Lossless and Perceptual Coding of Digital Audio

Peter Noll, Tilman Liebchen
Technische Universität Berlin, Fachgebiet Nachrichtenübertragung (formerly Fernmeldetechnik)
noll@nue.tu-berlin.de, liebchen@nue.tu-berlin.de

This paper is dedicated to Professor Dr. Werner Endres

on the occasion of his 90th birthday.

Abstract: We have seen rapid progress in high-quality compression of wideband audio signals. Today's coding algorithms can achieve substantially better compression than was thought possible only a few years ago. In the case of audio coding with its bandwidth of 20 kHz and more, the concept of perceptual coding has paved the way for significant bit rate reductions.

However, multiple coding can reveal originally masked distortions. In addition, reproduction of critical music items shows that even the best systems can not be considered as truly transparent. Therefore lossless audio coding has become a topic of high interest both for professional and customer applications.

This paper will explain approaches to lossless and lossy compression, both with emphasis on MPEG standards which have found a wide range of communications-based and storage-based applications. As an example for state-of the-art lossless coding, an overview of the forthcoming MPEG-4 *Audio Lossless Coding* (ALS) standard will be presented. On the other hand, it will be shown that the recent MPEG-4 *Advanced Audio Coding* (AAC) standard outperforms many other perceptual coding algorithms (including MP3 coders). Finally, we will 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

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 bit. 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 up to 192 kHz sampling and 24 to 32-bit amplitude resolution are already used.

Audio coding is employed in order to reduce bit rate compared to the PCM representation. In some applications coding will have to be *lossless*, with compression factors around two as will be shown shortly. For other applications, *perceptually transparent* coding will be sufficient, which allows to compress the audio data to less than a tenth of its original size.

The *Compact Disc* (CD) is today's *de facto standard* for disc-base delivery of digital audio. The CD uses the PCM format with 16-bit amplitude resolution and 44.1 kHz sampling rate,

resulting in a stereo bit rate of $2 \cdot 44100 \cdot 16 = 1.41$ Mbit/s. On the other hand, portable music players using MPEG-1 layer III (MP3) compression at 128 kbit/s have been increasingly successful. Table 1 compares parameters of the CD with other digital audio storage and transmission systems.

Application	Format	Sampling rate	Audio bit rate
Compact Disc (CD)	PCM	44.1 kHz	1.41 Mbit/s
MiniDisc (MD)	ATRAC	44.1 kHz	292 kbit/s
Digital Radio (DAB)	MPEG-1 Layer II	48 kHz	≤ 256 kbit/s
MP3 Players	MP3, WMA, AAC	44.1 kHz	128 kbit/s
Internet Streaming	MP3, WMA, RealAudio	≤ 44.1 kHz	≤ 128 kbit/s

Table 1: Bit rates for various digital audio schemes (stereophonic signals).

2 Lossless audio coding

Lossless audio coding permits the compression of digital audio data without any loss in quality due to a perfect reconstruction of the original signal. It is a topic of high interest for both professional and customer applications. While modern *lossy* coding standards such as MP3 or AAC can achieve high compression ratios with transparent subjective quality, they do not preserve every single bit of the original audio data. Thus, lossy coding methods are not suited for editing or archiving applications, since multiple coding or post-processing can reveal originally masked distortions.

2.1 The Principle

Applying lossless entropy coding methods such as Lempel-Ziv, Huffman or arithmetic coding directly to the audio signal is not very efficient due to the long-time correlations and the high range of values. Therefore, conventional *text* compression tools such as Winzip or gzip fail in the case of digital audio data. A decorrelation stage, which eliminates the statistical dependencies within the signal, leads to an almost uncorrelated source which is easier to code. A common method to achieve such decorrelation is linear prediction, where each sample of the original signal is predicted from previous samples. The difference between original and predicted version is called the residual. If prediction works well, the residual is a decorrelated signal and has smaller values than the original. The residual is usually coded using simple entropy coding methods such as Rice codes, which are a special case of Huffman codes.

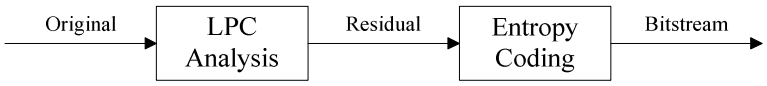

Fig. 1: Principle of lossless coding

Figure 1 shows the simplified diagram of a lossless encoder using linear predictive coding (LPC) followed by entropy coding. The corresponding decoder is shown in Figure 2.

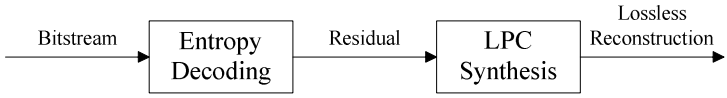

Fig. 2: Lossless decoding

The LPC synthesis filter basically adds the dependencies previously removed by the analysis filter, i.e. it correlates the signal again, leading to a lossless reconstruction of the original. Although the combination of analysis and synthesis filter is not lossless in general, lossless processing can be achieved if some basic conditions are observed.

