEG-1 Layers I and II

Block diagrams of Layer I and Layer II encoders are given in Fig. 1. An analysis filterbank splits the input signal with sampling rate F_s by dividing it into 32 equally spaced subband signals with sampling rate $F_s/32$. In each of the 32 subbands, 12 consecutive samples are assembled into blocks with the equivalent of 384 input samples. All of the samples within one block are normalized by a scale factor so that they all have absolute values less than one. The choice of a scale factor is done by first finding the sample with the maximum absolute value, and then comparing it to a scale factor table of 63 allowable values. After normalization, samples are quantized and coded under the control of a psychoacoustic model. Detailed psychoacoustic analysis is performed through the use of a 512 (Layer I) or 1024 (Layer II) point FFT in parallel with the subband decomposition. The bit-allocation unit determines the quantizer resolution according to the targeted bit rate and the perceptual information derived from the psychoacoustic model. Layer II introduces further compression with respect to Layer I through three modifications. First, the overall information is reduced by removing redundancy and irrelevance between the scale factors of three adjacent 12-sample blocks. Second, a quantization table with improved precision is provided. Third, the psychoacoustic analysis benefits from better frequency resolution because of the increased FFT size.

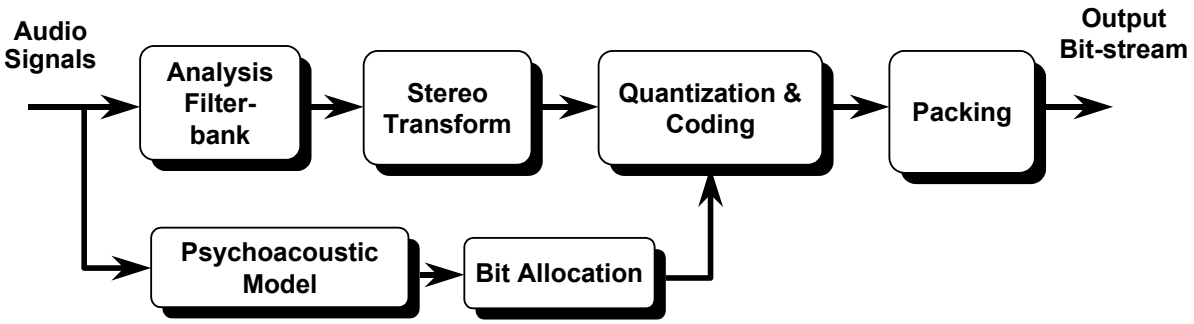

Fig. 1. MPEG-1 Layer I or II audio encoder.

2.1.2 MPEG-1 Layer III

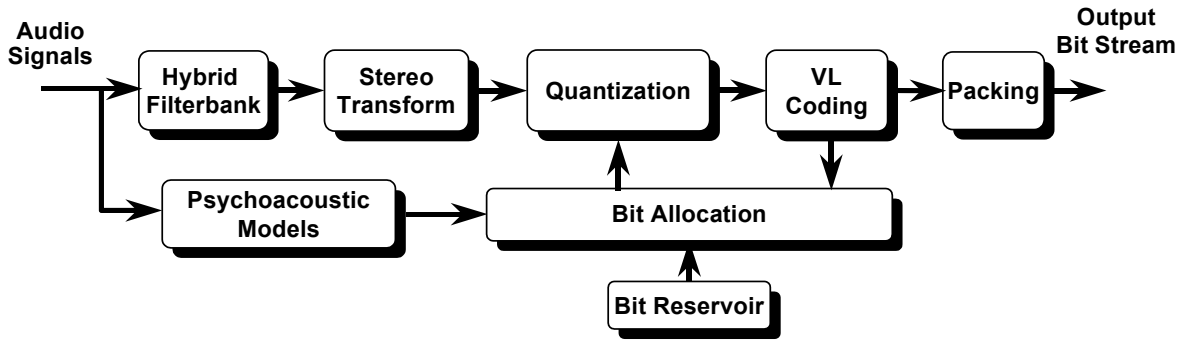

Fig. 2. MPEG-1 Layer III audio encoder.

The MPEG-1 Layer III audio coder introduces many new features, in particular a hybrid filterbank which is a cascade of two filterbanks. For notational convenience, the first filterbank is labeled as the Layer III 1st hybrid level and the second as the Layer III 2nd hybrid level . A block diagram of the Layer III encoder is given in Fig. 2. Although its 1st level is based on the same filterbank found in the other Layers, Layer III provides a higher frequency resolution by subdividing each of the 32 subbands with an 18-point modified discrete cosine transform (**MDCT**). Furthermore, the transform block size adapts to signal characteristics to ensure dynamic tradeoffs between time and frequency resolution. It also employs nonuniform quantization in conjunction with variable length coding for further savings in bit rates. One special feature of Layer III is the bit reservoir; it provides the vehicle to better fit the encoder's time-varying demand on code bits. The encoder can donate bits to a reservoir when it needs less than the average number of bits to code the samples in a frame. But in case the audio signals are hard to compress, the encoder can borrow bits from the reservoir to improve the fidelity.

2.2 MPEG-2

MPEG-2 differs from MPEG-1 in that it supports up to 5.1 channels, including five full-bandwidth channels of the 3/2 stereo, plus an optional low-frequency enhancement channel. This multichannel extension leads to an improved realism of auditory ambience not only for audio-only applications, but also for high-definition television (HDTV) and digital versatile disc (DVD). In addition, initial sampling rates can be extended downward to include 16, 22.05, and 24 kHz. Two coding standards within MPEG-2 are defined: the BC (Backward Compatible) standard preserves the backward compatibility with MPEG-1, and the AAC (Advanced Audio Coding) standard does not.

