Signal Processing for Multimedia
J.S. Byrnes (Ed.)
IOS Press, 1999

Wideband Audio

Peter Noll

Technische Universität Berlin, Germany
Einsteinufer 25 D-10587 Berlin
noll@ee.tu-berlin.de

ABSTRACT. This chapter covers key technologies in wideband audio coding including auditory masking, perceptual coding, frequency domain coding, and dynamic bit allocation. The MPEG standardization work is then described. MPEG algorithms have found a wide range of communications-based and storage-based applications. For example, European digital audio broadcast (DAB) makes use of MPEG-1. It will then be shown that the MPEG-2 *Advanced Audio Coding* (AAC) standard offers a powerful collection of very flexible tools for stereo and multichannel coding, and that AAC outperforms many other coding algorithms (including MPEG-1 coders). Finally, we address the current MPEG-4 speech and audio coding standardization work which merges the whole range of audio, from high fidelity audio coding and speech coding down to synthetic audio, synthetic speech and text-to-speech conversion.

1. Introduction

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. Wideband (high fidelity) audio representations, including multichannel audio, need bandwidths of at least 20 kHz. The conventional digital format of digital audio is PCM, with sampling rates of 32, 44.1, or 48 kHz and an amplitude resolution (PCM bits per sample) of 16 bits. Typical application areas for digital audio are in the fields of audio production, program distribution and exchange, digital sound broadcasting, digital storage, and various multimedia applications.

For archiving and processing of audio signals, highest quality formats with 96 kHz sampling and 24 to 30 bit amplitude resolution are under discussion. In some applications coding will have to be *lossless*—with compression factors around two to three. Upcoming Digital versatile Discs (DVD), with their capacity of 4.7 GB (single layer) or 8.5 GB (double layer), will be the appropriate storage devices for lossless-coded audio material. The capacity can be doubled if both sides of the discs are readable. [The capacities are for Read Only Memory (ROM) discs only. A Random Access Memory (RAM) disc has only 2.6 GB instead of 4.7 GB].

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 x 44.1 x 16 x 1000 = 1.41 Mb/s (see table 1). However, the CD needs significant overhead for a run-length-limited line code, 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 x 49/16 = 4.32 Mb/s. Table 1 compares parameters of the Compact Disc and the *Digital Audio Tape* (DAT) with those of two more recent storage systems, Philips *Digital Compact Cassette (DCC)* and Sony's 64 mm optical or magneto-optical *MiniDisc* (MD), and with the parameters of the European Digital Audio Broadcast (DAB). The source coding algorithms of DCC, MD, and DAB are, respectively, MPEG-1, ATRAC, and MPEG-1.

TABLE 1: Bit rates for various digital audio schemes (Stereophonic signals)
[DCC also supports sampling rates of 32 kHz and 48 kHz.
DAB supports a number of bit rates and different amounts of error protection (see later section)]

Applications	Format	Sampling rate	Audio bit rate	Overhead bit rate	Total bit rate
Compact Disc (CD)	PCM	44.1 kHz	1.41 Mb/s	2.91 Mb/s	4.32 Mb/s
Digital Audio Tape (DAT)	PCM	44.1 kHz	1.41 Mb/s	1.67 Mb/s	3.08 Mb/s
Digital Compact Cassette (DCC)	MPEG-1	48 kHz	384 kb/s	384 kb/s	768 kb/s
MiniDisc (MD)	ATRAC	44.1 kHz	292 kb/s	718 kb/s	1.01 Mb/s
Digital Audio Broadcast	MPEG-1	48 kHz	256 kb/s	256 kb/s	512 kb/s

We cover the audio coding standards MPEG-1 and MPEG-2 in some detail, since they have been the first international standards in the field of high quality digital audio compression. MPEG-1 covers coding of stereophonic audio signals at high sampling rates aiming at *transparent* quality, whereas MPEG-2 offers three extensions of the basic MPEG-1 digital audio standard:

- stereophonic audio coding at lower sampling rates,
- multichannel coding
- Advanced Audio Coding (AAC).

The very recent MPEG-2 AAC standard offers a collection of very flexible tools for various applications, and it offers the highest compression rates.

We also will include a short section on the current MPEG-4 work, which addresses standardization of audiovisual coding for applications ranging from mobile access low complexity multimedia terminals to high quality multichannel sound systems. The standard will allow for interactivity and universal accessibility, and will provide a high degree of flexibility and extensibility.

2. Key Technologies in Audio Coding

We have seen rapid progress in bit rate compression techniques for speech and audio signals [1–4]. 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 excellent coding quality can be obtained with bit rates of 0.5 b/sample for speech and wideband speech, and 1 b/sample for audio.

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 representations, and, finally, higher quality expectations. Indeed, the high quality of the Compact Disc with its 16-b/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.

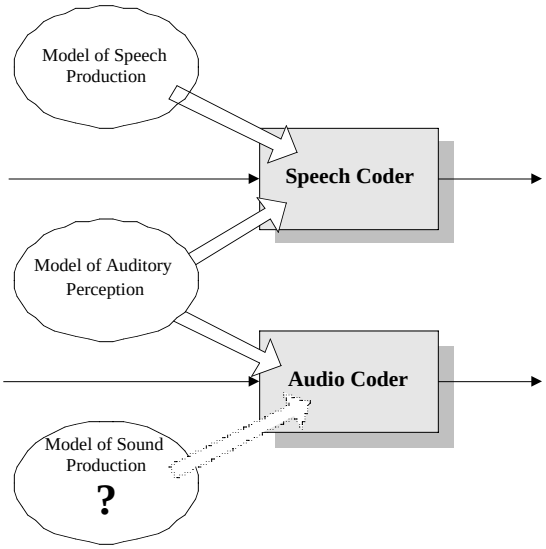

FIGURE 1: Efficient speech and audio compression by employing models of perception and production

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

2.1. *Perceptual Coding*

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 on the order of 50 to 100 Hz for signals below 500 Hz and up to 5000 Hz for signals at high frequencies. 25 critical bands covering frequencies 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 [5]. Such masking is largest 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; low-level signals below this threshold 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) [6], 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 figure 2: 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*. This depends on frequency and covers a dynamic range of more than 60 dB as shown in the lower curve of figure 2.

