Peter Noll. “MPEG Digital Audio Coding Standards.”
2000 CRC Press LLC. <http://www.engnetbase.com>.

MPEG Digital Audio Coding Standards

Peter Noll
Technical University of Berlin

40.1 Introduction

PCM Bit Rates

Typical audio signal classes are telephone speech, wideband speech, and wideband audio, all of which differ in bandwidth, dynamic range, and in listener expectation of offered quality. The quality of telephone-bandwidth speech is acceptable for telephony and for some videotelephony and video-conferencing services. Higher bandwidths (7 kHz for wideband speech) may be necessary to improve the intelligibility and naturalness of speech. Wideband (high fidelity) audio representation including multichannel audio needs bandwidths of at least 15 kHz.

The conventional digital format for these signals is PCM, with sampling rates and amplitude resolutions (PCM bits per sample) as given in Table 40.1.

The *compact disc* (CD) is today's *de facto standard* of digital audio *representation*. On a CD with its 44.1 kHz sampling rate the resulting stereo net bit rate is $2 \times 44.1 \times 16 \times 1000 \equiv 1.41$ Mb/s (see Table 40.2). However, the CD needs a significant overhead for a runlength-limited line code, which maps 8 information bits into 14 bits, for synchronization and for error correction, resulting in a 49-bit representation of each 16-bit audio sample. Hence, the total stereo bit rate is $1.41 \times 49/16 =$ 4.32Mb/s. Table 40.2 compares bit rates of the compact disc and the *digital audio tape* (DAT).

TABLE 40.1 Basic Parameters for Three Classes of Acoustic Signals

	Frequency range in Hz	Sampling rate in kHz	PCM bits per sample	PCM bit rate in kb/s
Telephone speech	300 - 3,400[a]	8	8	64
Wideband speech	50 - 7,000	16	8	128
Wideband audio (stereo)	10 - 20,000	48[b]	2 × 16	2 × 768

[a] Bandwidth in Europe; 200 to 3200 Hz in the U.S.

[b] Other sampling rates: 44.1 kHz, 32 kHz.

TABLE 40.2 CD and DAT Bit Rates

Storage device	Audio rate (Mb/s)	Overhead (Mb/s)	Total bit rate (Mb/s)
Compact disc (CD)	1.41	2.91	4.32
Digital audio tape (DAT)	1.41	1.05	2.46

Note: Stereophonic signals, sampled at 44.1 kHz; DAT supports also sampling rates of 32 kHz and 48 kHz.

For archiving and processing of audio signals, sampling rates of at least 2×44.1 kHz and amplitude resolutions of up to 24 b per sample are under discussion. Lossless coding is an important topic in order not to compromise audio quality in any way [1]. The digital versatile disk (DVD) with its capacity of 4.7 GB is the appropriate storage medium for such applications.

Bit Rate Reduction

Although high bit rate channels and networks become more easily accessible, low bit rate coding of audio signals has retained its importance. The main motivations for low bit rate coding are the need to minimize transmission costs or to provide cost-efficient storage, the demand to transmit over channels of limited capacity such as mobile radio channels, and to support variable-rate coding in packet-oriented networks.

Basic requirements in the design of low bit rate audio coders are first, to retain a high quality of the reconstructed signal with robustness to variations in spectra and levels. In the case of stereophonic and multichannel signals spatial integrity is an additional dimension of quality. Second, robustness against random and bursty channel bit errors and packet losses is required. Third, low complexity and power consumption of the codecs are of high relevance. For example, in broadcast and playback applications, the complexity and power consumption of audio decoders used must be low, whereas constraints on encoder complexity are more relaxed. Additional network-related requirements are low encoder/decoder delays, robustness against errors introduced by cascading codecs, and a graceful degradation of quality with increasing bit error rates in mobile radio and broadcast applications. Finally, in professional applications, the coded bit streams must allow editing, fading, mixing, and dynamic range compression [1].

We have seen rapid progress in bit rate compression techniques for speech and audio signals [2]–[7]. Linear prediction, subband coding, transform coding, as well as various forms of vector quantization and entropy coding techniques have been used to design efficient coding algorithms which can achieve substantially more compression than was thought possible only a few years ago. Recent results in speech and audio coding indicate that an excellent coding quality can be obtained with bit rates of 1 b per sample for speech and wideband speech and 2 b per sample for audio. Expectations over the next decade are that the rates can be reduced by a factor of four. Such reductions shall be based mainly on employing sophisticated forms of adaptive noise shaping controlled by psychoacoustic criteria. In storage and ATM-based applications additional savings are possible by employing variable-rate coding with its potential to offer a time-independent constant-quality performance.

Compressed digital audio representations can be made less sensitive to channel impairments than analog ones if source and channel coding are implemented appropriately. Bandwidth expansion has often been mentioned as a disadvantage of digital coding and transmission, but with today's data compression and multilevel signaling techniques, channel bandwidths can be reduced actually, compared with analog systems. In broadcast systems, the reduced bandwidth requirements, together with the error robustness of the coding algorithms, will allow an efficient use of available radio and TV channels as well as “taboo” channels currently left vacant because of interference problems.

MPEG Standardization Activities

Of particular importance for digital audio is the standardization work within the International Organization for Standardization (ISO/IEC), intended to provide international standards for audio-visual coding. ISO has set up a Working Group WG 11 to develop such standards for a wide range of communications-based and storage-based applications. This group is called MPEG, an acronym for *Moving Pictures Experts Group*.

MPEG’s initial effort was the MPEG Phase 1 (MPEG-1) coding standards *IS 11172* supporting bit rates of around 1.2 Mb/s for video (with video quality comparable to that of today’s analog video cassette recorders) and 256 kb/s for two-channel audio (with audio quality comparable to that of today’s compact discs) [8].

The more recent MPEG-2 standard *IS 13818* provides standards for high quality video (including High Definition TV) in bit rate ranges from 3 to 15 Mb/s and above. It provides also new audio features including low bit rate digital audio and multichannel audio [9].

Finally, the current MPEG-4 work addresses standardization of audiovisual coding for applications ranging from mobile access low complexity multimedia terminals to high quality multichannel sound systems. MPEG-4 will allow for interactivity and universal accessibility, and will provide a high degree of flexibility and extensibility [10].

MPEG-1, MPEG-2, and MPEG-4 standardization work will be described in Sections 40.3 to 40.5 of this paper. *Web information about MPEG* is available at different addresses. The official MPEG Web site offers crash courses in MPEG and ISO, an overview of current activities, MPEG requirements, workplans, and information about documents and standards [11]. Links lead to collections of frequently asked questions, listings of MPEG, multimedia, or digital video related products, MPEG/Audio resources, software, audio test bitstreams, etc.

40.2 Key Technologies in Audio Coding

First proposals to reduce wideband audio coding rates have followed those for speech coding. Differences between audio and speech signals are manifold; however, audio coding implies higher sampling rates, better amplitude resolution, higher dynamic range, larger variations in power density spectra, stereophonic and multichannel audio signal presentations, and, finally, higher listener expectation of quality. Indeed, the high quality of the CD with its 16-b per sample PCM format has made digital audio popular.

Speech and audio coding are similar in that in both cases quality is based on the properties of human auditory perception. On the other hand, speech can be coded very efficiently because a *speech production model* is available, whereas nothing similar exists for audio signals.

Modest reductions in audio bit rates have been obtained by instantaneous companding (e.g., a conversion of uniform 14-bit PCM into a 11-bit nonuniform PCM presentation) or by forward-adaptive PCM (block companding) as employed in various forms of *near-instantaneously companded audio multiplex* (NICAM) coding [ITU-R, Rec. 660]. For example, the British Broadcasting Corporation (BBC) has used the NICAM 728 coding format for digital transmission of sound in several European broadcast television networks; it uses 32-kHz sampling with 14-bit initial quantization followed by a compression to a 10-bit format on the basis of 1-ms blocks resulting in a total stereo bit rate of 728 kb/s [12]. Such adaptive PCM schemes can solve the problem of providing a sufficient dynamic range for audio coding but they are not efficient compression schemes because they do not exploit statistical dependencies between samples and do not sufficiently remove signal irrelevancies.

Bit rate reductions by fairly simple means are achieved in the interactive CD (CD-i) which supports 16-bit PCM at a sampling rate of 44.1 kHz and allows for three levels of adaptive differential PCM (ADPCM) with switched prediction and noise shaping. For each block there is a multiple choice of fixed predictors from which to choose. The supported bandwidths and b/sample-resolutions are 37.8 kHz/8 bit, 37.8 kHz/4 bit, and 18.9 kHz/4 bit.

In recent audio coding algorithms *four key technologies* play an important role: perceptual coding, frequency domain coding, window switching, and dynamic bit allocation. These will be covered next.

40.2.1 Auditory Masking and Perceptual Coding

Auditory Masking

The inner ear performs short-term critical band analyses where frequency-to-place transformations occur along the basilar membrane. The power spectra are not represented on a linear frequency scale but on limited frequency bands called *critical bands*. The auditory system can roughly be described as a bandpass filterbank, consisting of strongly overlapping bandpass filters with bandwidths in the order of 50 to 100 Hz for signals below 500 Hz and up to 5000 Hz for signals at high frequencies. Twenty-five critical bands covering frequencies of up to 20 kHz have to be taken into account.

Simultaneous masking is a frequency domain phenomenon where a low-level signal (the maskee) can be made inaudible (masked) by a simultaneously occurring stronger signal (the masker), if masker and maskee are close enough to each other in frequency [13]. Such masking is greatest in the critical band in which the masker is located, and it is effective to a lesser degree in neighboring bands. A *masking threshold* can be measured below which the low-level signal will not be audible. This masked signal can consist of low-level signal contributions, quantization noise, aliasing distortion, or transmission errors. The masking threshold, in the context of source coding also known as *threshold of just noticeable distortion* (JND) [14], varies with time. It depends on the sound pressure level (SPL), the frequency of the masker, and on characteristics of masker and maskee. Take the example of the masking threshold for the SPL = 60 dB narrowband masker in Fig. 40.1: around 1 kHz the four maskees will be masked as long as their individual sound pressure levels are below the masking threshold. The slope of the masking threshold is steeper towards lower frequencies, i.e., higher frequencies are more easily masked. It should be noted that the distance between masker and masking threshold is smaller in noise-masking-tone experiments than in tone-masking-noise experiments, i.e., noise is a better masker than a tone. In MPEG coders both thresholds play a role in computing the masking threshold.

Without a masker, a signal is inaudible if its sound pressure level is below the *threshold in quiet* which depends on frequency and covers a dynamic range of more than 60 dB as shown in the lower curve of Figure 40.1.

The qualitative sketch of Fig. 40.2 gives a few more details about the masking threshold: a critical band, tones below this threshold (darker area) are masked. The distance between the level of the masker and the masking threshold is called *signal-to-mask ratio (SMR)*. Its maximum value is at the left border of the critical band (point *A* in Fig. 40.2), its minimum value occurs in the frequency range of the masker and is around 6 dB in noise-masks-tone experiments. Assume a m-bit quantization of an audio signal. Within a critical band the quantization noise will not be audible as long as its signal-to-noise ratio SNR is higher than its SMR. Noise *and* signal contributions *outside* the particular critical band will also be masked, although to a lesser degree, if their SPL is below the masking threshold.