2.2.1 MPEG-2 BC

Regarding syntax and semantics, the differences between MPEG-1 and MPEG-2 BC are minor, except in the latter case for the new definition of a sampling frequency field, a bit rate index field, and a psychoacoustic model used in bit allocation tables. In addition, parameters of MPEG-2 BC have to be changed accordingly. With the extension of lower sampling rates, it is possible to compress two-channel audio signals to bit rates less than 64 kb/s with good quality. Backward compatibility implies that existing MPEG-1 audio decoders can deliver two main channels of the MPEG-2 BC coded bitstream. This is achieved by coding the left and right channels as MPEG-1, while the remaining channels are coded as ancillary data in the MPEG-1 bitstream.

2.2.2 MPEG-2 AAC

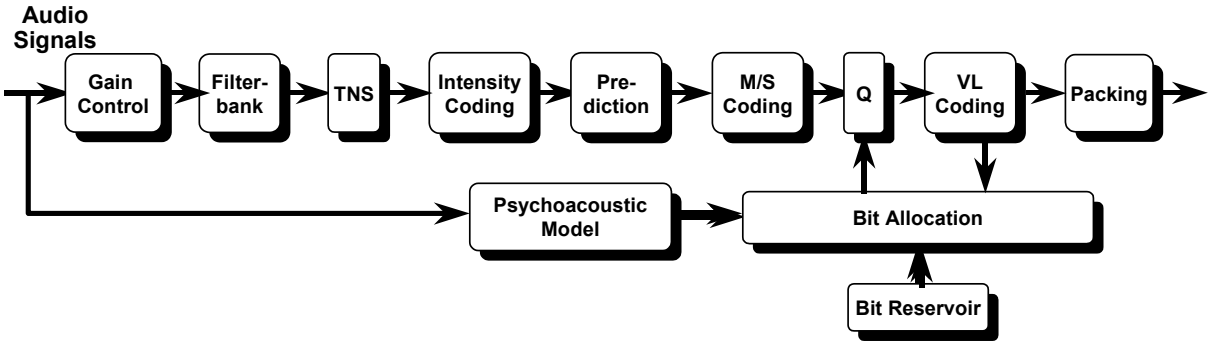

Fig. 3. MPEG-2 AAC audio encoder

MPEG-2 ACC provides the highest quality for applications where backward compatibility with MPEG-1 is not a constraint. While MPEG-2 BC provides good audio quality at data rates of 640-896 kb/s for five full-bandwidth channels, MPEG-2 AAC provides very good quality at less than half of that data rate. Block diagram of an AAC encoder is given in Fig. 3. The gain control tool splits the input signal into four equally spaced frequency bands, which are then flexibly encoded to fit into a variety of sampling rates. The pre-echo effect can also be alleviated through the use of the gain control tool. The filterbank transforms the signals from the time domain to the frequency domain. The temporal noise shaping (TNS) tool helps to control the temporal shape of the quantization noise. **Intensity coding** and the coupling reduce perceptually irrelevant information by combining multiple channels in high-frequency regions into a single channel. The prediction tool further removes the redundancies between adjacent frames. M/S coding removes stereo redundancy based on coding the sum and difference signal instead of the left and right channels. Other units, including quantization, variable length coding, psychoacoustic model, and bit allocation, are similar to those used in MPEG Layer III.

MPEG-2 AAC offers flexibility for different quality-complexity tradeoffs by defining three profiles: the main profile, the low-complexity profile, and the sampling rate scalable

(SRS) profile. Each profile builds on some combinations of different tools as listed in Table 2. The main profile yields the highest coding efficiency by incorporating all the tools with the exception of the gain control tool. The low complexity profile is used for applications where memory and computing power are constrained. The SRS profile offers a scalable complexity by allowing partial decoding of a reduced audio bandwidth.

2.3 MPEG-4

The MPEG-4 standard, which was finalized in 1999, integrates the whole range of audio from high-fidelity speech coding and audio coding down to synthetic speech and synth audio. The MPEG-2 ACC tool set within the MPEG-4 standard supports the compression of natural audio at bit rates ranging from 2 up to 64 kb/s. The MPEG-4 standard defines three types of coders: parametric coding, code-excited linear predictive (CELP) coding, and time/frequency

Tools	**Main**	**Low Complexity**	**SRS**
Variable-Length Decoding	✓	✓	✓
Inverse Quantizer	✓	✓	✓
M/S	✓	✓	✓
Prediction	✓	×	×
Intensity/Coupling	✓	×	×
TNS	✓	Limited	Limited
Filterbank	✓	✓	✓
Gain Control	×	×	✓

Table 2 Coding tools used in MPEG-2 AAC

(T/F) coding. For speech signals sampled at 8 kHz, parametric coding is used to achieve targeted bit rates between about 2 and 6 kb/s. For audio signals sampled at 8 and 16 kHz, CELP coding offers good quality at medium bit rates between about 6 and 24 kb/s.

T/F coding is typically applied to the bit rates starting at about 16 kb/s for audio signals

with bandwidths above 8 kHz. T/F coding is developed based on the coding tools used in MPEG-2 AAC with some add-ons. One is referred to as the twin-VQ (vector quantization), which makes combined use of an interleaved VQ and LPC (Linear Predictive Coding) spectral estimation. In addition, the introduction of bit-sliced arithmetic coding (BSAC) offers noiseless transcoding of an AAC stream into a fine granule scalable stream between 16 and 64 kb/s per channel. BSAC enables the decoder to stop anywhere between 16 kb/s and the bit rate arranged in 1-kb/s steps.