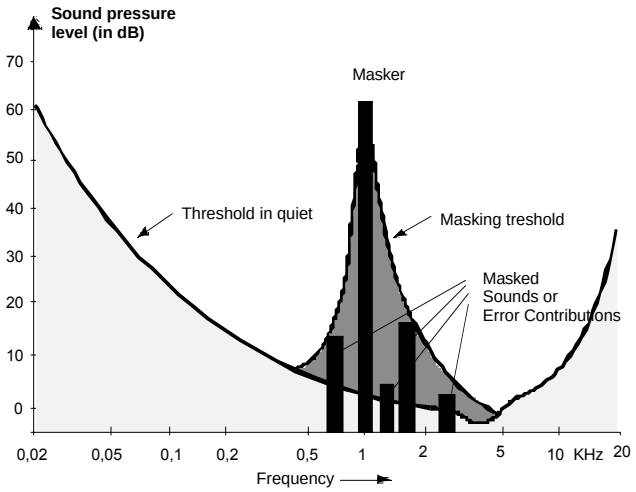

FIGURE 2: Threshold in quiet and masking threshold
Acoustical events in the gray areas will not be audible.

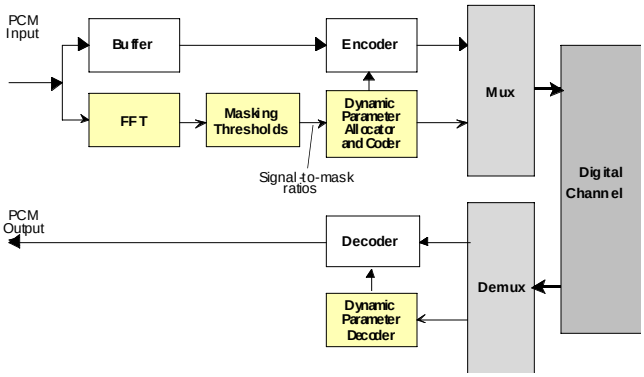

FIGURE 3: Block diagram of perception-based coders

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.

The dependence of human auditory perception on frequency and on 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 is not important for perception. 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. This can be done in different ways. However, frequency weightings based on linear filtering, 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 3 depicts the structure of a *perception-based coder* that exploits auditory masking. The encoding process is controlled by the signal-to-mask ratio (SMR), the ratio of short-term signal power within each frequency band to the masking threshold. From the SMR 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 (1024-point, say) FFT-based spectral analysis of the audio block to be coded. In general, any coding scheme may be used that can be dynamically controlled by such perceptual information. Frequency domain coders (see next section) are of particular interest since they

offer a direct method for noise shaping.

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, since 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, since such models are not normative elements of the standard.

2.2. *Frequency Domain Coding*

Frequency domain coders with dynamic allocations of bits (and hence of quantization noise contributions) to subbands or transform coefficients offer an easy and accurate way to control the quantization noise [1], [7]. *Hybrid filterbanks, i.e.*, combinations of discrete transform and filterbank implementations, have frequently been used in speech and audio coding. 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 audio coders use a subband approach in Layer I and Layer II, and a hybrid filterbank in Layer III.

2.3. *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 the short-term spectrum of the audio coding block on a block-by-block basis [7]. Such a *dynamic bit allocation* is used in all recent audio coding algorithms. They do not assign bits in order to minimize the overall mean-squared reconstruction error, but instead assign bits such that the overall *perceptual quality* is maximized.

3. ISO/MPEG-1 Audio Coding

3.1. *Structure and Layers*

The MPEG-1 audio coding standard [8–10] has already become a universal standard in diverse fields, such as consumer electronics, professional audio processing, telecommunications, and broadcasting. It offers a subjective reproduction quality that is equivalent to compact disc (CD) quality (16 bit PCM) at stereo rates given in table 2 for many types of music.

TABLE 2: Approximate (conservative) MPEG-1 bit rates for transparent representations of audio signals and corresponding compression factors (compared to CD bit rate)
* Average bit rate; variable bit rate coding assumed.

MPEG-1 audio coding	Approximate stereo bit rates for transparent quality	Compression factor
Layer I	384 kb/s	4
Layer II	256 kb/s	6
Layer III	192 kb/s*	8

Structure. The basic structure follows that of perception-based coders (see figure 3). 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. The number of quantizer bits is obtained from a dynamic bit allocation algorithm that is controlled by a *psychoacoustic model* (see below). The subband codewords, the scalefactor, and the 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.

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 point of view, the higher layers incorporate the main building blocks of the lower layers. A standard *full MPEG-1 audio decoder* is able to decode bit streams of all three layers. More typical are MPEG-1 audio *Layer X decoders* (X = I, II, or III).

Psychoacoustic Models. We have already mentioned that the adaptive bit allocation algorithm is controlled by a psychoacoustic model. This model computes signal-to-mask ratios (SMR), taking into 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 [7]. 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. 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 as the sum of all individual masking thresholds and the absolute masking threshold.