Defining SNR(m) as the signal-to-noise ratio resulting from an m-bit quantization, the perceivable distortion in a given subband is measured by the *noise-to-mask ratio*

$$\mathrm{NMR}\,(\mathrm{m}) = \mathrm{SMR} - \mathrm{SNR}\,(\mathrm{m})\ (\text{in dB}).$$

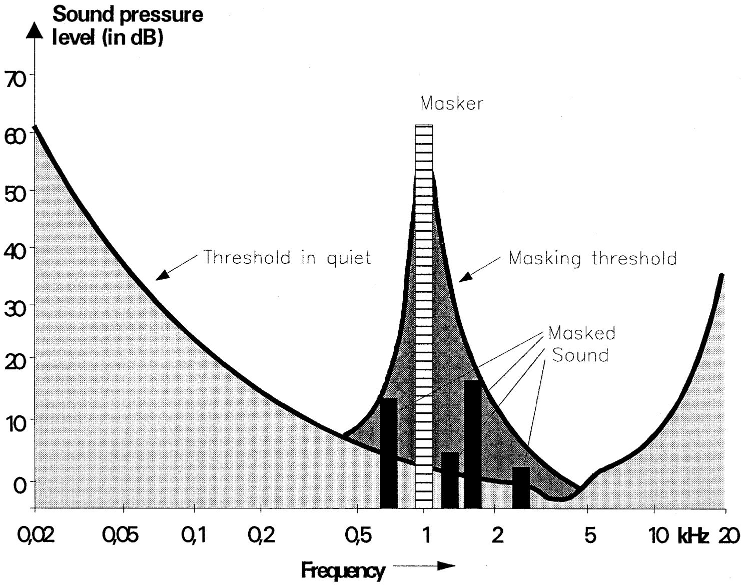

FIGURE 40.1: Threshold in quiet and masking threshold. Acoustical events in the shaded areas will not be audible.

The noise-to-mask ratio NMR(m) describes the difference in dB between the signal-to-*mask* ratio and the signal-to-*noise* ratio to be expected from an m-bit quantization. The NMR value is also the difference (in dB) between the level of quantization noise and the level where a distortion may just become audible in a given subband. Within a critical band, coding noise will not be audible as long as NMR(m) is negative.

We have just described masking by only one masker. If the source signal consists of many simultaneous maskers, each has its own masking threshold, and a *global masking threshold* can be computed that describes the threshold of just noticeable distortions as a function of frequency.

In addition to simultaneous masking, the time domain phenomenon of *temporal masking* plays an important role in human auditory perception. It may occur when two sounds appear within a small interval of time. Depending on the individual sound pressure levels, the stronger sound may mask the weaker one, even if the maskee precedes the masker (Fig. 40.3)!

Temporal masking can help to mask pre-echoes caused by the spreading of a sudden large quantization error over the actual coding block. The duration within which *pre-masking* applies is significantly less than one tenth of that of the *post-masking* which is in the order of 50 to 200 ms. Both pre- and postmasking are being exploited in MPEG/Audio coding algorithms.

Perceptual Coding

Digital coding at high bit rates is dominantly waveform-preserving, i.e., the amplitude-vs.-time waveform of the decoded signal approximates that of the input signal. The difference signal between input and output waveform is then the basic error criterion of coder design. Waveform coding principles have been covered in detail in [2]. At lower bit rates, facts about the production and perception of audio signals have to be included in coder design, and the error criterion has to be in favor of an output signal that is useful to the human receiver rather than favoring an output signal that follows and preserves the input waveform. Basically, an efficient source coding algorithm will (1) remove redundant components of the source signal by exploiting correlations between its

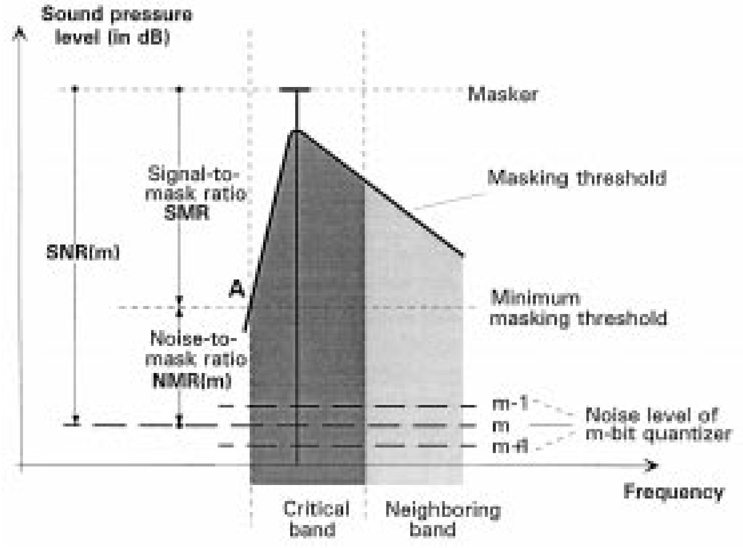
Sound pressure level (in dB)
Masker
Signal-to-mask ratio SMR
Masking threshold
SNR(m)
A
Minimum masking threshold
Noise-to-mask ratio NMR(m)
m-1
m
m+1
Noise level of m-bit quantizer
Frequency
Critical band
Neighboring band

FIGURE 40.2: Masking threshold and signal-to-mask ratio (SMR). Acoustical events in the shaded areas will not be audible.

samples and (2) remove components that are irrelevant to the ear. Irrelevancy manifests itself as unnecessary amplitude or frequency resolution; portions of the source signal that are masked do not need to be transmitted.

The dependence of human auditory perception on frequency and the accompanying perceptual tolerance of errors can (and should) directly influence encoder designs; *noise-shaping techniques* can emphasize coding noise in frequency bands where that noise perceptually is not important. To this end, the noise shifting must be dynamically adapted to the actual short-term input spectrum in accordance with the signal-to-mask ratio which can be done in different ways. However, frequency weightings based on linear filtering, as typical in speech coding, cannot make full use of results from psychoacoustics. Therefore, in wideband audio coding, noise-shaping parameters are dynamically controlled in a more efficient way to exploit simultaneous masking and temporal masking.

Figure 40.4 depicts the structure of a *perception-based coder* that exploits auditory masking. The

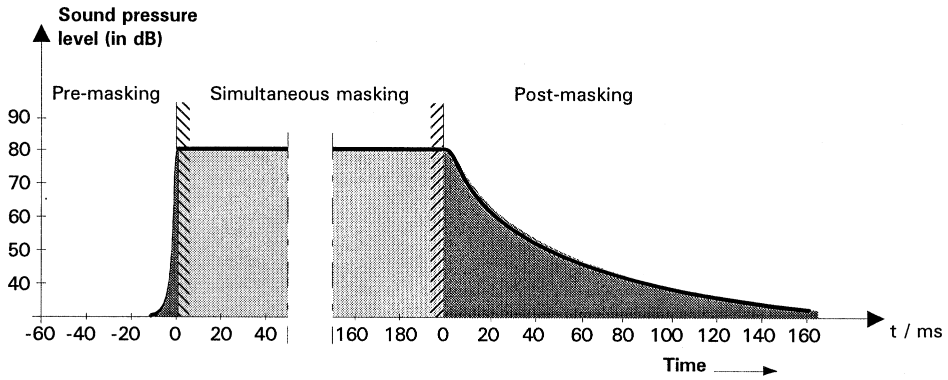

FIGURE 40.3: Temporal masking. Acoustical events in the shaded areas will not be audible.

encoding process is controlled by the SMR vs. frequency curve from which the needed amplitude resolution (and hence the bit allocation and rate) in each frequency band is derived. The SMR is typically determined from a high resolution, say, a 1024-point FFT-based spectral analysis of the audio block to be coded. Principally, any coding scheme can be used that can be dynamically controlled by such perceptual information. Frequency domain coders (see next section) are of particular interest because they offer a direct method for noise shaping. If the frequency resolution of these coders is high enough, the SMR can be derived directly from the subband samples or transform coefficients without running a FFT-based spectral analysis in parallel [15, 16].

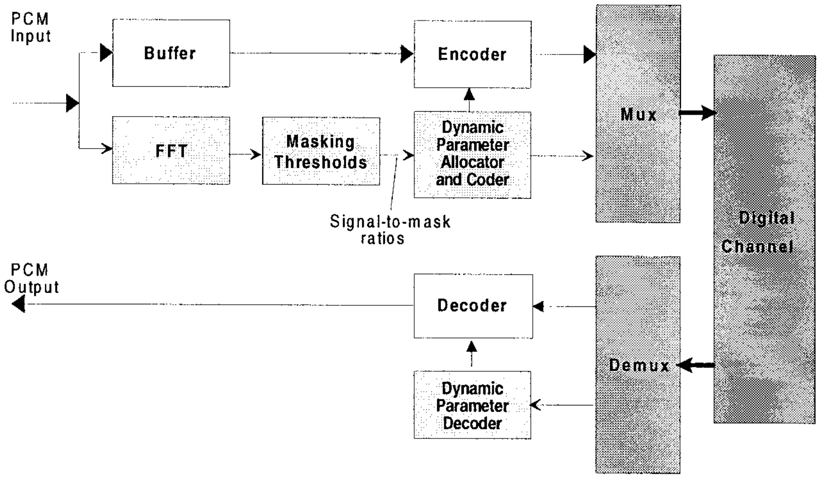

FIGURE 40.4: Block diagram of perception-based coders.

If the necessary bit rate for a complete masking of distortion is available, the coding scheme will be perceptually transparent, i.e., the decoded signal is then subjectively indistinguishable from the source signal. In practical designs, we cannot go to the limits of just noticeable distortion because postprocessing of the acoustic signal by the end-user and multiple encoding/decoding processes in transmission links have to be considered. Moreover, our current knowledge about auditory masking is very limited. Generalizations of masking results, derived for simple and stationary maskers and for limited bandwidths, may be appropriate for most source signals, but may fail for others. Therefore, as an additional requirement, we need a sufficient safety margin in practical designs of such perception-based coders. It should be noted that the MPEG/Audio coding standard is open for better encoder-located psychoacoustic models because such models are not normative elements of the standard (see Section 40.3).

40.2.2 Frequency Domain Coding

As one example of dynamic noise-shaping, quantization noise feedback can be used in predictive schemes [17, 18]. However, frequency domain coders with dynamic allocations of bits (and hence of quantization noise contributions) to subbands or transform coefficients offer an easier and more accurate way to control the quantization noise [2, 15].

In all frequency domain coders, redundancy (the non-flat short-term spectral characteristics of the source signal) and irrelevancy (signals below the psychoacoustical thresholds) are exploited to reduce the transmitted data rate with respect to PCM. This is achieved by splitting the source spectrum into frequency bands to generate nearly uncorrelated spectral components, and by quantizing these separately. Two coding categories exist, *transform coding* (TC) and *subband coding* (SBC). The differentiation between these two categories is mainly due to historical reasons. Both use an analysis filterbank in the encoder to decompose the input signal into subsampled spectral components. The spectral components are called subband samples if the filterbank has low frequency resolution, otherwise they are called spectral lines or transform coefficients. These spectral components are recombined in the decoder via synthesis filterbanks.

In *subband coding*, the source signal is fed into an analysis filterbank consisting of M bandpass filters which are contiguous in frequency so that the set of subband signals can be recombined additively to produce the original signal or a close version thereof. Each filter output is critically decimated (i.e., sampled at twice the nominal bandwidth) by a factor equal to M, the number of bandpass filters. This decimation results in an aggregate number of subband samples that equals that in the source signal. In the receiver, the sampling rate of each subband is increased to that of the source signal by filling in the appropriate number of zero samples. Interpolated subband signals appear at the bandpass outputs of the synthesis filterbank. The sampling processes may introduce aliasing distortion due to the overlapping nature of the subbands. If perfect filters, such as two-band quadrature mirror filters or polyphase filters, are applied, aliasing terms will cancel and the sum of the bandpass outputs equals the source signal in the absence of quantization [19]–[22]. With quantization, aliasing components will not cancel ideally; nevertheless, the errors will be inaudible in MPEG/Audio coding if a sufficient number of bits is used. However, these errors may reduce the original dynamic range of 20 bits to around 18 bits [16].