3. OTHER AUDIO CODING STANDARDS

The audio data on a compact disc is typically sampled at 44.1 kHz that requires an uncompressed data rate of 1.41 Mb/s for stereo sound with 16 bit pulse code modulation (PCM). Lower bit rates than those given by 16-bit PCM format are mandatory in order to support a circuit realization that is compact and has low-power consumption, two key enabling factors of equipment portability for the user. The digital compact cassette (DCC) developed by Philips is one of the first commercially available forms of perceptual coded media. To offer backward compatibility for playback of analog compact cassettes, DCC's combination of tape speed and symbol spacing yields a raw data rate of only 768 kb/s, half of that is used for error correcting redundancy. Another example is Sony's MiniDisc (MD) that allows us to store a full CD's worth of music on a disc only half the diameter. DCC and MD systems make use of perceptual coding techniques to achieve the necessary compression ratios of 4:1 and 5:1, respectively. Dolby AC-3 is currently the audio coding standard for the United States Grand Alliance HDTV system and has been widely adopted for DVD films. Dolby AC-3 can reproduce various playback configurations from one channel up to 5.1 channels: left, right, center, left-surrounding, right-surrounding, and low-frequency enhancement channels.

3.1 Philips PASC

Philips' DCC incorporates the PASC (Precision Adaptive Subband Coding) algorithm that is capable of compressing two-channel stereo audio to 384 kb/s with near CD quality [Lokhoff,1992]. PASC can be considered as a simplified version of ISO/MPEG-1 Layer I; it does not require a side-chain FFT analysis for the estimation of masking threshold. The PASC encoder creates 32 subband representations of the audio signal, which are then quantized coded according to the bit allocation derived from a psychoacoustic model. The first generation PASC encoder performs a very simple psychoacoustic analysis based on the outputs of the filterbank. By measuring the average power level of 12 samples, the masking levels of that particular subband and all the adjacent subbands can be estimated with the help of an empirically derived *32x32* matrix, which is described in the DCC standard. The algorithm assumes the 32 frequencies of this matrix are positioned on the edges of the subband spectra, the most conservative approach.

Every block of 12 samples is converted to a floating-point notation; the mantissa determines resolution and the exponent controls dynamic range. As in MPEG-1 Layer I, the scale factor is determined and coded as a 6-bit exponent; it is valid for 12 samples within a block. The algorithm assigns each sample a mantissa with a variable length of 2 to 15 bits, depending on the ratio of the maximum signal to the masking threshold, plus an additional 4 bits for allocation information detailing the length of a mantissa.

3.2 Sony ATRAC

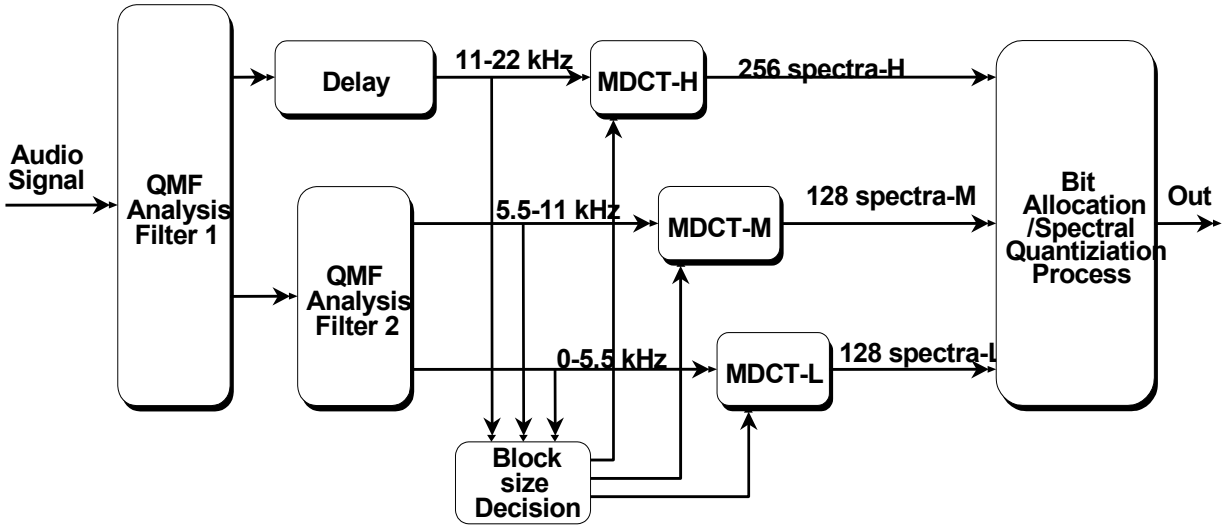

Fig. 4. ATRAC audio encoder.

The ATRAC (Adaptive TRansform Acoustic Coding) algorithm was developed by Sony to support 74 minutes of recording and playing time on a 64-mm MiniDisc [Tsutsui, 1992]. It supports coding of 44.1 kHz two-channel audio at a rate of 256 kb/s. The key to ATRAC's efficiency is that psychoacoustic principles are applied to both the bit allocation and the time-frequency mapping. The encoder (Fig. 4) begins with two stages of quadrature mirror filters (QMFs) to divide the audio signal into three subbands which cover the ranges of 0-5.5 kHz, 5.5-11.0 kHz, and 11.0-22.0 kHz. These subbands are then transformed from the time domain to the frequency domain using the modified discrete cosine transform (MDCT). In addition, the transform block size adapts to signal characteristics to ensure dynamic tradeoffs between time and frequency resolution. The default transform block size is 11.6 ms, but in case of predicted pre-echoes the block size is switched to 1.45 ms in the high-frequency band and to 2.9 ms in the low- and mid-frequency bands. Following the time-frequency analysis, transform coefficients are grouped nonuniformly into 52 block floating units (BFUs) in accordance with the ear's **critical band** partitions. Transform coefficients are quantized using

two parameters: word length and scale factor. The scale factor defines the full-scale range of the quantization and the word length defines the resolution within that scale. Each of the 52 BFUs has the same word length and scale factor, reflecting the psychoacoustic similarity within each critical band.