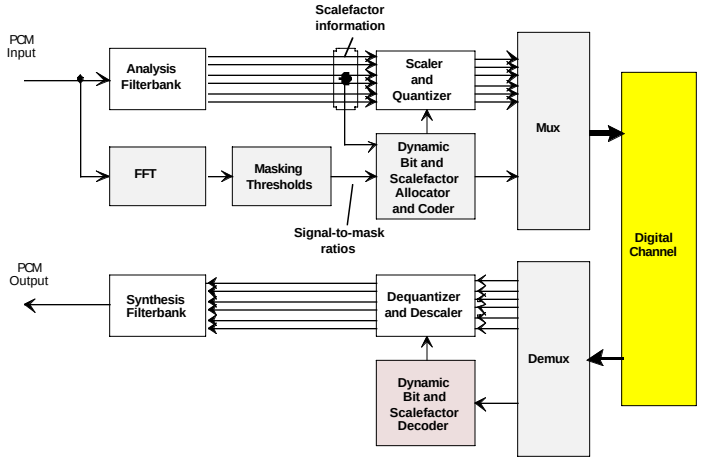

FIGURE 4: Structure of MPEG-1/Audio encoder and decoder, layers I and II

3.2. *Layer II*

Layer I and Layer II coders map the digital audio input into 32 subbands via equally spaced bandpass filters (figure 4). A polyphase filter structure is used for the frequency mapping; its filters have 512 coefficients. Polyphase structures are computationally very efficient since 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 2.1). At a 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.

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 scaled and quantized spectral subband components are transmitted to the receiver together with scalefactor and bit allocation information.

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.

3.3. *Layer III*

Layer III of the MPEG-1/Audio coding standard introduces many new features, 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.

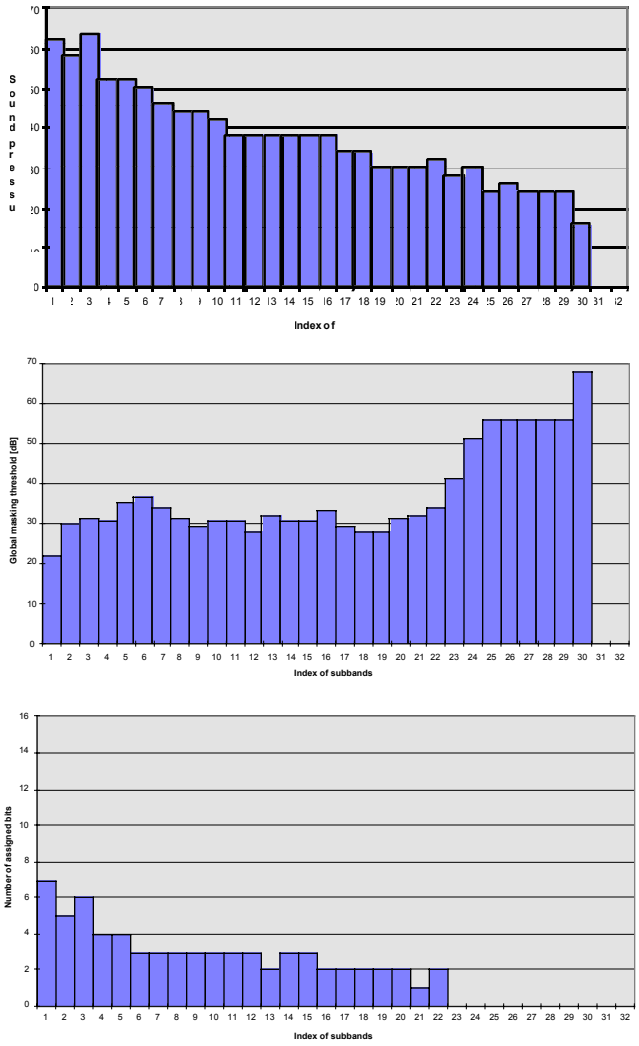

FIGURE 5: Parameters for of a given audio block
(a) Power spectrum (b) Global masking threshold (c) Bit allocation

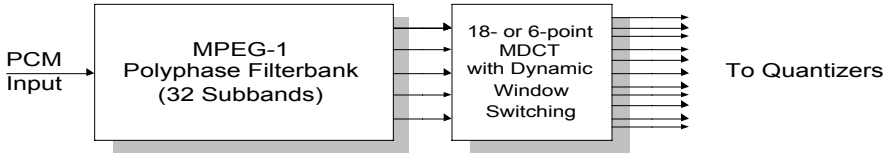

FIGURE 6: Hybrid filterbank of MPEG-1 Layer III encoder

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-point or 18-point modified DCT block transform, with 50% overlap; hence, the windows contain, respectively, 12 or 36 subband samples (see figure 6).

The maximum number of frequency components is 32 x 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. In principle, a pre-echo is assumed when an instantaneous demand for a high number of bits occurs. Depending on the nature of potential pre-echoes, all 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 do-

main alias cancellation feature of the MDCT.

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 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.

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.

3.4. *Example: Digital Audio Broadcast (DAB)*

The DAB system employs the MPEG-1 Layer II standard with its range of bit rates between 32 and 384 kb/s. The transmission scheme was especially designed to allow mobile reception under additive noise and multipath propagation conditions. DAB is a multicarrier system employing an efficient unequal bit error protection strategy. Therefore DAB provides excellent audio quality over a wide range of carrier-to-noise ratios (C/N), unlike analog FM where the quality of the audio signal follows directly the C/N of the channel (see figure 7). Poor reception conditions exist not only at the borders of service coverage areas, but also within the coverage area due to shadowing effects etc.