In *transform coding*, a block of input samples is linearly transformed via a discrete transform into a set of near-uncorrelated transform coefficients. These coefficients are then quantized and transmitted in digital form to the decoder. In the decoder, an inverse transform maps the signal back into the time domain. In the absence of quantization errors, the synthesis yields exact reconstruction. Typical transforms are the Discrete Fourier Transform or the Discrete Cosine Transform (DCT), calculated via an FFT, and modified versions thereof. We have already mentioned that the decoder-based inverse transform can be viewed as the synthesis filterbank, the impulse responses of its bandpass filters equal the basis sequences of the transform. The impulse responses of the analysis filterbank are just the time-reversed versions thereof. The finite lengths of these impulse responses may cause so-called block boundary effects. State-of-the-art transform coders employ a *modified DCT* (MDCT) filterbank as proposed by Princen and Bradley [21]. The MDCT is typically based on a 50% overlap between successive analysis blocks. Without quantization they are free from block boundary effects, have a higher transform coding gain than the DCT, and their basis functions correspond to better bandpass responses. In the presence of quantization, block boundary effects are deemphasized due to the doubling of the filter impulse responses resulting from the overlap.

Hybrid filterbanks, i.e., combinations of discrete transform and filterbank implementations, have frequently been used in speech and audio coding [23, 24]. One of the advantages is that different frequency resolutions can be provided at different frequencies in a flexible way and with low complexity. A high spectral resolution can be obtained in an efficient way by using a cascade of a filterbank (with its short delays) and a linear MDCT transform that splits each subband sequence further in frequency content to achieve a high frequency resolution. MPEG-1/Audio coders use a subband approach in layers I and II, and a hybrid filterbank in layer III.

40.2.3 Window Switching

A crucial part in frequency domain coding of audio signals is the appearance of *pre-echoes*, similar to copying effects on analog tapes. Consider the case that a silent period is followed by a percussive sound, such as from castanets or triangles, within the same coding block. Such an onset (“attack”) will cause comparably large instantaneous quantization errors. In TC, the inverse transform in the *decoding* process will distribute such errors over the block; similarly, in SBC, the decoder bandpass filters will spread such errors. In both mappings pre-echoes can become distinctively audible, especially at low bit rates with comparably high error contributions. Pre-echoes can be masked by the time domain effect of pre-masking if the time spread is of short length (in the order of a few milliseconds). Therefore, they can be reduced or avoided by using blocks of short lengths. However, a larger percentage of the total bit rate is typically required for the transmission of side information if the blocks are shorter. A solution to this problem is to switch between block sizes of different lengths as proposed by Edler (*window switching*) [25], typical block sizes are between N = 64 and N = 1024. The small blocks are only used to control pre-echo artifacts during nonstationary periods of the signal, otherwise the coder switches back to long blocks. It is clear that the block size selection has to be based on an analysis of the characteristics of the actual audio coding block. Figure 40.5 demonstrates the effect in transform coding: if the block size is N = 1024 [Fig. 40.5(b)] pre-echoes are clearly (visible and) audible whereas a block size of 256 will reduce these effects because they are limited to the block where the signal attack and the corresponding quantization errors occur [Fig. 40.5(c)]. In addition, pre-masking can become effective.

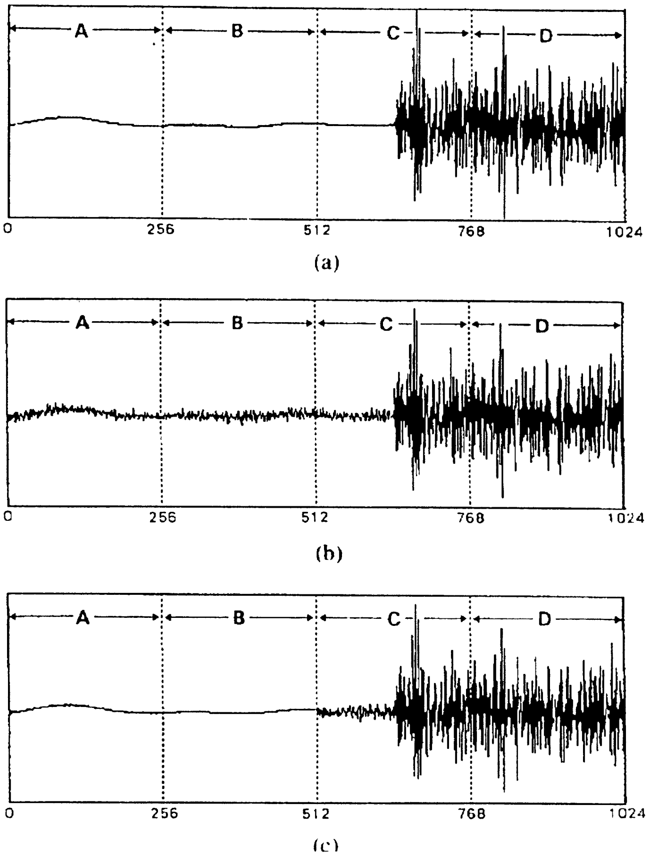

FIGURE 40.5: Window switching. (a) Source signal, (b) reconstructed signal with block size N = 1024, and (c) reconstructed signal with block size N = 256. (*Source:* Iwadare, M., Sugiyama, A., Hazu, F., Hirano, A., and Nishitani, T., IEEE J. Sel. Areas Commun., 10(1), 138-144, Jan. 1992.)

40.2.4 Dynamic Bit Allocation

Frequency domain coding significantly gains in performance if the number of bits assigned to each of the quantizers of the transform coefficients is adapted to short-term spectrum of the audio coding block on a block-by-block basis. In the mid-1970s, Zelinski and Noll introduced *dynamic bit allocation* and demonstrated significant SNR-based and subjective improvements with their adaptive transform coding (ATC, see Fig. 40.6 [15, 27]). They proposed a DCT mapping and a dynamic bit allocation algorithm which used the DCT transform coefficients to compute a DCT-based short-term spectral envelope. Parameters of this spectrum were coded and transmitted. From these parameters, the short-term spectrum was estimated using linear interpolation in the log-domain. This estimate was then used to calculate the optimum number of bits for each transform coefficient, both in the encoder and decoder.

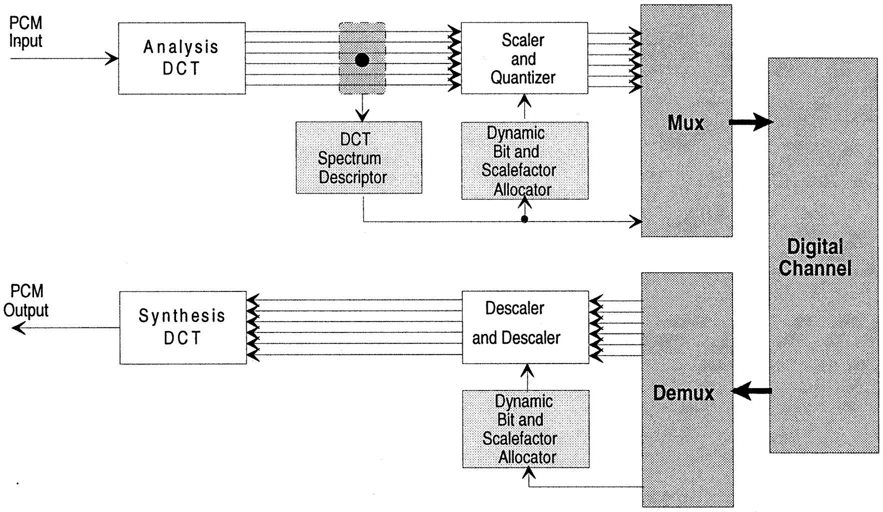

FIGURE 40.6: Conventional adaptive transform coding (ATC).

That ATC had a number of shortcomings, such as block boundary effects, pre-echoes, marginal exploitation of masking, and insufficient quality at low bit rates. Despite these shortcomings, we find many of the features of the conventional ATC in more recent frequency domain coders.

MPEG/Audio coding algorithms, described in detail in the next section, make use of the above key technologies.

40.3 MPEG-1/Audio Coding

The MPEG-1/Audio coding standard [8], [28]–[30] is about to become a universal standard in many application areas with totally different requirements in the fields of consumer electronics, professional audio processing, telecommunications, and broadcasting [31]. The standard combines features of *MUSICAM* and *ASPEC coding algotithms* [32, 33]. Main steps of development towards the MPEG-1/Audio standard have been described in [30, 34]. The MPEG-1/Audio standard represents the state of the art in audio coding. Its subjective quality is equivalent to CD quality (16-bit PCM) at stereo rates given in Table 40.3 for many types of music. Because of its high dynamic range, MPEG-1/audio has potential to exceed the quality of a CD [31, 35].

TABLE 40.3 Approximate MPEG-1 Bit Rates for Transparent Representations of Audio Signals and Corresponding Compression Factors (Compared to CD Bit Rate)

MPEG-1 audio coding	Approximate stereo bit rates for transparent quality	Compression factor
Layer I	384 kb/s	4
Layer II	192 kb/s	8
Layer III	128 kb/s[a]	12

[a] Average bit rate; variable bit rate coding assumed.

40.3.1 The Basics

Structure

The basic structure follows that of perception-based coders (see Fig. 40.4). In the first step, the audio signal is converted into spectral components via an analysis filterbank; layers I and II make use of a subband filterbank, layer III employs a hybrid filterbank. Each spectral component is quantized and coded with the goal to keep the quantization noise below the masking threshold. The number of bits for each subband and a scalefactor are determined on a block-by-block basis, each block has 12 (layer I) or 36 (layers II and III) subband samples (see Section 40.2). The number of quantizer bits is obtained from a dynamic bit allocation algorithm (layers I and II) that is controlled by a *psychoacoustic model* (see below). The subband codewords, scalefactor, and bit allocation information are multiplexed into one bitstream, together with a header and optional ancillary data. In the decoder, the synthesis filterbank reconstructs a block of 32 audio output samples from the demultiplexed bitstream.

MPEG-1/Audio supports sampling rates of 32, 44.1, and 48 kHz and bit rates between 32 kb/s (mono) and 448 kb/s, 384 kb/s, and 320 kb/s (stereo; layers I, II, and III, respectively). Lower sampling rates (16, 22.05, and 24 kHz) have been defined in MPEG-2 for better audio quality at bit rates at, or below, 64 kb/s per channel [9]. The corresponding maximum audio bandwidths are 7.5, 10.3, and 11.25 kHz. The syntax, semantics, and coding techniques of MPEG-1 are maintained except for a small number of parameters.

Layers and Operating Modes

The standard consists of three layers I, II, and III of increasing complexity, delay, and subjective performance. From a hardware and software standpoint, the higher layers incorporate the main building blocks of the lower layers (Fig. 40.7). A standard *full MPEG-1/Audio decoder* is able to decode bit streams of all three layers. The standard also supports MPEG-1/Audio *layer X decoders* (X = I, II, or III). Usually, a layer II decoder will be able to decode bitstreams of layers I and II, a layer III decoder will be able to decode bitstreams of all three layers.