The bit allocation algorithm determines the word length with the aim of keeping the quantization noise below the masking threshold. One suggested algorithm makes combined use of fixed and variable bits. The algorithm assigns each BFU variable bits according to the logarithm of the transform coefficients. Fixed bits are mainly allocated to low-frequency BFU regions; this reflects the ear's decreasing sensitivity toward higher frequencies. The total bit allocation $b_{tot}(k)$ is the weighted sum of the fixed bit $b_{fix}(k)$ and the variable bit $b_{var}(k)$. Thus, for each BFU k, $b_{tot}(k)=T\ b_{var}(k)+ (1-T)\ b_{fix}(k)$. The weight T describes the tonality of the signal, taking a value close to 1 for pure tones; and a value close to 0 for white noise. To ensure a fixed data rate, an offset b_{off} is subtracted from $b_{tot}(k)$ to yield the final bit allocation $b(k)=$ integer$[b_{tot}(k)- b_{off}]$. As a result, the ATRAC encoder output contains MDCT block size mode, word length and scale factor for each BFU, and quantized spectral coefficients.

3.3 Dolby AC-3

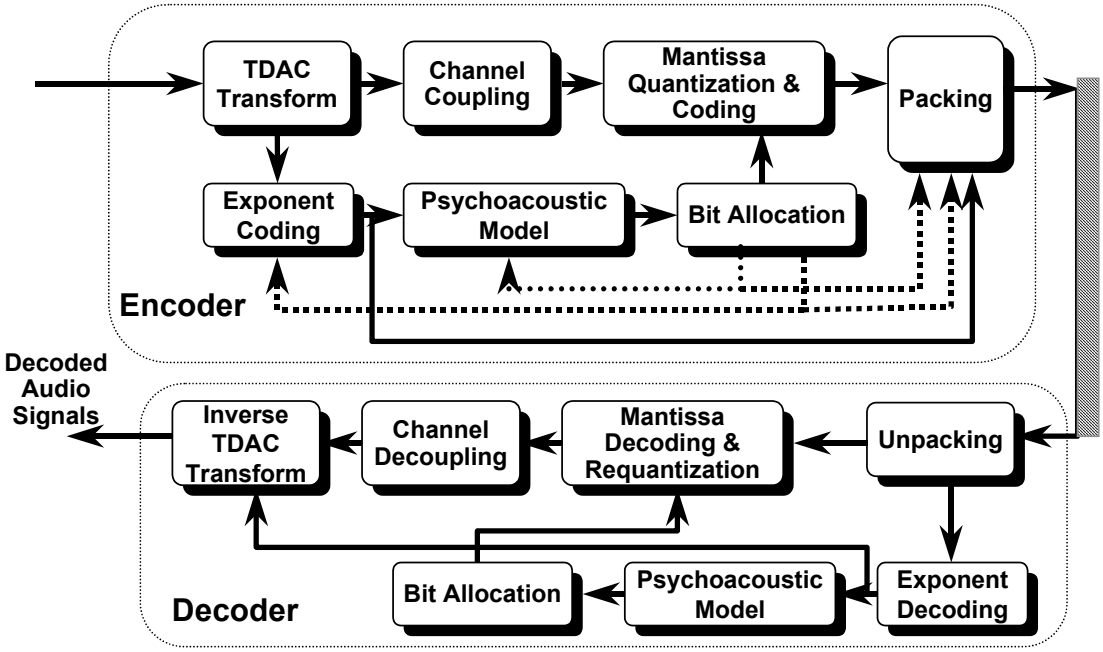

Fig. 5. Dolby AC-3 encoder and decoder.

As illustrated in Fig. 5, Dolby AC-3 encoder first employs a MDCT to transform the audio signals from the time domain to frequency domain. Then, adjacent transform coefficients are grouped into nonuniform subbands which approximate the critical bands of human auditory system. Transform coefficients within one subband are converted to a floating-point representation, with one or more mantissas per exponent. The exponents are encoded by a suitable strategy according to the required time and frequency resolution and fed into the psychoacoustic model. Then, the psychoacoustic model calculates the perceptual resolution according to the encoded exponents and the proper perceptual parameters. Finally, both the perceptual resolution and the available bits are used to decide the mantissa quantization.

One distinctive feature of Dolby AC-3 is the intimate relationship among exponent coding, psychoacoustic models, and the bit allocation. This relationship can be described most conveniently by the hybrid backward/forward bit allocation. The encoded exponents provide an estimate of the spectral envelope which, in turn, is used in the psychoacoustic model to

determine the mantissa quantization. While most audio encoders need to transmit side information about the mantissa quantization, the AC-3 decoder can automatically derive the quantizer information from the decoded exponents and limited perceptual parameters. The basic problem with this approach is that the exponents are subject to limited time-frequency resolution and hence fail to provide a detailed psychoacoustic analysis. The tradeoff between the bit merits of transmitting side information and the constraint psychoacoustic precision decides the coding efficiency of Dolby AC-3.

4. ARCHITECTURAL OVERVIEW

The principles of the perceptual coding can be considered according to eight aspects: the time/frequency mapping, quantization and coding, psychoacoustic model, channel correlation and irrelevancy, long-term correlation, pre-echo control, and bit-allocation. This section provides an overview of these standards through examination of the eight aspects.