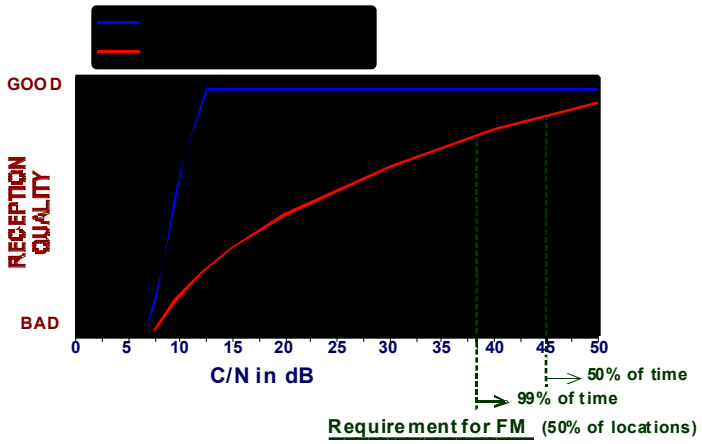

FIGURE 7: Comparison of Analog FM and Digital Audio Broadcast [11]
Rayleigh-Fading Channel, Rural I Model, speed of mobile 50 km/h

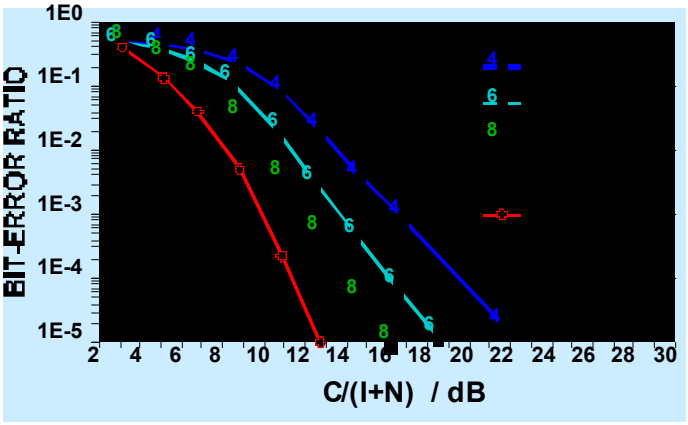

FIGURE 8: Residual bit error rates [11]
Rayleigh-Fading Channel, Rural 1 Model, speed of mobile: 50 km/h

TABLE 3: Protection of DAB audio frames

Protection class 1	high error protection	DAB header, CRC codeword, bit allocation, scalefactor select info
Protection class 2	medium error protection	Scalefactors of the subbands
Protection class 3	lower error protection	Subband samples
Protection class 4	medium error protection	CRC codewords for scalefactor- select info program associated data (PAD)

The DAB system employs convolutional coding with a constraint length L = 7 and soft-decision Viterbi decoding with 64 states. The minimum applicable code rate for DAB is 1/4, *i.e.*, 300% redundancy is added to the bitstream. Less protection is obtained by puncturing code bits of the rate 1/4 mother code. The punctured code bits are not transmitted and they are treated as erased bits in the decoding process. Therefore the decoder need not be modified in case of puncturing.

The DAB specification provides 24 code rates from 8/9 to 8/32. Figure 8 shows the *residual* BER vs. C/N performance for code rates between 2/3 and 1/3 [11]. These curves have been obtained by simulations employing Rayleigh-fading channels. Parts of the audio bitstream are coded with different code rates, whereby the error protection depends on the perceptual distortion in case of bit errors. A so-called *protection profile* is used to define the different protection classes and the corresponding bit rates. DAB has standardized 64 of these protection profiles. In the DAB encoding process four protection classes are used for each DAB audio frame of length 24 ms.

In addition to protection profiles, five *protection levels* are supported. These levels define the amount of average error protection for different types of applications. Examples are a code rate of 3/4 for audio signal distribution over cable networks and a code rate of 2/5 for mobile reception at high vehicle speed. Due to the convolutional coding there is a smooth transition between the four protection classes. Table 3 describes their parameters.

Infrequently, deep fading occurs and algebraic channel coding will break down. In these cases concealment techniques have to be evoked to allow for a graceful degradation of quality since residual bit errors can lead to very annoying distortions. Muting in case of loss of control information and repetition of parameters of the previous frame are typical means to conceal infrequently occurring residual errors. Since concealment strategies are to be implemented in the receiver, they are not part of the DAB specifications.

4. MPEG Advanced Audio Coding

The MPEG-2 AAC standard employs high resolution filter banks, prediction techniques, and noiseless coding [12], [13]. It is based on recent evaluations and definitions of *tools (or modules)* each having been selected from a number of proposals. The self-contained tools include optional preprocessing, a filterbank, a perceptual model, temporal noise shaping, intensity multichannel coding, time-domain prediction, M/S stereo coding, quantization, noiseless coding, and a bit stream multiplexer (see figure 9). The filterbank is a 1024-point modified discrete cosine transform, the perceptual model is taken from MPEG-1 (model 2).

The temporal noise shaping (TNS) tool plays an important role in improving the overall performance of the coder (see figure 10). It performs a prediction of the spectral coefficients of each audio frame. Instead of the coefficients, the prediction residual is transmitted. TNS is very effective in case of transient audio signals since such transients (signal “attacks”) imply high predictability in the spectral domain. (Recall that “peaky” spectra lead to high predictability in the time domain). Therefore, the TNS tool controls the time dependence of the quantization noise.