Stereo Redundancy Coding

MPEG-1/Audio supports four *modes:* mono, stereo, dual with two separate channels (useful for bilingual programs), and joint stereo. In the optimal joint stereo mode, interchannel dependencies are exploited to reduce the overall bit rate by using an irrelevancy reducing technique called *intensity stereo*. It is known that above 2 kHz and within each critical band, the human auditory system bases its perception of stereo imaging more on the temporal envelope of the audio than on its temporal fine structure. Therefore, the MPEG audio compression algorithm supports a stereo redundancy

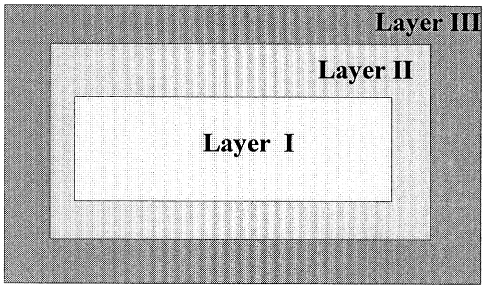

FIGURE 40.7: Hierarchy of layers I, II, and III of MPEG-1/Audio.

coding mode called *intensity stereo coding* which reduces the total bit rate without violating the spatial integrity of the stereophonic signal.

In intensity stereo mode, the encoder codes some upper-frequency subband outputs with a single sum signal L + R (or some linear combination thereof) instead of sending independent left (L) and right (R) subband signals. The decoder reconstructs the left and right channels based only on the single L + R signal and on independent left and right channel scalefactors. Hence, the spectral shape of the left and right outputs is the same within each intensity-coded subband but the magnitudes are different [36]. The optional joint stereo mode will only be effective if the required bit rate exceeds the available bit rate, and it will only be applied to subbands corresponding to frequencies of around 2 kHz and above.

Layer III has an additional option: in the mono/stereo (M/S) mode the left and right channel signals are encoded as middle (L + R) and side (L − R) channels. This latter mode can be combined with the joint stereo mode.

Psychoacoustic Models

We have already mentioned that the adaptive bit allocation algorithm is controlled by a psychoacoustic model. This model computes SMR taking into a account the short-term spectrum of the audio block to be coded and knowledge about noise masking. The model is only needed in the encoder which makes the decoder less complex; this asymmetry is a desirable feature for audio playback and audio broadcasting applications.

The normative part of the standard describes the decoder and the meaning of the encoded bitstream, but the encoder is not standardized thus leaving room for an evolutionary improvement of the encoder. In particular, *different psychoacoustic models can be used* ranging from very simple (or none at all) to very complex ones based on quality and implementability requirements. Information about the short-term spectrum can be derived in various ways, for example, as an accurate estimate from an FFT-based spectral analysis of the audio input samples or, less accurate, directly from the spectral components as in the conventional ATC [15]; see also Fig. 40.6. Encoders can also be optimized for a certain application. All these encoders can be used with complete compatibility with all existing MPEG-1/Audio decoders.

The informative part of the standard gives two examples of FFT-based models; see also [8, 30, 37]. Both models identify, in different ways, tonal and non-tonal spectral components and use the corresponding results of tone-masks-noise and noise-masks-tone experiments in the calculation of the global masking thresholds. Details are given in the standard, experimental results for both psychoacoustic models are described in [37]. In the informative part of the standard a 512-point FFT is proposed for layer I, and a 1024-point FFT for layers II and III. In both models, the audio input samples are Hann-weighted. *Model 1*, which may be used for layers I and II, computes for each masker its individual masking threshold, taking into account its frequency position, power, and tonality information. The global masking threshold is obtained as the sum of all individual masking thresholds and the absolute masking threshold. The SMR is then the ratio of the maximum signal level within a given subband and the minimum value of the global masking threshold in that given subband (see Fig. 40.2).

Model 2, which may be used for all layers, is more complex: tonality is assumed when a simple prediction indicates a high prediction gain, the masking thresholds are calculated in the cochlea domain, i.e., properties of the inner ear are taken into account in more detail, and, finally, in case of potential pre-echoes the global masking threshold is adjusted appropriately.

40.3.2 Layers I and II

MPEG layer I and II coders have very similar structures. The layer II coder achieves a better performance, mainly because the overall scalefactor side information is reduced exploiting redundancies between the scalefactors. Additionally, a slightly finer quantization is provided.

Filterbank

Layer I and II coders map the digital audio input into 32 subbands via equally spaced bandpass filters (Figs. 40.8 and 40.9). A polyphase filter structure is used for the frequency mapping; its filters have 512 coefficients. Polyphase structures are computationally very efficient because a DCT can be used in the filtering process, and they are of moderate complexity and low delay. On the negative side, the filters are equally spaced, and therefore the frequency bands do not correspond well to the critical band partition (see Section 40.2.1). At 48-kHz sampling rate, each band has a width of 24000/32 = 750 Hz; hence, at low frequencies, a single subband covers a number of adjacent critical bands. The subband signals are resampled (critically decimated) at a rate of 1500 Hz. The impulse response of subband k, $h_{\mathrm{sub}(k)}(n)$, is obtained by multiplication of the impulse response of a single *prototype lowpass filter*, $h(n)$, by a modulating function which shifts the lowpass response to the appropriate subband frequency range:

$$\begin{aligned} h_{\mathrm{sub}(k)}(n) &= h(n)\cos\left[\frac{(2k+1)\pi n}{2M} + \varphi(k)\right] ; \\ & \quad M = 32\ ;\ k = 0, 1, \ldots, 31\ ;\ n = 0, 1, \ldots, 511 \end{aligned}$$

The prototype lowpass filter has a 3-dB bandwidth of $750/2 = 375$ Hz, and the center frequencies are at odd multiples thereof (all values at 48 kHz sampling rate). The subsampled filter outputs exhibit a significant overlap. However, the design of the prototype filter and the inclusion of appropriate phase shifts in the cosine terms result in an aliasing cancellation at the output of the decoder synthesis filterbank. Details about the coefficients of the prototype filter and the phase shifts $\varphi(k)$ are given in the ISO/MPEG standard. Details about an efficient implementation of the filterbank can be found in [16] and [37], and, again, in the standardization documents.

Quantization

The number of quantizer levels for each spectral component is obtained from a dynamic bit allocation rule that is controlled by a psychoacoustic model. The bit allocation algorithm selects one uniform midtread quantizer out of a set of available quantizers such that both the bit rate requirement and the masking requirement are met. The iterative procedure minimizes the NMR in each subband. It starts with the number of bits for the samples and scalefactors set to zero. In each iteration step, the quantizer SNR(m) is increased for the one subband quantizer producing the largest value of the NMR at the quantizer output. (The increase is obtained by allocating one more bit). For that purpose, NMR(m) = SMR − SNR(m) is calculated as the difference (in dB) between the actual quantization

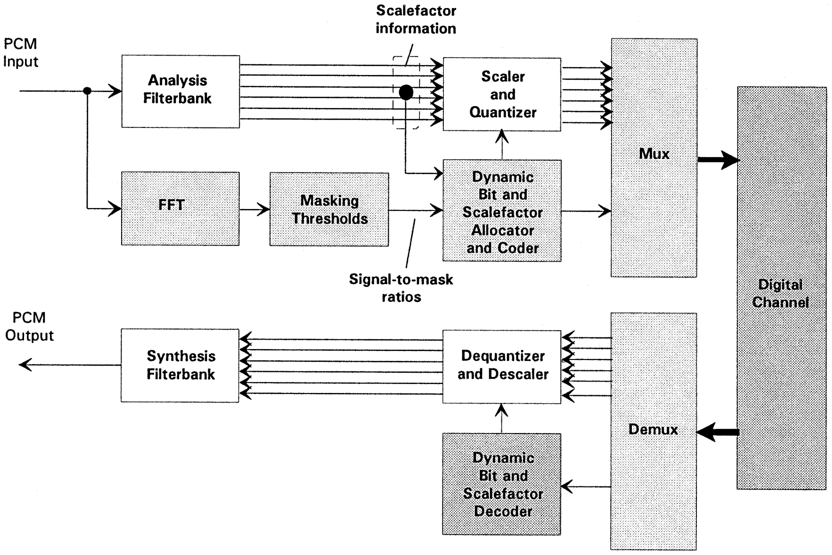

FIGURE 40.8: Structure of MPEG-1/Audio encoder and decoder, layers I and II.

noise level and the minimum global masking threshold. The standard provides tables with estimates for the quantizer SNR(m) for a given m.

Block companding is used in the quantization process, i.e., blocks of decimated samples are formed and divided by a *scalefactor* such that the sample of largest magnitude is unity. In *layer I* blocks of 12 decimated and scaled samples are formed in each subband (and for the left and right channel) and there is one bit allocation for each block. At 48-kHz sampling rate, 12 subband samples correspond to 8 ms of audio. There are 32 blocks, each with 12 decimated samples, representing $32 \times 12 = 384$ audio samples.

In *layer II* in each subband a 36-sample *superblock* is formed of three consecutive blocks of 12 decimated samples corresponding to 24 ms of audio at 48 kHz sampling rate. There is one bit allocation for each 36-sample superblock. All 32 superblocks, each with 36 decimated samples, represent, altogether, $32 \times 36 = 1152$ audio samples. As in layer I, a scalefactor is computed for each 12-sample block. A redundancy reduction technique is used for the transmission of the scalefactors: depending on the significance of the changes between the three consecutive scalefactors, one, two, or all three scalefactors are transmitted, together with a 2-bit *scalefactor select information*. Compared with layer I, the bit rate for the scalefactors is reduced by around 50% [30]. Figure 40.9 indicates the block companding structure.

The scaled and quantized spectral subband components are transmitted to the receiver together with scalefactor, scalefactor select (layer II), and bit allocation information. Quantization with block companding provides a very large dynamic range of more than 120 dB. For example, in layer II uniform midtread quantizers are available with 3, 5, 7, 9, 15, 31, . . . , 65535 levels for subbands of low index (low frequencies). In the mid and high frequency region, the number of levels is reduced significantly. For subbands of index 23 to 26 there are only quantizers with 3, 5, and 65535 (!) levels available. The 16-bit quantizers prevent overload effects. Subbands of index 27 to 31 are not transmitted at all. In order to reduce the bit rate, the codewords of three successive subband samples resulting from quantizing with 3-, 5, and 9-step quantizers are assigned one common codeword. The savings in bit rate is about 40% [30].

Figure 40.10 shows the time-dependence of the assigned number of quantizer bits in all subbands

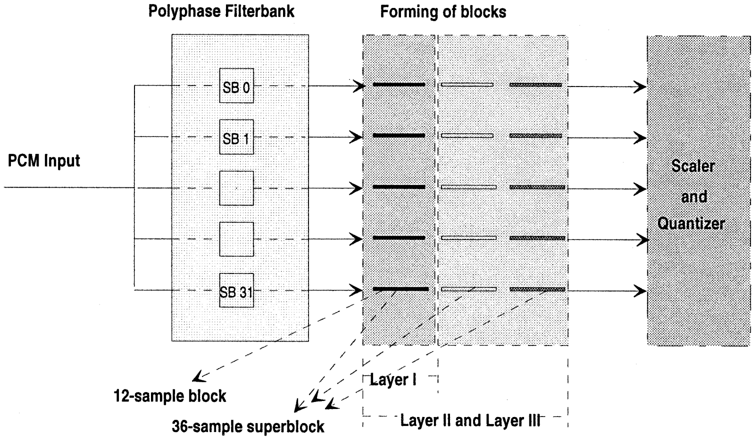

FIGURE 40.9: Block companding in MPEG-1/Audio coders.

for a layer II encoded high quality speech signal. Note, for example, that quantizers with ten or more bits resolution are only employed in the lowest subbands, and that no bits have been assigned for frequencies above 18 kHz (subbands of index 24 to 31).