4.1 Psychoacoustic Modeling

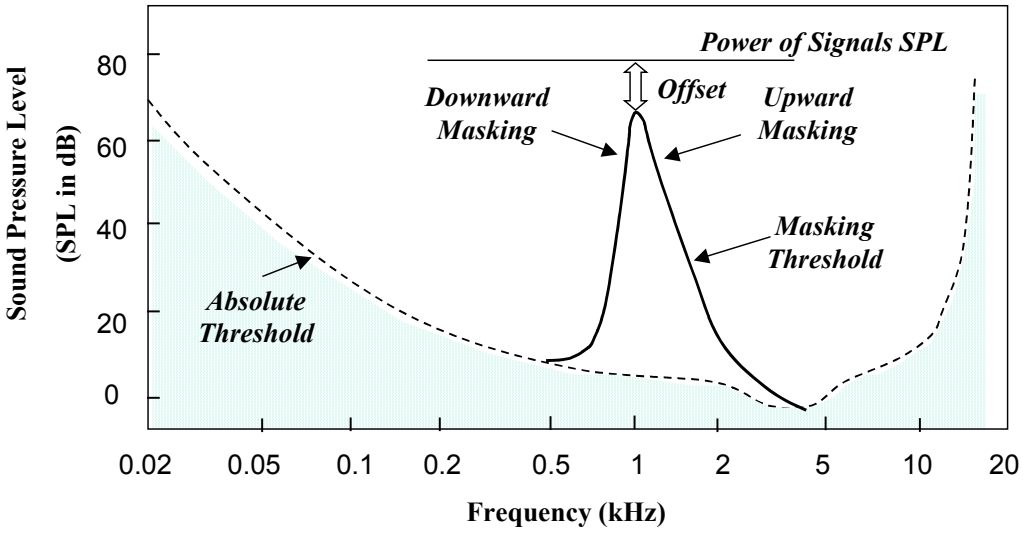

Fig. 6. Masking threshold of a masker centered at 1 kHz.

Most perceptual coders rely, at least to some extent, on the psychoacoustic models to reduce the subjective impairments of quantization noise. The encoder analyzes the incoming

audio signals to identify perceptually important information by incorporating several psychoacoustic principles of the human ear [Zwicker, 1990]. One is the critical-band spectral analysis, which accounts for the ear's poorer discrimination in the higher frequency region than in lower ones. Further investigations indicated that a good choice of spectral resolution is around 20 Hz, which has been implemented in MPEG-2 AAC and MPEG-4. The phenomenon of masking is another effect that occurs whenever a strong signal (masker) makes a spectral or temporal neighborhood of weaker signals inaudible. To illustrate this, Fig. 6 shows an example of the masking threshold produced by a masker centered at 1 kHz. The absolute threshold in the (dashed line) is also included to indicate the minimum audible intensity level in quiet surroundings. Notice that the slope of the masking curve is less steep on the high-frequency side; i.e., higher frequencies are more easily masked. The offset between masker and masking threshold is varied with respect to the tonality of the masker; it has a smaller value for noise-like masker (about 5.5 dB) than tone-like masker (above 25 dB).

The encoder performs the psychoacoustic analysis based on either a FFT analysis (in MPEG) or the output of the filterbank (in AC-3 and PASC). The psychoacoustic model used in AC-3 is specially designed; it does not provide a means to differentiate the masking effects produced by either the tonal or the noise masker. MPEG provides two examples of psychoacoustic models, the first of which we will now describe. The calculation starts with a precise spectral analysis on 512 (Layer I) or 1024 (Layer II) input samples to generate the magnitude spectrum. The spectral lines are then examined to discriminate between noise-like and tone-like maskers by taking the local maximum of magnitude spectrum as an indicator of tonality. Among all the labeled maskers, only those above the absolute threshold are retained for further calculation. Using rules known from psychoacoustics, the individual masking thresholds for the relevant maskers are then calculated dependent on frequency position, loudness level, and the nature of tonality. Finally, we obtain the global masking threshold from the upward and downward slopes of the individual masking thresholds of tonal and

nontonal maskers and from the absolute threshold in quiet.

4.2 Time-Frequency Mapping

Since psychoacoustic interpretation is mainly described in frequency domain, the time-frequency mapping is incorporated into the encoder for further signal analysis. The time-frequency mapping can be implemented either through PQMF [ISO, 1992], time-domain aliasing cancellation (TDAC) filters [Prince, 1987] or the modified discrete cosine transform [ISO, 1992]. All of them can be referred to as the cosine modulated filterbanks (CMFBs) [Shlien, 1997][Liu, 1998]. The process of CMFBs consists of two steps: the window-and-overlapping addition (WOA) followed by the modulated cosine transform (MCT). The WOA performs a windowing multiplication and addition with overlapping audio blocks. Its complexity is O(k) per audio sample, where k is the overlapping factor of audio blocks. In general, the sidelobe attenuation of the filterbank increases with the factor. For example, the factor k is 16 for MPEG-1 Layer II and is 2 for AC-3.

CMFBs In Standards	Overlap Factor k	Number of Bands N	Frequency Resolution at 48 kHz	Sidelobe Atten.
MPEG Layers I & II	16	32	750 Hz	96 dB
MPEG 2nd hybrid level of Layer III	2	18	41.66 Hz	23dB
MPEG-2 AAC, MPEG-4 T/F coding	2	1024	23.40 Hz	19dB
Dolby AC-3	2	256	93.75 Hz	18 dB