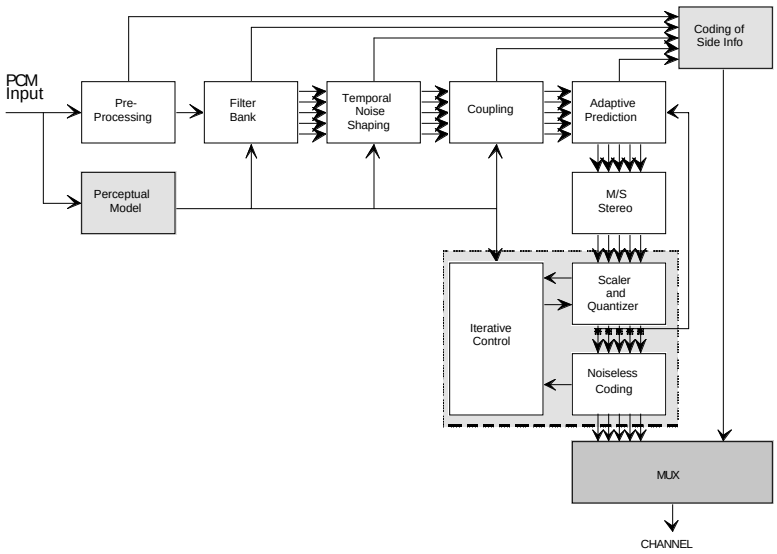

FIGURE 9: Structure of MPEG-2 Advanced Audio Coder (AAC)

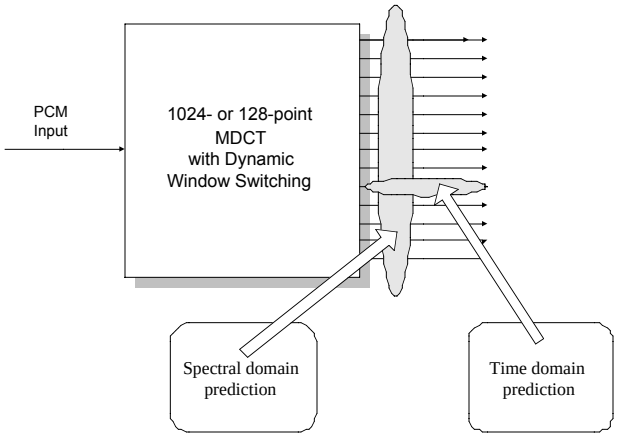

FIGURE 10: Spectral and time domain prediction in MPEG-2 Advanced Audio Coding (AAC)

Time domain prediction is applied to subsequent subband samples in a given subband in order to further improve coding efficiency, in particular for stationary sounds (see figure 10). Second-order backward-adaptive predictors are used for this purpose. Finally, for quantization and noiseless coding, an iterative method is employed so as to keep the quantization noise in all critical bands below the global masking threshold.

Profiles. In order to serve different needs, the standard provides three profiles: (i) the main profile offers highest quality, (ii) the low complexity profile works without prediction, (iii) and the sampling-rate-scaleable profile offers lowest complexity. For example, in its *main profile* the filterbank is a 1024 line modified discrete cosine transform (MDCT) with 50% overlap (blocklength of 2048 samples). The filterbank is switchable to eight 128 line MDCTs (blocklengths of 256 samples). Hence, 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. The *low complexity profile* does not employ temporal noise shaping and time domain prediction, whereas in the *sampling-rate-scaleable profile* a hybrid filterbank is applied.

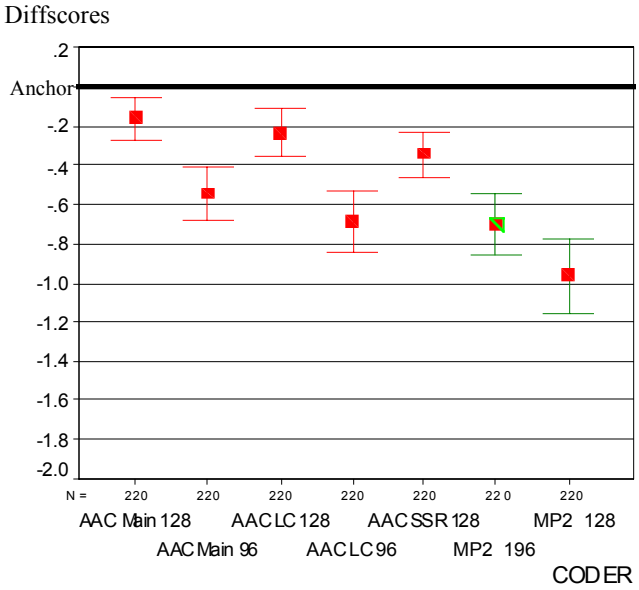

FIGURE 11: Subjective quality of AAC and MPEG-1 audio coders [MPEG document].

The MPEG-2 AAC standard offers high quality at lowest possible bit rates; it will find many applications, for both consumer and professional use. The following figure shows MOS differences, with diffscore = 0 for the compact disc reference. For example, the AAC coder operating at 128 kb/s stereo rate is close to the MOS value of the reference (with a diffscore of around –0.18). At that rate, the MPEG-1 Layer 3 coder (MP3 128) has a diffscore of almost –1. Note also that the AAC main coder performs better at a rate of 96 kb/s than the MPEG-1 Layer 1 coder at twice the rate (MP2 192).