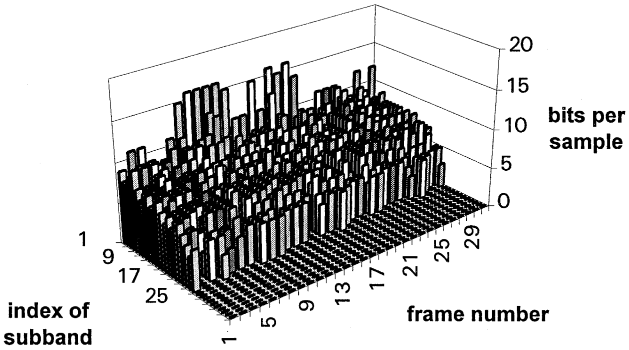

FIGURE 40.10: Time-dependence of assigned number of quantizer bits in all subbands for a layer II encoded high quality speech signal.

Decoding

The decoding is straightforward: the subband sequences are reconstructed on the basis of blocks of 12 subband samples taking into account the decoded scalefactor and bit allocation information. If a subband has no bits allocated to it, the samples in that subband are set to zero. Each time the subband samples of all 32 subbands have been calculated, they are applied to the *synthesis filterbank*, and 32 consecutive 16-bit PCM format audio samples are calculated. If available, as in bidirectional communications or in recorder systems, the encoder (analysis) filterbank can be used in a reverse mode in the decoding process.

40.3.3 Layer III

Layer III of the MPEG-1/Audio coding standard introduces many new features (see Fig. 40.11), in particular a switched hybrid filterbank. In addition, it employs an analysis-by-synthesis approach, an advanced pre-echo control, and nonuniform quantization with entropy coding. A buffer technique, called *bit reservoir*, leads to further savings in bit rate. Layer III is the only layer that provides mandatory decoder support for variable bit rate coding [38].

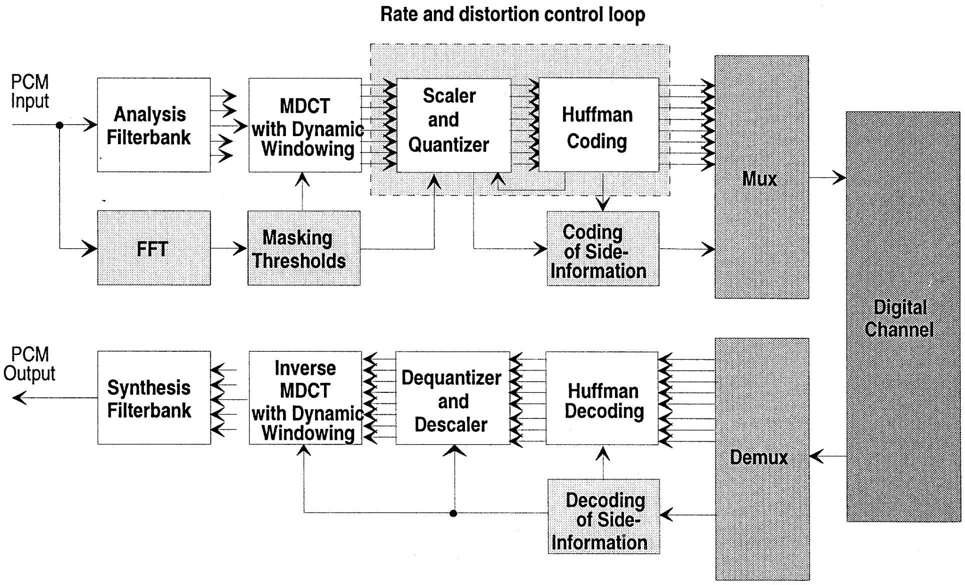

FIGURE 40.11: Structure of MPEG-1/Audio encoder and decoder, layer III.

Switched Hybrid Filterbank

In order to achieve a higher frequency resolution closer to critical band partitions, the 32 subband signals are subdivided further in frequency content by applying, to each of the subbands, a 6- or 18-point modified DCT block transform, with 50% overlap; hence, the windows contain, respectively, 12 or 36 subband samples. The maximum number of frequency components is $32 \times 18 = 576$ each representing a bandwidth of only $24000/576 = 41.67$ Hz. Because the 18-point block transform provides better frequency resolution, it is normally applied, whereas the 6-point block transform provides better time resolution and is applied in case of expected pre-echoes (see Section 40.2.3). In principle, a pre-echo is assumed, when an instantaneous demand for a high number of bits occurs. Depending on the nature of potential, all pre-echoes or a smaller number of transforms are switched. Two special MDCT windows, a start window and a stop window, are needed in case of transitions between short and long blocks and vice versa to maintain the time domain alias cancellation feature of the MDCT [22, 25, 37]. Figure 40.12 shows a typical sequence of windows.

Quantization and Coding

The MDCT output samples are nonuniformly quantized thus providing both smaller mean-squared errors and masking because larger errors can be tolerated if the samples to be quantized

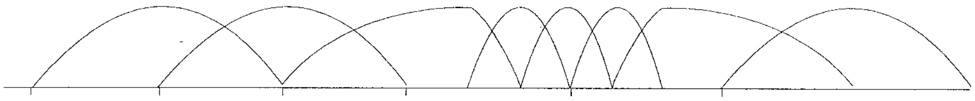

FIGURE 40.12: Typical sequence of windows in adaptive window switching.

are large. Huffman coding, based on 32 code tables, and additional run-length coding are applied to represent the quantizer indices in an efficient way. The encoder maps the variable wordlength codewords of the Huffman code tables into a constant bit rate by monitoring the state of a bit reservoir. The bit reservoir ensures that the decoder buffer neither underflows nor overflows when the bitstream is presented to the decoder at a constant rate.

In order to keep the quantization noise in all critical bands below the global masking threshold (noise allocation) an *iterative analysis-by-synthesis method* is employed whereby the process of scaling, quantization, and coding of spectral data is carried out within two nested iteration loops. The decoding follows that of the encoding process.

40.3.4 Frame and Multiplex Structure

Frame Structure

Figure 40.13 shows the frame structure of MPEG-1/Audio coded signals, both for layer I and layer II. Each frame has a header; its first part contains 12 synchronisation bits, 20 bit system information, and an optional 16-bit cyclic redundancy check code. Its second part contains side information about the bit allocation and the scalefactors (and, in layer II, scalefactor information). As main information, a frame carries a total of 32 × 12 subband samples (corresponding to 384 PCM audio input sample — equivalent to 8 ms at a sampling rate of 48 kHz) in layer I, and a total of 32 × 36 subband samples in layer II (corresponding to 1152 PCM audio input samples — equivalent to 24 ms at a sampling rate of 48 kHz). Note that the layer I and II frames are autonomous: each frame contains all information necessary for decoding. Therefore, each frame can be decoded independently from previous frames, it defines an entry point for audio storage and audio editing applications. Please note that the lengths of the frames are not fixed, due to (1) the length of the main information field, which depends on bit-rate and sampling frequency, (2) the side information field which varies in layer II, and (3) the ancillary data field, the length of which is not specified.

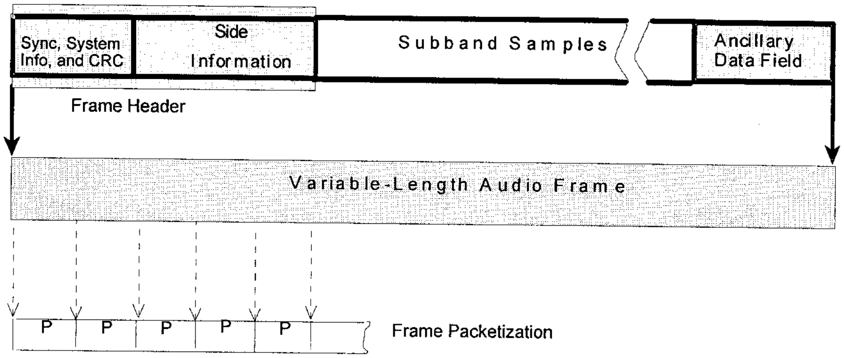

FIGURE 40.13: MPEG-1 frame structure and packetization. Layer I: 384 subband samples; layer II: 1152 subband samples; packets P: 4-byte header; 184-byte payload field (see also Fig. 40.14).

Multiplex Structure

We have already mentioned that the systems part of the MPEG-1 coding standard IS 11172 defines a *packet structure for multiplexing* audio, video, and ancillary data bitstreams in one stream. The variable-length MPEG frames are broken down into packets. The packet structure uses 188-byte packets consisting of a 4-byte header followed by 184 bytes of payload (see Fig. 40.14). The header includes a sync byte, a 13-bit field called packet identifier to inform the decoder about the type of data, and additional information. For example, a *1-bit payload unit start indicator* indicates if the payload starts with a frame header. No predetermined mix of audio, video, and ancillary data bitstreams is required, the mix may change dynamically, and services can be provided in a very flexible way. If additional header information is required, such as for periodic synchronization of audio and video timing, a variable-length *adaptation header* can be used as part of the 184-byte payload field.

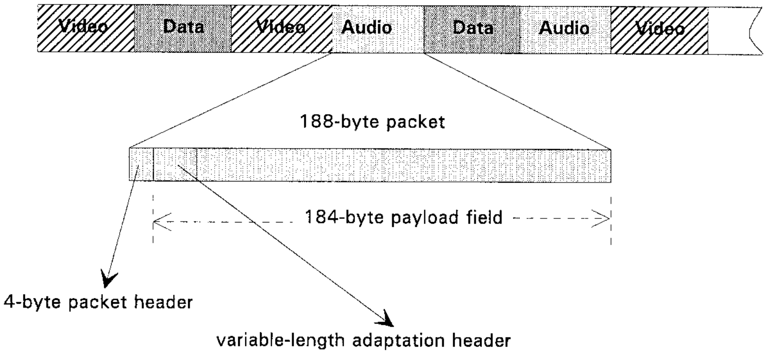

FIGURE 40.14: MPEG packet delivery.

Although the lengths of the frames are not fixed, the interval between frame headers is constant (within a byte) throughout the use of padding bytes. The MPEG systems specification describes how MPEG-compressed audio and video data streams are to be multiplexed together to form a single data stream. The terminology and the fundamental principles of the systems layer are described in [39].

40.3.5 Subjective Quality

The standardization process included extensive subjective tests and objective evaluations of parameters such as complexity and overall delay. The MPEG (and equivalent ITU-R) listening tests were carried out under very similar and carefully defined conditions with around 60 experienced listeners, approximately 10 test sequences were used, and the sessions were performed in stereo with both loudspeakers and headphones. In order to detect even small impairments, the 5-point ITU-R impairment scale was used in all experiments. Details are given in [40] and [41]. *Critical test items were chosen in the tests to evaluate the coders by their worst case (not average) performance*. The subjective evaluations, which have been based on triple stimulus/hidden reference/double blind tests, have shown very similar and stable evaluation results. In these tests the subject is offered three signals, A,B, and C (triple stimulus). A is always the unprocessed source signal (the reference). B and C, or C and B, are the reference and the system under test (hidden reference). The selection is neither known to the subjects nor to the conductors(s) of the test (double blind test). The subjects have to decide if B or C is the reference and have to grade the remaining one.

The MPEG-1/Audio coding standard has shown an excellent performance for all layers at the rates given in Table 40.3. It should be mentioned again that the standard leaves room for encoder-based improvements by using better psychoacoustic models. Indeed, many improvements have been achieved since the first subjective results had been carried out in 1991.