Table 4 CMFBs used in current audio coding standards

Classes	MCT Transform Pair	CMFBs in Standards
Polyphase Filterbank	$X_k = \sum_{i=0}^{N-1} x_i \cos(\frac{\pi}{N}(i-\frac{N}{4})(2k+1))$ $x_i = \sum_{i=0}^{N/2-1} X_k \cos(\frac{\pi}{2N}(i+\frac{N}{4})(2k+1))$ for k=0, 1 ..., N/2-1 and i=0, 1, ..., N-1	MPEG Layers I and II (N=64), MPEG Layer III 1st hybrid level (N=64)
TDAC Filterbank	$X_k = \sum_{i=0}^{N-1} x_i \cos(\frac{\pi}{2N}(2i+1+\frac{N}{2})(2k+1))$ $x_i = \sum_{k=0}^{N/2-1} X_k \cos(\frac{\pi}{2N}(2i+1+\frac{N}{2})(2k+1))$ for k=0, 1 ..., N/2-1 and i=0, 1, ..., N-1	MPEG-2—AAC (N=4096), MPEG-4- T/F Coding (N=4096) MPEG Layer III 2nd hybrid level (N=36), AC-3 Long Transform (N=512)
TDAC-Variant Filterbank	$X_k = \sum_{i=0}^{N-1} x_i \cos(\frac{\pi}{2N}(2i+1)(2k+1))$ $x_i = \sum_{k=0}^{N/2-1} X_k \cos(\frac{\pi}{2N}(2i+1)(2k+1))$ for k=0, 1 ..., N/2-1 and i=0, 1, ..., N-1	AC-3 Short Transform 1 (N=256)
	$X_k = \sum_{i=0}^{N-1} x_i \cos(\frac{\pi}{2N}(2i+1+N)(2k+1))$ $x_i = \sum_{k=0}^{N/2-1} X_k \cos(\frac{\pi}{2N}(2i+1+N)(2k+1))$ for k=0, 1 ..., N/2-1 and i=0, 1, ..., N-1	AC-3 Short Transform 2 (N=256)

Table 3 Comparison of filterbank properties

The complexity of MCT is O(2N) per audio sample, where N is the number of bands.

The range of N is from 18 for MPEG-1 Layer III to 2048 for the MPEG-2 advanced audio coding. Table 4 compares the properties of CMFBs used in audio coding standards. Due to the high complexity of MCT, fast algorithms have been developed following the similar concepts behind the fast Fourier transform. As listed in Table 4, the MCTs used in current audio standards can be classified into three different types: time-domain aliasing cancellation (TDAC), variant of the TDAC filterbank, and the **polyphase filterbank**.

4.3 Quantization

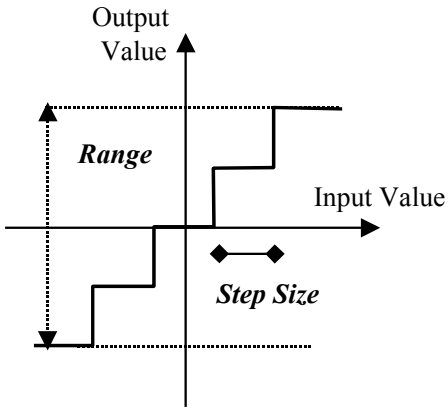

Fig. 7. Quantizer characteristics.

For perceptual audio coding, quantization involves representing the outputs of the filterbank by a finite number of levels with the aim of minimizing the subjective impairments of quantization noise. The characteristics of a quantizer can be specified by means of the step size and the range as shown in Fig. 7. While a uniform quantizer has the same step size throughout the input range, a nonuniform quantizer does not. For a uniform quantizer, the range and the step size determine the number of levels required for the output value. In contrast to a uniform quantizer, the quantization noise of a nonuniform quantizer is varied with respect to the input value. Such a design is more relevant to the human auditory system in the sense that the ear's ability to decipher two sounds with different volumes decreases with the sound pressure levels.

For a uniform quantization with fixed bit rate, the range directly affects the quantization error. Hence, an accurate estimation of the range leads to a good control of the quantization error. On the other hand, for nonuniform quantization, the quantization noise depends more on the input values instead of the ranges. In the current audio standards, uniform quantizers are used in MPEG Layer I, Layer II and Dolby AC-3. For MPEG Layers I and II, the scale factor

helps to determine the range of a quantizer. For Dolby AC-3, the exponent, which accounts for the range of a uniform quantizer, is adaptive with the time and frequency. Table 5 lists the quantization schemes used in the audio coding standards. For MPEG Layer III, MPEG-2 AAC, and MPEG-4 T/F coding, the ranges of nonuniform quantizers are not adaptive with the frequency.

Till now, we only considered the scalar quantization situation where one sample is quantized at a time. On the other hand, vector quantization (VQ) involves representing a block of input samples at a time. Twin-VQ has been adopted in MPEG-4 T/F coding as an alternative to scalar quantization for higher coding efficiency.

Standards	Quantization Types	Range adaptation with time at 48 kHz	Range adaptation with frequency at 48 kHz
MPEG Layer I	Uniform	8 ms	750 Hz
MPEG Layer II	Uniform	8 ms – 24 ms	750 Hz
MPEG Layers III	Nonuniform	24 ms	No
Dolby AC-3	Uniform	5.3 ms – 32 ms	93.75Hz – 375 Hz
MPEG-2 AAC	Nonuniform	21.33 ms	No

Table 5 Quantization schemes used in audio coding standards.

4.4 Variable Length Coding

Variable length coding has been used in MPEG Layer III and MPEG-2 AAC to enhance the coding efficiency. Since the quantization leads to a result that small values have higher probability of occurrence than the large values, the coding of small values with longer bit length and large values with shorter length can improve the coding efficiency. However, the variable length coding leads to two implementation problems. The first problem is on the decoder complexity. The variable length coding leads to the boundary vague of the coding

contents. Logically, the decoder has to decode the packing values bit-by-bit and symbol-by-symbol, which has induced much higher complexity than that by the fixed-length coding. Usually, a parallel or multiple table-look-up has been adopted to solve this problem. The second problem is on the bit-allocation. The variable bit requirement makes it difficult to allocate the limited bits to various parameters. This problem will be discussed in Subsection 4.8.