5. MPEG Multichannel Audio Coding

5.1. *Multichannel audio representations*

A logical further step in digital audio is the definition of multichannel audio representation systems to create a realistic surround-sound field, 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 (dialogue with enhanced intelligibility). ITU-R and other international groups have 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 and two side/rear surround channels (LS and RS) augmenting the L and R channels. Such a configuration offers a surround-sound field with a stable frontal sound image and a large listening area. ITU-R Recommendation 775 provides a set of downwards mixing equations if the number of loudspeakers is to be reduced (*downwards compatibility).*

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, components of the multichannel signal which are irrelevant with respect to the spatial perception of the stereophonic presentation, *i.e.*, 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. However, a careful design is needed, otherwise such joint coding may produce artifacts.

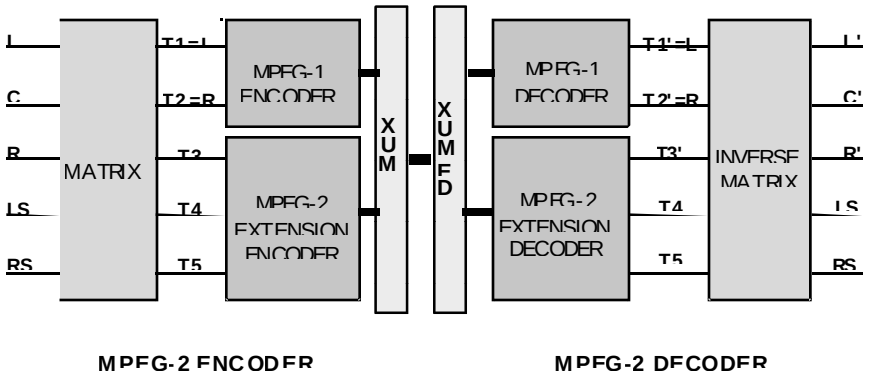

FIGURE 12: Compatibility of MPEG-2 multichannel audio bit streams

5.2 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 backwards-compatible with MPEG-1 Audio [14]. *Forward compatibility* means that an MPEG-2 multichannel decoder is able to properly decode MPEG-1 mono- or stereophonic signals, *backwards compatibility* means that existing MPEG-1 stereo decoders, which only handle two-channel audio, are able to reproduce a meaningful basic 2/0 stereo signal from an MPEG-2 multichannel bit stream so as to serve the need of users with simple mono or stereo equipment. *Non-backwards compatible* multichannel coders will not be able to feed a meaningful bit stream into an MPEG-1 stereo decoder. On the other hand, *non-backwards compatible* codecs have more freedom in producing high quality reproduction of audio signals.

With backwards 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 DAB system, which will require MPEG-1 stereo decoders in the first generation but may offer multichannel audio at a later point.

5.2.1 Backwards-Compatible MPEG-2 Audio Coding

Backwards compatibility 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, L0 and R0, respectively. The signals L0 and R0 are transmitted in MPEG-1 format in transmission channels T1 and T2. Channels T3, T4, and T5 together form the *multichannel extension signal* (figure 12). They have to be chosen so 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. Note that the user with stereophonic equipment can use signals T1’ = L0’ and T2’ = R0’ directly.

Matrixing is obviously necessary to provide backwards compatibility; however, if used in connection with perceptual coding, “unmasking” of quantization noise may appear

5.2.2. *Nonbackwards-Compatible MPEG-2 Advanced Audio Coding*

A second standard within MPEG-2 supports applications which do not require compatibility with the existing MPEG-1 stereo format. Therefore, matrixing and dematrixing are not necessary, and the corresponding potential artifacts disappear (see figure 13). Within the MPEG-2 standard the multichannel advanced audio coding technique (AAC) has been proposed which offers highest performance at comparably low overall bit rates.

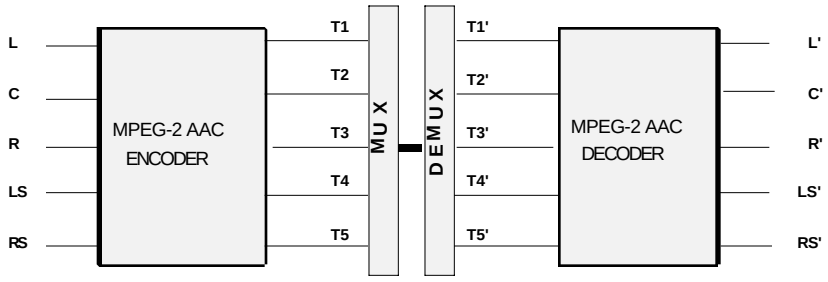

FIGURE 13: MPEG-2 Advanced Audio Coding (multichannel configuration)

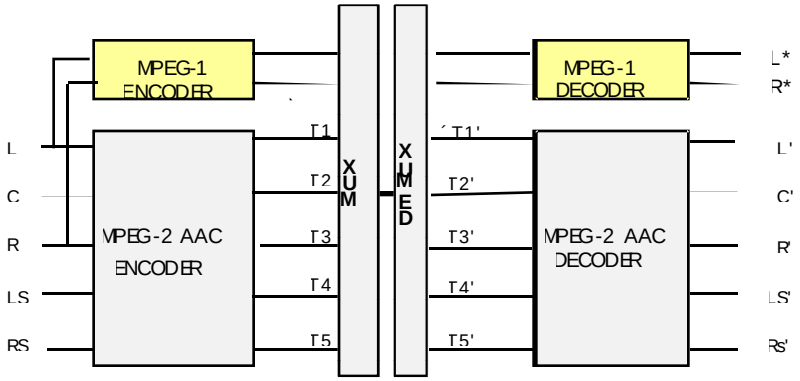

FIGURE 14: Backwards-compatible MPEG-2 multichannel audio coding (simulcast mode)

Recently, extensive formal subjective tests have been carried out to compare MPEG-2 AAC versions and a backward-compatible MPEG-2 Layer II coder [15]. All coders showed very good performance, with a slight advantage of the *non-backwards-compatible* 320 kb/s MPEG-2 AAC coder compared with the *backwards-compatible* 640 (!) kb/s MPEG-2 Layer II coder. From these subjective tests it has become clear that the concept of backwards-compatibility implies the need for significantly higher bit rates. The AAC coder satisfies requirements for ITU-R broadcast quality at 320 kbit/s for 5 channels.