40.4 MPEG-2/Audio Multichannel Coding

A logical further step in digital audio is the definition of a multichannel audio representation system to create a convincing, lifelike soundfield both for audio-only applications and for audiovisual systems, including video conferencing, videophony, multimedia services, and electronic cinema. Multichannel systems can also provide multilingual channels and additional channels for visually impaired (a verbal description of the visual scene) and for hearing impaired (dialog with enhanced intelligibility). ITU-R has recommended a five-channel loudspeaker configuration, referred to as 3/2-stereo, with a left and a right channel (L and R), an additional center channel C, two side/rear surround channels (LS and RS) augmenting the L and R channels, see Fig. 40.15 [ITU-R Rec. 775]. Such a configuration offers an improved realism of auditory ambience with a stable frontal sound image and a large listening area.

Multichannel digital audio systems support p/q presentations with p front and q back channels, and also provide the possibilities of transmitting two independent stereophonic programs and/or a number of commentary or multilingual channels. Typical combinations of channels include.

• 1 channel	1/0-configuration:	centre (mono)
• 2 channels	2/0-configuration:	left, right (stereophonic)
• 3 channels	3/0-configuration:	left, right, centre
• 4 channels:	3/1-configuration	left, right, centre, mono-surround
• 5 channels:	3/2-configuration:	left, right, centre, surround left, surround right

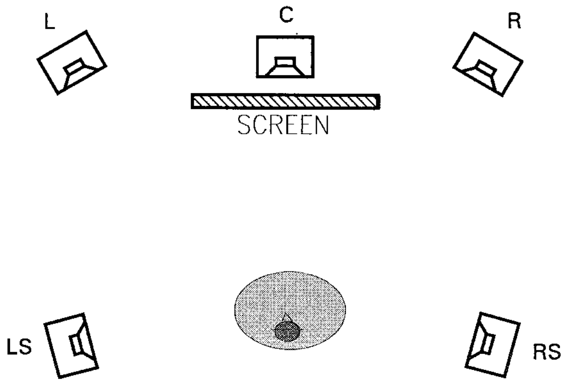

FIGURE 40.15: 3/2 Multichannel loudspeaker configuration.

ITU-R Recommendation 775 provides a set of downward mixing equations if the number of loudspeakers is to be reduced (*downward compatibility*). An additional *low frequency enhancement (LFE-or subwoofer-) channel* is particularly useful for HDTV applications, it can be added, optionally, to any of the configurations. The LFE channel extends the low frequency content between 15 and 120 Hz in terms of both frequency and level.

One or more loudspeakers can be positioned freely in the listening room to reproduce this LFE signal. (Film industry uses a similar system for their digital sound systems).[1]

In order to reduce the overall bit rate of multichannel audio coding systems, redundancies and irrelevancy, such as interchannel dependencies and interchannel masking effects, respectively, may be exploited. In addition, stereophonic-irrelevant components of the multichannel signal, which do not contribute to the localization of sound sources, may be identified and reproduced in a monophonic format to further reduce bit rates. State-of-the-art multichannel coding algorithms make use of such effects. A careful design is needed, otherwise such joint coding may produce artifacts.

40.4.1 MPEG-2/Audio Multichannel Coding

The second phase of MPEG, labeled MPEG-2, includes in its audio part two multichannel audio coding standards, one of which is forward- and backward-compatible with MPEG-1/Audio [8], [42]–[45]. *Forward compatibility* means that an MPEG-2 multichannel decoder is able to properly decode MPEG-1 mono or stereophonic signals, *backward compatibility* (BC) means that existing MPEG-1 stereo decoders, which only handle two-channel audio, is able to reproduce a meaningful basic 2/0 stereo signal from a MPEG-2 multichannel bit stream so as to serve the need of users with simple mono or stereo equipment. *Non-backward compatible* (NBC) multichannel coders will not be able to feed a meaningful bit stream into a MPEG-1 stereo decoder. On the other hand, NBC codecs have more freedom in producing a high quality reproduction of audio signals.

With backward compatibility, it is possible to introduce multichannel audio at any time in a smooth way without making existing two-channel stereo decoders obsolete. An important example is the European Digital Audio Broadcast system, which will require MPEG-1 stereo decoders in the first generation but may offer multichannel audio at a later point.

40.4.2 Backward-Compatible (BC) MPEG-2/Audio Coding

BC implies the use of compatibility matrices. A down-mix of the five channels (“matrixing”) delivers a correct basic 2/0 stereo signal, consisting of a left and a right channel, LO and RO, respectively. A typical set of equations is

$$\begin{aligned} LO &= \alpha\ (L + \beta^{\cdot} C + \delta^{\cdot} LS) \\ RO &= \alpha\ (R + \beta^{\cdot} C + \delta^{\cdot} RS) \end{aligned} \qquad \alpha = \tfrac{1}{1+\sqrt{2}}\ ;\ \beta = \delta = \sqrt{2}$$

Other choices are possible, including $LO = L$ and $RO = R$. The factors α, β, and δ attenuate the signals to avoid overload when calculating the compatible stereo signal (LO, RO). The signals LO and RO are transmitted in MPEG-1 format in transmission channels $T1$ and $T2$. Channels $T3$, $T4$, and $T5$ together form the *multichannel extension signal* (Fig. 40.16). They have to be chosen such that the decoder can recompute the complete 3/2-stereo multichannel signal. Interchannel redundancies and masking effects are taken into account to find the best choice. A simple example is $T3 = C$, $T4 = LS$, and $T5 = RS$. In MPEG-2 the matrixing can be done in a very flexible and even time-dependent way.

BC is achieved by transmitting the channels LO and RO in the subband-sample section of the MPEG-1 audio frame and all multichannel extension signals $T3$, $T4$, and $T5$ in the first part of the MPEG-1/Audio frame reserved for ancillary data. This ancillary data field is ignored by MPEG-1

[1] A 3/2-configuration with five high-quality full-range channels plus a subwoofer channel is often called a 5.1 system.

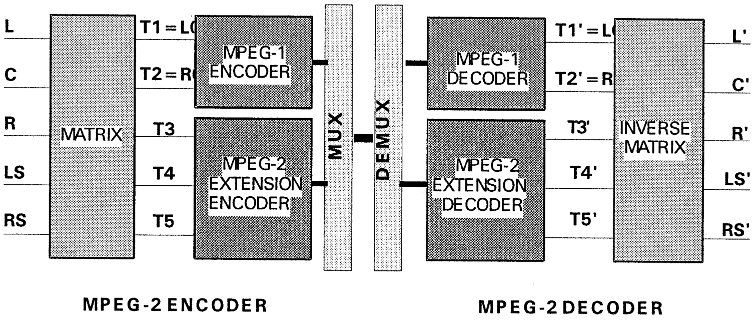

FIGURE 40.16: Compatibility of MPEG-2 multichannel audio bit streams.

decoders (see Fig. 40.17). The length of the ancillary data field is not specified in the standard. If the decoder is of type MPEG-1, it uses the 2/0-format front left and right down-mix signals, LO' and RO', directly (see Fig. 40.18). If the decoder is of type MPEG-2, it recomputes the complete 3/2-stereo multichannel signal with its components L', R', C', LS', and RS' via "dematrixing" of LO', RO', $T3'$, $T4'$, and $T5'$ (see Fig. 40.16).

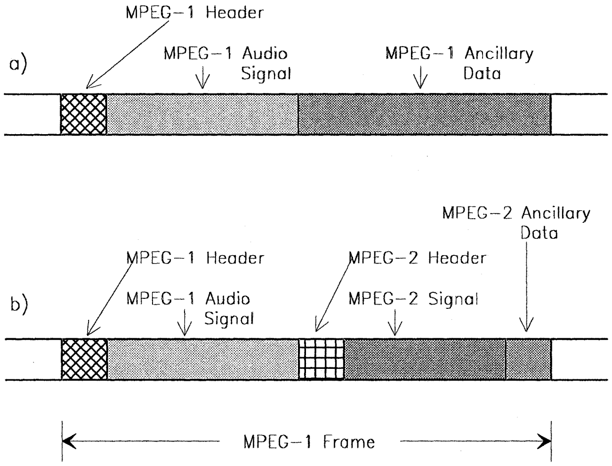

FIGURE 40.17: Data format of MPEG audio bit streams. a.) MPEG-1 audio frame; b.) MPEG-2 audio frame, compatible with MPEG-1 format.

Matrixing is obviously necessary to provide BC; however, if used in connection with perceptual coding, "unmasking" of quantization noise may appear [46]. It may be caused in the dematrixing process when sum and difference signals are formed. In certain situations, such a masking sum or difference signal component can disappear in a specific channel. Since this component was supposed to mask the quantization noise in that channel, this noise may become audible. Note that the masking signal will still be present in the multichannel representation but it will appear on a different loudspeaker. Measures against "unmasking" effects have been described in [47].

MPEG-1 decoders have a bit rate limitation (384 kb/s in layer II). In order to overcome this limitation, the MPEG-2 standard allows for a second bit stream, the extension part, to provide

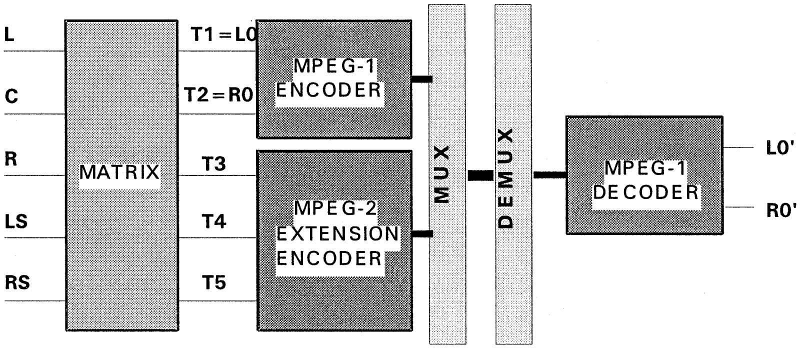

FIGURE 40.18: MPEG-1 stereo decoding of MPEG-2 multichannel bit stream.

compatible multichannel audio at higher rates. Figure 40.19 shows the structure of the bit stream with extension.

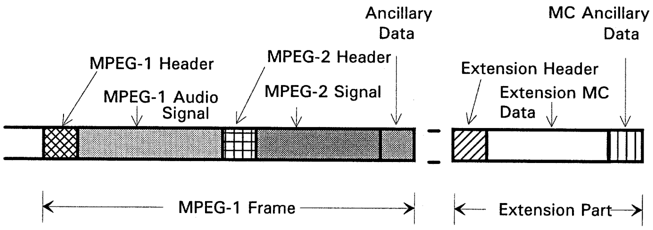

FIGURE 40.19: Data format of MPEG-2 audio bit stream with extension part.

40.4.3 Advanced/MPEG-2/Audio Coding (AAC)

A second standard within MPEG-2 supports applications that do not request compatibility with the existing MPEG-1 stereo format. Therefore, matrixing and dematrixing are not necessary and the corresponding potential artifacts disappear (see Fig. 40.20). The advanced multichannel coding mode will have the sampling rates, audio bandwidth, and channel configurations of MPEG-2/Audio, but shall be capable of operating at bit rates from 32kb/s up to a bit rate sufficient for high quality audio.

The last two years have seen extensive activities to optimize and standardize a MPEG-2 AAC algorithm. Many companies around the world contributed advanced audio coding algorithms in a collaborative effort to come up with a flexible high quality coding standard [44]. The MPEG-2 AAC standard employs high resolution filter banks, prediction techniques, and Huffman coding.