4.5 Multichannel Correlation and Irrelevancy

Standards	Stereo Correlation & Irrelevancy	Multichannel Transform
MPEG-1 Layer I	Intensity Coding	No
MPEG-1 Layer II	Intensity Coding	No
MPEG-1 Layer III	M/S Coding and Intensity coding	No
MPEG-2 Layers I	Intensity Coding	Matrixing
MPEG-2 Layers II	Intensity Coding	Matrixing
MPEG-2 Layers III	Intensity and M/S Coding	Matrixing
Dolby AC-3	Coupling and Matrixing	M/S and Coupling
MPEG-2 AAC	M/S and Intensity coding	M/S and Intensity coding
MPEG-4 T/F coding	M/S and Intensity coding	M/S and Intensity coding

Table 6 Multichannel coding for the audio standards

Since the signals from different channels are usually recorded from the same sound source, it is a natural speculation that lots of correlation exists among these channels. Furthermore, there is stereo irrelevancy for the multichannel signals. In particular, localization of the stereophonic image within a critical band for frequencies above 2 kHz is based more on the temporal envelope of the audio signal than on its temporal fine structure. Recognizing this, audio coding standards have developed techniques for dealing with the correlation and irrelevancy. For stereo audio, there are two types of coding techniques: middle/side (M/S)

coding and intensity coding. M/S coding transforms the left and right channels into the sum and difference signals to remove the correlation. In intensity coding mode the encoder combines some high-frequency parts of two-channel audio into a single summed signal by exploiting the stereo irrelevancy.

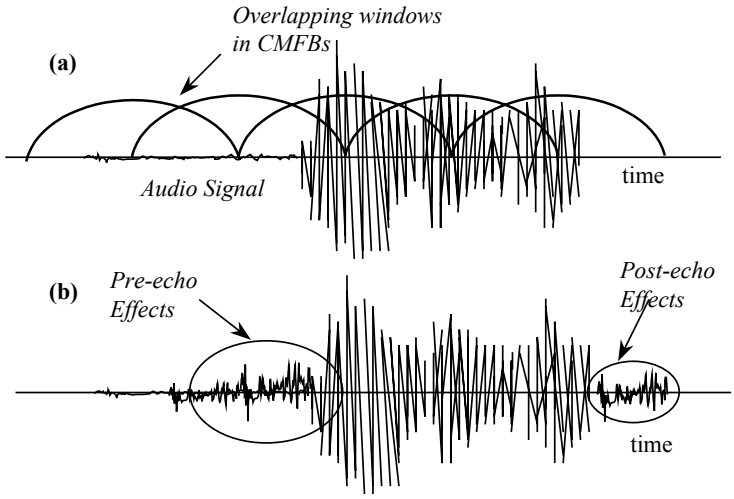

Fig. 8. Pre-echo effects. (a) Original signal (b) Reconstructed signal with pre-echo and post-echo.

For channel numbers greater than two, like M/S coding, the signals in the channels can be transformed into another signals with the same channel number to remove their correlation. This is typically done by using an *N-by-N* matrix, where N is the channel number. As with intensity coding, the signals of the N channels in high-frequency regions can also be combined into one channel, which is called the coupling scheme in multichannel scenario. MPEG-2 AAC separates all the channels in pairs and then applies the M/S and intensity coding to them. Table 6 lists the multichannel coding schemes used in various audio standards.

4.6 Long-term Correlation

When dealing with stationary signals, an audio encoder can benefit from the correlation between adjacent frames to achieve further coding gain. In the current audio coding standards, MPEG-2 AAC and MPEG-4 T/F coding are two typical examples that make use of a second-order backward prediction tool to exploit the inter-frame correlation. The importance of the prediction tool increases with the sampling rates.

4.7 Pre-echo Control

Pre-echoes occur when a region of low power is followed by a signal with a sharp attack as showed in Fig. 8(a). Using a CMFB with its window length covering these two regions, the inverse transform will spread quantization noise evenly throughout the whole region. Hence, the quantization noise in the region of low power will be too large to be masked. This is called the pre-echo effect as shown in Fig. 8(b). Pre-echoes can be masked by the pre-masking phenomenon of human auditory system [Zwicker, 1990] only if the window length is sufficiently small (about 1-4 ms). Similarly, the post-echo effect can be found in the situation where a signal with a sharp attack is followed by a region of low power. Fortunately, the post-echo effect is not so serious because the duration within which post-masking applies is in the order of 50 to 200 ms. A solution to the pre-echo problem is to switch between window sizes of different lengths. This approach has been implemented in MPEG Layer III, Dolby AC-3, MPEG-2 AAC, and MPEG-4 T/F coding. Table 7 lists the long and the short window used in various audio coding standards. In addition to the block size switching, MPEG-2 AAC and MPEG-4 T/F coding make use of the temporal noise shaping tool to control the pre-echo effect. This tool performs a forward prediction in the frequency domain and hence leads to a temporal shaping in time domain. Through this shaping, the small signal and the large signal can be shaped to similar amplitudes which, in turn, helps to reduce the pre-echo effects.

Standards	Length of long window at 48 kHz sampling frequency	Length of short window at 48 kHz sampling frequency
MPEG Layer III	48 ms	16 ms
Dolby AC-3	10.67 ms	2.66 ms
MPEG-2 AAC	42.67 ms	2.66 ms
MPEG-4 T/F Coding	42.67 ms	2.66 ms

Table 7 Length of the long and short windows.

4.8 Bit-Allocation

The bit allocation is aimed to assign suitable parameters to the encoder to achieve the best audio quality under the restricted bit number. Hence control over the quality and the bit number are two fundamental requirements for the bit allocation. The complexity of the task depends on the difficulties to have the quality and bit control. For MPEG Layers I and II, both the quality and the bit requirement are controlled by a uniform quantizer. Hence the bit allocation is just to apportion the total number of bits available for the quantization of the subband signals to minimize the audibility of the quantization noise.