5.3 Backwards compatibility via 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 a full non-backwards-compatible multichannel bit stream to support backwards compatibility without matrixing techniques (figure 14).

6. MPEG-4 Audio Coding

6.1. *MPEG-4 Coding of Audiovisual Scenes*

Activities within MPEG-4 have aimed at proposals for a broad field of applications including multimedia. It is clear that communication services, interactive services and broadcast services will overlap in future applications. The new standard, which will become an international standard in early 1999, takes into account that *a growing part of information is read, seen and heard in interactive ways*. It will support new forms of communications, in particular

- Internet
- Multimedia
- Mobile Communications.

FIGURE 15: Audiovisual scene [16]

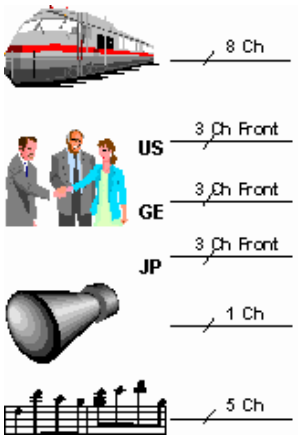

FIGURE 16: Audio channels for the audiovisual scene of Figure 15 [16]

Indeed, MPEG-1 and MPEG-2 have some main disadvantages: they offer only very limited interaction and control over the presentation and configuration of the system. In addition, integration of natural and synthetic content is difficult, and access and transmission across heterogeneous networks is not well-supported.

MPEG 4 is different: it represents an audiovisual scene as a composition of (potentially meaningful) objects and supports the evolving ways in which audiovisual material is produced, delivered, and consumed. For example, computer-generated content becomes part in the production of an audiovisual scene. In addition, interaction of objects with scene is possible. For example, it will be possible to associate a URL to a person in a scene.

6.2. *MPEG-4 Audio Coding*

In the case of *audio*, MPEG-4 will merge the whole range of audio from high fidelity audio coding and speech coding down to synthetic speech and synthetic audio, supporting applications from high-fidelity audio systems down to mobile-access multimedia terminals. The following figures indicate the potential of MPEG-4: figure 15 describes an audiovisual scene with a number of audio "objects": the noise of an incoming train, an announcement, and a conversation. Figure 16 lists the corresponding objects in detail.

For example, the noise of the train can be described by an eight-channel representation. On the other hand, if the necessary bandwidth is not available, a one-channel representation—or no representation at all—could be used instead. Such a form of scalability will be very useful in future applications whenever audiovisual signals have to be transmitted to and via receivers of differing complexity and channels of differing capacity. In the case of the announcement, one-channel, pseudo 3-D and echo effects could be added. The background music may have an AAC format, or it is of synthetic origin.

MPEG-4 offers tools which can be combined to satisfy specific user requirements [17]. A number of such configurations will be standardized. A syntactic description is 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. In the case of audio and speech the current toolset supports compression at monophonic bit rates ranging from 2 to 64 kb/s. Three *core coders* are used:

- a parametric coding scheme ("vocoder") for low bit rate speech coding (2 to 10 kbit/s)
- a CELP-based analysis-by-synthesis coding scheme for medium bit rates (4 to 16 kb/s)
- a transform-based coding scheme for higher bit rates (up to 64 kbit/s).

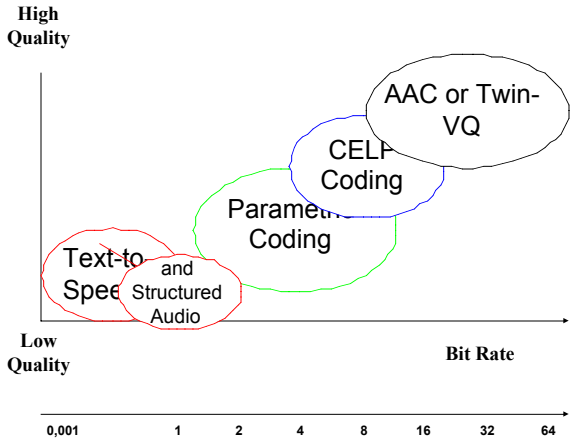

FIGURE 17: Range of qualities and bit rates in MPEG-4.

MPEG-4 will not only offer simple means of manipulation of coded data such as time scale control and pitch change, but also flexible access to coded data and subsets thereof, *i.e.*, scalability of bit rate, bandwidth, complexity, and error robustness. In addition, MPEG-4 supports not only natural audio coding at rates between 2 and 64 kb/s**,** but also text-to-speech conversion (TTS) and structured audio. Natural sounding TTS is obtained by combining conventional TTS synthesis with additional prosodic parameters. The standard also offers an interface between TTS and facial animation for synthetic face models to be driven from speech.

Ultra-low bit rate coding of sound is achieved by coding and transmitting parameters of a sound model. MPEG-4 standardizes a sound language and related tools for structured coding of synthetic music and sound effects at rates of 0.01 to 10 kb/s. MPEG-4 does not standardize a particular set of synthesis methods, but a signal-processing language for describing synthesis methods. Any current or future sound-synthesis method may be described in the MPEG-4 structured audio format. The language is entirely normative and standardized, so that every piece of music will sound exactly the same on every compliant MPEG-4 decoder.