Modules

The MPEG-2 AAC standard is based on recent evaluations and definitions of basic modules each having been selected from a number of proposals. The self-contained modules include:

- optional preprocessing
- time-to-frequency mapping (filterbank)

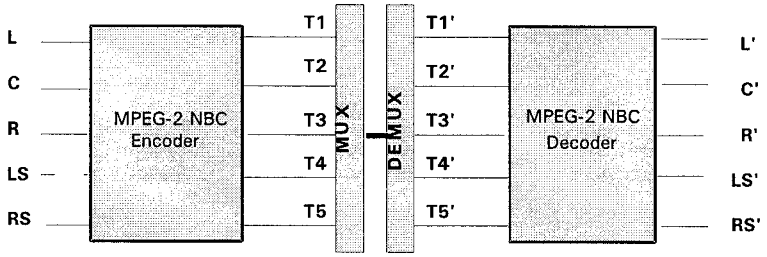

FIGURE 40.20: Non-backward-compatible MPEG-2 multichannel audio coding (advanced audio coding).

- psychoacoustic modeling
- prediction
- quantization and coding
- noiseless coding
- bit stream formatter

Profiles

In order to serve different needs, the standard will offer three profiles:

1. main profile
2. low complexity profile
3. sampling-rate-scaleable profile

For example, in its main profile, the filter bank is a modified discrete cosine transform of blocklength 2048 or 256, it allows for a frequency resolution of 23.43 Hz and a time resolution of 2.6 ms (both at a sampling rate of 48 kHz). In the case of the long blocklength, the window shape can vary dynamically as a function of the signal; a temporal noise shaping tool is offered to control the time dependence of the quantization noise; time domain prediction with second order backward-adaptive linear predictors reduces the bit rate for coding subsequent subband samples in a given subband; iterative non-uniform quantization and noiseless coding are applied.

The low complexity profile does not employ temporal noise shaping and time domain prediction, whereas in the sampling-rate-scaleable profile a preprocessing module is added that allows for samplig rates of 6, 12, 18, and 24 kHz. The default configurations of MPEG-2 AAC include 1.0, 2.0, and 5.1 (mono, stereo, and five channel with LFE-channel). However, 16 configurations can be defined in the encoder. A detailed description of the MPEG-2 AAC multichannel standard can be found in the literature [44].

The above listed selected modules define the MPEG-2/AAC standard which became International Standard in April 1997 as an extension to MPEG-2 (*ISO/MPEG 13818 - 7*). The standard offers high quality at lowest possible bit rates between 320 and 384 kb/s for five channels, it will find many applications, both for consumer and professional use.

40.4.4 Simulcast Transmission

If bit rates are not of high concern, a *simulcast transmission* may be employed where a full MPEG-1 bitstream is multiplexed with the full MPEG-2 AAC bit stream in order to support BC without matrixing techniques (Fig. 40.21).

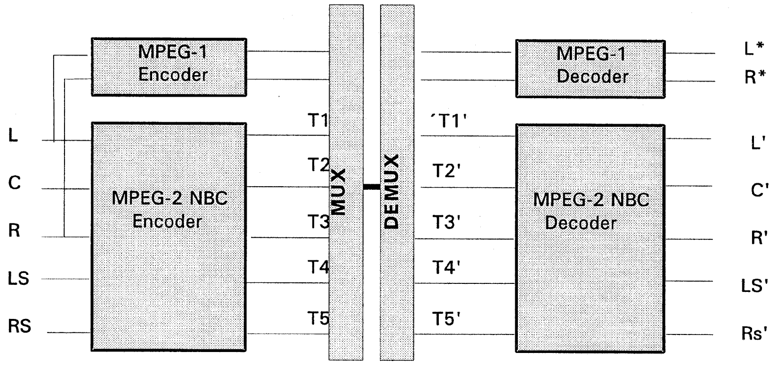

FIGURE 40.21: BC MPEG-2 multichannel audio coding (simulcast mode).

40.4.5 Subjective Tests

First subjective tests, independently run at German Telekom and BBC (UK) under the umbrella of the MPEG-2 standardization process had shown a satisfactory average performance of NBC and BC coders. The tests had been carried out with experienced listeners and critical test items at low bit rates (320 and 384 kb/s). However, all codecs showed deviations from transparency for some of the test items [48, 49]. Very recently [50], extensive formal subjective tests have been carried out to compare MPEG-2 AAC coders, operating, respectively, at 256 and 320 kb/s, and a BC MPEG-2 layer II coder,[2] operating at 640 kb/s. All coders showed a very good performance, with a slight advantage of the 320 kb/s MPEG-2 AAC coder compared with the 640 kb/s MPEG-2 layer II BC coder. The performances of those coders are indistinguishable from the original in the sense of the EBU definition of *indistinguishable quality* [51].

40.5 MPEG-4/Audio Coding

Activities within MPEG-4 aim at proposals for a broad field of applications including multimedia. MPEG-4 will offer higher compression rates, and it will merge the whole range of audio from high fidelity audio coding and speech coding down to synthetic speech and synthetic audio. In order to represent, integrate, and exchange pieces of audio-visual information, MPEG-4 offers standard tools which can be combined to satisfy specific user requirements [52]. A number of such configurations may be standardized. A syntactic description will be used to convey to a decoder the choice of tools made by the encoder. This description can also be used to describe new algorithms and download their configuration to the decoding processor for execution. The current toolset supports audio and speech compression at monophonic bit rates ranging from 2 to 64 kb/s. Three *core coders* are used:

1. a parametric coding scheme for low bit rate speech coding
2. an analysis-by-synthesis coding scheme for medium bit rates (6 to 16 kb/s)
3. a subband/transform-based coding scheme for higher bit rates.

These three coding schemes have been integrated into a so-called verification model that describes the operations both of encoders and decoders, and that is used to carry out simulations and optimizations.

[2] A 1995 version of this latter coder was used, therefore its test results do not reflect any subsequent enhancements.

In the end, the verification model will be the embodiment of the standard [52]. Let us also note that MPEG-4 will offer new functionalities such as time scale changes, pitch control, edibility, database access, and scalibility, which allows extraction from the transmitted bitstream of a subset sufficient to generate audio signals with lower bandwidth and/or lower quality depending on channel capacity or decoder complexity. MPEG-4 will become an international standard in November 1998.

40.6 Applications

MPEG/Audio compression technologies will play an important role in consumer electronics, professional audio, telecommunications, broadcasting, and multimedia. A few, but typical application fields are described in the following.

Main applications will be based on delivering digital audio signals over terrestrial and satellite-based digital *broadcast and transmission systems* such as subscriber lines, program exchange links, cellular mobile radio networks, cable-TV networks, local area networks, etc. [53]. For example, in narrowband *Integrated Services Digital Networks* (ISDN) customers have physical access to one or two 64-kb/s B channels and one 16-kb/s D channel (which supports signaling but can also carry user information). Other configurations are possible including $p \times 64$ kb/s ($p = 1, 2, 3, \ldots$) services. ISDN rates offer useful channels for a practical distribution of stereophonic and multichannel audio signals.

Because ISDN is a bidirectional service, it also provides upstream paths for future on-demand and interactive audiovisual *just-in-time audio* services. The backbone of digital telecommunication networks will be broadband (B-) ISDN with its cell-oriented structure. Cell delays and cell losses are sources of distortions to be taken into account in designs of digital audio systems [54].

Lower bit rates than those given by the 16-bit PCM format are mandatory if audio signals are to be stored efficiently on *storage media*—although the upcoming digital versatile disk (DVD) with its capacity of 4.7 GB relieves the pressure for extreme compression factors. In the field of digital storage on digital audio tape and (re-writeable) disks, a number of MPEG-based consumer products have recently reached the audio market. Of these products, Philips *Digital Compact Cassette* (DCC) essentially makes use of layer I of the MPEG-1/Audio coder employing its 384 kb/s stereo rate; its audio coding algorithm is called PASC (*Precision Audio Subband Coding*) [16]. The DCC encoder obtains an estimate of the short-term spectrum directly from the 32 subbands.

In the movie theater world, a 7.1-channel configuration is becoming popular due to an improved front-back stability of the stereo image and an improved impression of spaciousness. A scalable 7.1-channel reproduction is applied in the digital video disc (DVD). It is based on the MPEG-1 and MPEG-2 standards by down-mixing the 7-channel signal into a 5-channel signal, and a subsequent down-mixing of the latter one into a 2-channel signal [55]. The 2-channel signal, three contributions from the 5-channel signal, and two contributions from the 7-channel signal can then be transmitted or stored. The decoder uses the 2-channel signal directly, or it employs matrixing to reconstruct 5- or 7-channel signals. Other formats are possible, such as storing a 5-channel signal and an additional stereo signal in simulcast mode, without down-mixing the stereo signal from the multichannel signal.

A further example is solid state audio playback systems (e.g., for announcements) with the compressed data stored on chip-based memory cards or smart cards. One example is NEC's prototype *Silicon Audio Player* which uses a one-chip MPEG-1/Audio layer II decoder and offers 24 min of stereo at its recommended stereo bit rate of 192 kb/s [56].

A number of decisions concerning the introduction of *digital audio broadcast* (DAB) and digital video broadcast (DVB) services have been made recently. In Europe, a project group named Eureka 147 has worked out a DAB system able to cope with the problems of digital broadcasting [57]–[59]. ITU-R has recommended the MPEG-1/Audio coding standard after it had made extensive subjective tests. Layer II of this standard is used for program emission, the Layer III version is recommended for commentary links at low rates. The sampling rate is 48 kHz in all cases, the ancillary data field is used for program associated data (PAD information). The DAB system has a significant bit rate overhead for error correction based on punctured convolutional codes in order to support *source-adapted channel coding*, i.e., an unequal error protection that is in accordance with the sensitivity of individual bits or a group of bits to channel errors [60]. Additionally, error concealment techniques will be applied to provide a *graceful degradation* in case of severe errors. In the U.S. a standard has not yet been defined. Simulcasting analog and digital versions of the same audio program in the FM terrestrial band (88 to 108 MHz) is an important issue (whereas the European solution is based on new channels) [61].

As examples of satellite-based digital broadcasting, we mention the Hughes DirecTV satellite subscription television system and ADR (Astra Digital Radio) both of which make use of MPEG-1 layer II. As a further example, the Eutelsat SaRa system will be based on layer III coding.

Advanced digital TV systems provide HDTV delivery to the public by terrestrial broadcasting and a variety of alternate media and offer full-motion high resolution video and high quality multichannel surround audio. The overall bit rate may be transmitted within the bandwidth of an analog UHF television channel. The U.S. *Grand Alliance HDTV* system and the European *Digital Video Broadcast (DVB)* system both make use of the MPEG-2 video compression system and of the MPEG-2 transport layer which uses a flexible ATM-like packet protocol with headers/descriptors for multiplexing audio and video bit streams in one stream with the necessary information to keep the streams synchronized when decoding. The systems differ in the way the audio signal is compressed: the Grand Alliance system will use Dolby's AC-3 transform coding technique [62]–[64], whereas the DVB system will use the MPEG-2/Audio algorithm.

40.7 Conclusions

Low bit rate digital audio is applied in many different fields, such as consumer electronics, professional audio processing, telecommunications, and broadcasting. Perceptual coding in the frequency domain has paved the way to high compression rates in audio coding. ISO/MPEG-1/Audio coding with its three layers has been widely accepted as an international standard. Software encoders, single DSP chip implementations, and computer extensions are available from a number of suppliers.