For MPEG Layer III and MPEG-2 AAC, control over the quality and the bit rate is difficult. This is mainly due to the fact that they both use a nonuniform quantizer whose quantization noise is varied with respect to the input values. In other words, it fails to control the quality by assigning quantizer parameters according to the perceptually allowable noise. In addition, the bit-rate control issue can be examined from the variable length coding used in MPEG Layer III and MPEG-2 AAC. The variable length coding assigns variable bit-length to different values, which means that the bits consumed should be obtained from the quantization results, and can not be from the quantizer parameters alone. Thus, the bit allocation is one of the main tasks leading to the high complexity of the encoder.

For Dolby AC-3, it is also difficult to determine the bit allocation. As mentioned above, AC-3 adapts its range according to the specified exponent strategy. There are 3072 possible strategies for the six blocks in a frame. These strategies affect the temporal resolution and the spectral resolution of the quantization ranges. These encoded exponents also affect the analysis result of the psychoacoustic model, which is a special feature of the hybrid coding in Dolby AC-3. The exponents and the resultant psychoacoustic results determine the quantization results. Hence the intimate relation among the exponents, the psychoacoustic

models, and the quantization has led to high complexity in bit allocation. This issues and the solution on the bit allocation in Dolby AC-3 has been analyzed in the paper [Liu, 1998].

5. CONCLUSIONS

With the standardization efforts of MPEG, audio coding has made historic success in the area of multimedia and consumer applications. A variety of compression algorithms has been provided ranging from MPEG-1 Layer I, which allows for an efficient decoder implementation, to MPEG-2 AAC, which provides high audio quality at a rate of 384 kb/s for five full-bandwidth channels. All of the important compression schemes build on the proven frequency-domain coding techniques in conjunction with block companding and perceptual bit allocation strategies. They are designed not just for using statistical correlation to remove redundancies but also to eliminate the perceptual irrelevancy by applying psychoacoustic principles. While current audio coders still have room for improvement in terms of bit rates and quality, the main focus of current and future work has been switched to offer new functionalities such as scalability and editability, and thereby opening the way to new applications.

Defining Terms

Critical band: Psychoacoustic measure in the spectral domain that corresponds to the frequency selectivity of the human ear.

Intensity stereo: A method of exploiting stereo irrelevance or redundancy in stereophonic audio programs based on retaining at high frequencies only the energy envelope of the right and left channels.

MDCT: Modified Discrete Cosine Transform that corresponds to the Time Domain Aliasing Cancellation Filterbank.

Polyphase filterbank: A set of equal bandwidth filters with special phase interrelationships,

allowing for an efficient implementation of the filterbank.

References

ISO/IEC JTC1/SC29/WG11. 1992. Information technology- Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mb/s— IS 11172 (Part3, Audio).

ISO/IEC JTC1/SC29/WG11. 1994. Information technology- Generic coding of moving pictures and associated audio information— IS 13818 (Part 3, Audio).

ISO/IEC JTC1/SC29/WG11. 1997. Information Technology- Generic coding of moving pictures and associated audio information— IS 13818 (Part 7, Advanced audio coding).

ISO/IEC JTC1/SC29/WG11. 1999. Information Technology- Coding of audiovisual objects—ISO/IEC.D 4496 (Part 3, Audio).

Liu, C.M. and Lee, W.J. 1998. A unified fast algorithm for the current audio coding standards. Proc. Audio Engineering Society Conv. preprint 4729.

Liu, C.M., Lee, S.W., and Lee, W.C. 1998. Bit allocation method for Dolby AC-3 encoder. *IEEE Trans. on Consumer Electronics.* 44(3): 883-887.

Lokhoff, G.C.P. 1992. Precision Adaptive Subband Coding (PASC) for the Digital Compact Cassette (DCC). *IEEE Trans. Consumer Electronics.* 38(4): 784-789.

Prince, J.P., Johnson, A.W., and Bradley, A.B. 1987. Subband/transform coding using filterbank design based on time domain aliasing cancellation. Proc. ICASSP: 2161-2164.

Shlien, S. 1997. The Modulated lapped transform, its time-varying forms, and its application to audio coding standards. *IEEE Trans. on Speech and Audio Process.* 5(4): 359-366.

Todd, C.C., Davidson, G.A., Davis, M.F., Fielder, L.D., Link, B.D., and Vernon S. 1994. AC-3: Flexible perceptual coding for audio transmission and storage. Proc. Audio Engineering Society 96th Conv. preprint 3796. Amsterdam, May.

Tsutsui, K., Suzuki, H., Shimoyoshi, O., Sonohara, M., Akagiri, K., and Heddle, R.M. 1992.

ATRAC: Adaptive Transform Acoustic Coding for MiniDisc. Proc. Audio Engineering Society 93rd Conv. preprint 3456. San Francisco, Oct.

Zwicker, E. and Fasti, H. 1990. *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin.

Further Information

Additional technical information can be found in the following references.

1. Brandenburg, K. and Bosi, M. 1997. Overview of MPEG audio: Current and future standards for low-bit-rate audio coding. *J. Audio Eng. Soc.* 45(1/2):4-19.
2. Noll P. 1997. MPEG digital audio coding. *IEEE Signal Processing Magazine*. 14(5):59-81.
3. Pan, D. 1995. A tutorial on MPEG/Audio compression, *IEEE Multimedia Magazine*, 2(2): 60-74.
4. Pohlmann, K.G. 1995. *Principles of Digital Audio*, 3rd ed. McGraw-Hill, New York.