Figure 17 indicates the range of bit rates offered by the new standard.

6.3. *Example: Coding of Low-Complexity Music*

Transform-based audio coders show very good performance at bit rates down to 16 kb/s, whereas speech coders perform clearly better at rates below 16 kb/s. Currently, a number of speech coders are available with good performance in the range of bit rates between 2.4 kb/s and 16 kb/s. Both coder classes do not offer solutions for audio coding at 4–16 kb/s. The MPEG-4 standard contains a concept for music of low complexity content [18]. It is assumed that such audio material can be decomposed in

- harmonic tones (fundamental frequency plus harmonic partials)
- individual sinusoids
- and noise.

In MPEG documents the term HILN is used to describe such a segmentation. (HILN stands for “Harmonics and Individual Lines plus Noise”). Model parameters of these audio objects are estimated, quantized and coded in an iterative analysis-by-synthesis approach which is controlled by a perception model. Figure 18 shows the block diagram of such an object-based audio encoder.

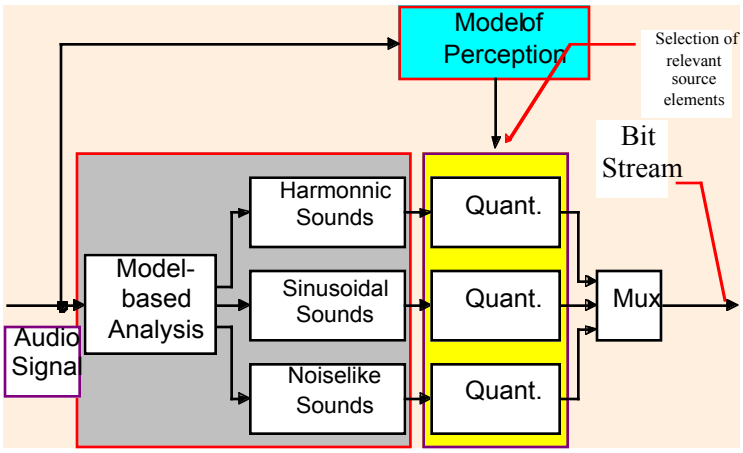

FIGURE 18: Object-based audio encoder [18].

We note in passing that this coding structure permits, in an easy way, new functionalities such as change of pitch and/or playback speed. The pitch can be changed by multiplying all frequency parameters by a given factor, the speed can be modified by changing the length of the synthesized time frames [18].

7. Conclusions

Low bit rate wideband 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 for high compression rates in audio coding. The MPEG audio coding standards have been widely accepted as state-of-the-art coders. All MPEG audio coders are controlled by psychoacoustic models which may be improved, thus leaving room for an evolutionary improvement of these coders.

The MPEG-4 standard merges the whole range of audio from high fidelity audio coding and speech coding down to text-to-speech conversion and synthetic audio. MPEG-4 offers new functionalities such as time scale changes, pitch control, database access, and scalability, which allows one to extract from the transmitted bitstream a subset sufficient to generate audio signals with lower bandwidth and/or lower quality, depending on channel capacity or decoder complexity. MPEG-4 will be the future multimedia standard.

References

[1] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice Hall, 1984.

[2] A. S. Spanias, Speech Coding: A Tutorial Review, *Proc. of the IEEE*, Vol. 82, No. 10, pp. 1541–1582, Oct. 1994.

[3] A. Gersho, Advances in Speech and Audio Compression, *Proc. of the IEEE*, vol. 82, No.6, pp. 900–918, 1994.

[4] P. Noll, Digital Audio Coding for Visual Communications, *Proc. of the IEEE*, vol. 83, No. 6, June 1995.

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

[6] N. S. Jayant, J. D. Johnston, and R. Safranek, Signal Compression Based on Models of Human Perception, *Proc. of the IEEE*, vol. 81, No. 10, pp. 1385–1422, 1993.

[7] R. Zelinski and P. Noll, Adaptive Transform Coding of Speech Signals, *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. ASSP-25, pp. 299–309, August 1977.

[8] ISO/IEC JTCl/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] K. Brandenburg and G. Stoll, The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio, *Journal of the Audio Engineering Society (AES)*, Vol. 42, No. 10, pp. 780–792, Oct. 1994.
[10] P. Noll, MPEG Audio Coding Standards, *IEEE Signal Processing Magazine*, Sept. 1997.
[11] C. Weck The Error Protection of DAB, *AES Conference DAB—The Future of Radio*, London,1995.
[12] ISO/IEC JTCl/SC29, Information Technology—Generic Coding of Moving Pictures and Associated Audio Information - IS 13818 (Part 7), 1997.
[13] M. Bosi et al, ISO/IEC MPEG-2 Advanced Audio Coding, *J. Audio Eng. Soc.*, Vol 45, No. 10, pp. 789–814, 1997.
[14] ISO/IEC JTCl/SC29, Information Technology—Generic Coding of Moving Pictures and Associated Audio Information - IS 13818 (Part 3, Audio), 1994.
[15] ISO/IEC/JTC1/SC29, Report on the Formal Subjective Listening Tests of MPEG-2 NBC Multichannel Audio Coding, Document N1371, Oct. 1996.
[16] ISO/IEC/JTC1/SC29, MPEG Document 97/2304.
[17] ISO/IEC/JTC1/SC29, Description of MPEG-4, Document N1410, Oct. 1996.
[18] H. Purnhagen, B. Edler, C. Ferekidis, Object-Based Analysis/Synthesis Audio Coder for Very Low Bit Rates, 104th *Audio Engineering Society Convention (AES)*, Amsterdam, 1998].