In the area of broadcasting and mobile radio systems, services are moving to portable and handheld devices, and new, third generation mobile communication networks are evolving. Coders for these networks must not only operate at low bit rates but must be stable in burst-error and packet- (cell-) loss environments. Error concealment techniques will play a significant role. Due to the lack of available bandwidth, traditional channel coding techniques may not be able to sufficiently improve the reliability of the channel.

MPEG/Audio coders are controlled by psychoacoustic models which may be improved thus leaving room for an evolutionary improvement of codecs. In the future, we will see new solutions for encoding. A better understanding of binaural perception and of stereo presentation will lead to new proposals.

Digital multichannel audio improves stereophonic images and will be of importance both for audio-only and multimedia applications. MPEG-2/audio offers both BC and NBC coding schemes to serve different needs. Ongoing research will result in enhanced multichannel representations by making better use of interchannel correlations and interchannel masking effects to bring the bit rates further down. We can also expect solutions for special presentations for people with impairments of hearing or vision which can make use of the multichannel configurations in various ways.

Emerging activities of the ISO/MPEG expert group aim at proposals for audio coding which will offer higher compression rates, and which will merge the whole range of audio from high fidelity audio coding and speech coding down to synthetic speech and synthetic audio (ISO/IEC MPEG-4).

Because the basic audio quality will be more important than compatibility with existing or upcoming standards, this activity will open the door for completely new solutions.

References

[1] Bruekers, A.A.M.L. et al., Lossless coding for DVD audio, 101th Audio Engineering Society Convention, Los Angeles, Preprint 4358, 1996.

[2] Jayant, N.S. and Noll, P., *Digital coding of waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[3] Spanias, A.S., Speech coding: A tutorial review, *Proc. IEEE,* 82(10), 1541–1582, Oct.94.

[4] Jayant, N.S., Johnston, J.D. and Shoham, Y., Coding of wideband speech, *Speech Commun.*, 11, 127–138, 1992.

[5] Gersho, A., Advances in speech and audio compression, *Proc. IEEE,* 82(6), 900–918, 1994.

[6] Noll, P., Wideband speech and audio coding, *IEEE Commun. Mag.*, 31(11), 34–44, 1993.

[7] Noll, P., Digital audio coding for visual communications, *Proc. IEEE,* 83(6), June 1995.

[8] ISO/IEC JTC1/SC29, Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s–IS 11172 (Part 3, Audio), 1992.

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

[10] ISO/MPEG, Doc. N0821, Proposal Package Description - Revision 1.0, Nov. 1994.

[11] WWW — official MPEG home page: address http://drogo.cselt.stet.it/mpeg/. Important link: http:/www.vol.it/MPEG/

[12] Hathaway, G.T., A NICAM digital stereophonic encoder, in *Audiovisual Telecommunications* Nigthingale, N.D. Ed., Chapman & Hall, 1992, 71 - 84.

[13] Zwicker, E. and Feldtkeller, R., *Das Ohr als Nachrichtenempfänger,* S. Hirzel Verlag, Stuttgart, 1967.

[14] Jayant, N.S., Johnston, J.D. and Safranek, R., Signal compression based on models of human perception, *Proc. IEEE,* 81(10), 1385–1422, 1993.

[15] Zelinski, R. and Noll, P., Adaptive transform coding of speech signals, *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, ASSP-25, 299–309, Aug. 1977.

[16] Hoogendorn, A., Digital compact cassette, *Proc. IEEE,* 82(10), 1479–1489, Oct. 1994.

[17] Noll, P., On predictive quantizing schemes, *Bell System Tech. J.,* 57, 1499–1532, 1978.

[18] Makhoul, J. and Berouti, M., Adaptive noise spectral shaping and entropy coding in predictive coding of speech. *IEEE Trans. on Acoustics, Speech, and Signal Processing,* 27(1), 63–73, Feb. 1979.

[19] Esteban, D. and Galand, C., Application of quadrature mirror filters to split band voice coding schemes, *Proc. ICASSP,* 191–195, 1987.

[20] Rothweiler, J.H., Polyphase quadrature filters, a new subband coding technique, *Proc. Intl. Conf. ICASSP'83,* 1280–1283, 1983.

[21] Princen, J. and Bradley, A., Analysis/synthesis filterbank design based on time domain aliasing cancellation, *IEEE Trans. on Acoust. Speech, and Signal Process.*, ASSP-34, 1153–1161, 1986.

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

[23] Yeoh, F.S. and Xydeas, C.S., Split-band coding of speech signals using a transform technique, *Proc. ICC,* 3, 1183–1187, 1984.

[24] Granzow, W., Noll, P. and Volmary, C., Frequency-domain coding of speech signals, (in German), NTG-Fachbericht No. 94, VDE-Verlag, Berlin, 150–155, 1986.

[25] Edler, B., Coding of audio signals with overlapping block transform and adaptive window functions, (in German), *Frequenz,* 43, 252–256, 1989.

[26] Iwadare, M., Sugiyama, A., Hazu, F., Hirano, A. and Nishitani, T., A 128 kb/s hi-fi audio CODEC based on adaptive transform coding with adaptive block size, *IEEE J. on Sel. Areas in Commun.*, 10(1), 138–144, Jan. 1992.

[27] Zelinski, R. and Noll, P., Adaptive Blockquantisierung von Sprachsignalen, Technical Report No. 181, Heinrich-Hertz-Institut für Nachrichtentechnik, Berlin, 1975.

[28] van der Waal, R.G., Brandenburg, K. and Stoll, G., Current and future standardization of high-quality digital audio coding in MPEG, *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1993.

[29] Noll, P. and Pan, D., ISO/MPEG audio coding, *Intl. J. High Speed Electronics and Systems,* 1997.

[30] Brandenburg, K. and Stoll, G., The ISO/MPEG-audio codec: A generic standard for coding of high quality digital audio, *J. Audio Eng. Soc. (AES)*, 42(10), 780–792, Oct. 1994.

[31] van de Kerkhof, L.M. and Cugnini, A.G., The ISO/MPEG audio coding standard, *Widescreen Review*, 1994.

[32] Dehery, Y.F., Stoll, G. and Kerkhof, L.v.d., MUSICAM source coding for digital sound, 17th International Television Symposium, Montreux, Record 612–617, june 1991.

[33] Brandenburg, K., Herre, J., Johnston, J.D., Mahieux, Y. and Schroeder, E.F., ASPEC: Adaptive spectral perceptual entropy coding of high quality music signals, *90th. Audio Engineering Society-Convention, Paris,* Preprint 3011, 1991.

[34] Musmann, H.G., The ISO audio coding standard, *Proc. IEEE Globecom*, Dec. 1990.

[35] van der Waal, R.G., Oomen, A.W.J. and Griffiths, F.A., Performance comparison of CD, noise-shaped CD and DCC, in *Proc. 96th Audio Engineering Society Convention,* Amsterdam, Preprint 3845, 1994.

[36] Herre, J., Brandenburg, K. and Lederer, D., Intensity stereo coding, *96th Audio Engineering Society Convention*, Amsterdam, Preprint no. 3799, 1994.

[37] Pan, D., A tutorial on MPEG/audio compression, *IEEE Trans. on Multimedia*, 2(2), 60–74, 1995.

[38] Brandenburg, K. et al., Variable data-rate recording on a PC using MPEG-audio layer III, *5th Audio Engineering Society Convention*, New York, 1993.

[39] Sarginson, P.A., MPEG-2: Overview of the system layer, *BBC Research and Development Report*, BBC RD 1996/2, 1996.

[40] Ryden, T., Grewin, C. and Bergman, S., The SR report on the MPEG audio subjective listening tests in Stockholm April/May 1991, ISO/IEC JTC1/SC29/WG 11: Doc.-No. MPEG 91/010, May 1991.

[41] Fuchs, H., Report on the MPEG/audio subjective listening tests in Hannover, ISO/IEC JTC1/SC29/WG 11: Doc.-No. MPEG 91/331, Nov. 1991.

[42] Stoll, G. et al., Extension of ISO/MPEG-audio layer II to multi-channel coding: The future standard for broadcasting, telecommunication, and multimedia application, *94th Audio Engineering Society Convention,* Berlin, Preprint no. 3550, 1993.

[43] Grill, B. et al., Improved MPEG-2 audio multi-channel encoding, *96th Audio Engineering Society Convention*, Amsterdam, Preprint 3865, 1994.

[44] Bosi, M. et al., ISO/IEC MPEG-2 advanced audio coding, *101th Audio Engineering Society Convention*, Los Angeles, Preprint 4382, 1996.

[45] Johnston J.D. et al., NBC-audio - stereo and multichannel coding methods, *101th Audio Engineering Society Convention,* Los Angeles, Preprint 4383, 1996.

[46] Ten Kate, W.R.Th. et al., Matrixing of bit rate reduced audio signals, *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP'92), 2, II-205–II-208, 1992.

[47] ten Kate, W.R.Th., Compatibility matrixing of multi-channel bit-rate-reduced audio signals, *96th Audio Engineering Society Convention*, Preprint 3792, Amsterdam, 1994.

[48] Feige, F. and Kirby, D., Report on the MPEG/audio multichannel formal subjective listening tests, ISO/IEC JTC1/SC29/WG 11: Doc. N 0685, March 1994.

[49] Meares, D. and Kirby, D., Brief subjective listening tests on MPEG-2 backwards compatible multichannel audio codecs, ISO/IEC JTC1/SC29/WG 11: Aug. 1994.

[50] ISO/IEC/JTC1/SC29, Report on the formal subjective listening tests of MPEG-2 NBC multichannel audio coding, Document N1371, Oct. 1996.

[51] ITU-R Document TG 10-2/3, Oct. 1991.

[52] /IEC/JTC1/SC29, Description of MPEG-4, Document N1410, Oct. 1996.

[53] Burpee, D.S. and Shumate, P.W., Emerging residential broadband telecommunications, *Proc. IEEE*, 82(4), 604–614, 1994.

[54] Jayant, N.S., High quality networking of audio-visual information, *IEEE Commun. Mag.*, 84–95, 1993.

[55] Ten Kate, W.R.Th., Akagiri, K., van de Kerkhof, L.M. and Kohut, M. J., Scalability in MPEG audio compression. From stereo via 5.1-channel surround sound to 7.1-channel augmented sound fields, *100th Audio Engineering Society Convention*, Copenhagen, 1996, Preprint 4196.

[56] Sugiyama, A. et al., A new implementation of the silicon audio player based on an MPEG/Audio decoder LSI, Technical Report DSP94-99 (1994-12) of the IEICE, 39–45, 1994.

[57] Lau, A. and Williams, W.F., Service planning for terrestrial digital audio broadcasting, *EBU Technical Review*, 4–25, 1992.

[58] Plenge, G., DAB—A new sound broadcasting systems: status of the development—routes to its introduction, *EBU Review*, April 1991.

[59] ETSI, European Telecommunication Standard, Draft prETS 300 401, Jan. 1994.

[60] Weck, Ch., The error protection of DAB, *Audio Engineering Society-Conference "DAB - The Future of Radio"*, London, May 1995.

[61] Jurgen, R.D., Broadcasting with digital audio, *IEEE Spectrum*, 52–59, March 1996.

[62] Todd, C. et al., AC-3: Flexible perceptual coding for audio transmission and storage, *96th Audio Engineering Society Convention*, Amsterdam, Preprint 3796, 1994.

[63] Hopkins, R., Choosing an American digital HDTV terrestrial broadcasting system, *Proc. IEEE*, 82(4), 554–563, 1994.

[64] The grand alliance, *IEEE Spectrum* 36–45, April 1995.