2.2 MPEG-4 Audio Lossless Coding (ALS)

As an addition to the MPEG-4 audio standard [1], Audio Lossless Coding (ALS) will define methods for lossless coding of audio signals with arbitrary sampling rates, resolutions of up to 32 bit, and up to 256 channels [2]. In July 2003, the lossless codec from Technical University of Berlin was chosen as the first working draft. Since then, further improvements and extensions have been integrated [3][4]. MPEG-4 ALS is expected to become an international standard by the end of 2005.

The ALS encoder consists of several basic elements. Figure 3 shows the typical processing for one input channel of audio data.

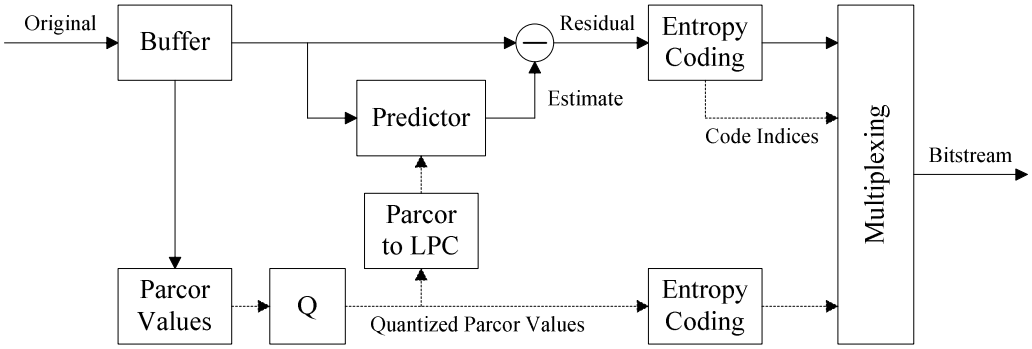

Fig. 3: MPEG-4 ALS Encoder

A buffer stores one block of input samples, and an optimum set of parcor coefficients is calculated for each block. The number of coefficients, i.e. the order of the predictor, can be adaptively chosen as well. The quantized parcor values are entropy coded for transmission, and converted to LPC coefficients for the prediction filter which calculates the prediction residual. The residual is entropy coded using different entropy codes. The indices of the chosen codes have to be transmitted as side information. Finally, a multiplexing unit combines coded residual, code indices, predictor coefficients and other additional information to form the compressed bitstream.

The decoder (Fig. 4) is significantly less complex than the encoder, since no adaptation has to be carried out. The decoder merely decodes the entropy coded residual and the parcor values, converts them into LPC coefficients, and applies the inverse prediction filter to calculate the lossless reconstruction signal.

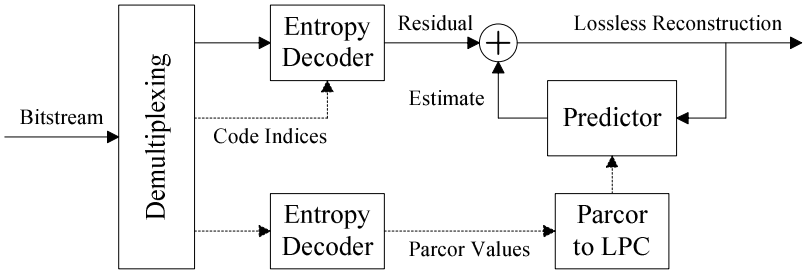

Fig. 4: MPEG-4 ALS Decoder

2.3 Compression Results

The following, the compression ratio is defined as

C = Original File Size / Compressed File Size

Table 2 compares the compression ratios of the popular open source codec FLAC and the MPEG-4 ALS codec from TU Berlin [3]. Higher values mean better compression. Almost 1 GB of stereo waveform data was used to measure the average compression ratios for different audio formats.

Audio Format	FLAC (open source)	MPEG-4 ALS (TU Berlin)
48 kHz / 16-bit	2.06	2.24
48 kHz / 24-bit	1.46	1.59
96 kHz / 24-bit	1.76	2.16
192 kHz / 24-bit	-	2.66

Table 2: Compression ratios for different audio formats
(192 kHz material is not supported by FLAC).

The compression ratio typically decreases with higher amplitude resolutions, but improves with higher sampling rates. The results also show that ALS outperforms FLAC for all formats, particularly for high-definition material (96 kHz / 24-bit).

2.4 Other Lossless Codecs

There are some proprietary lossless audio codecs that have not been included in the above comparison for different reasons. The commercial *MLP* codec is mainly used for the production of DVD-Audio content, but (probably due to its pricing) has not been established in other areas. The *Apple Lossless* codec only supports 16-bit audio data. For the 48/16 test material, it achieves a compression ratio of 2.03. Microsoft's *WMA Lossless* codec obviously converts any audio data to 44.1 kHz, 16-bit, before actually encoding it. Thus, lossless reconstruction is not possible for other sampling rates or resolutions.

2.5 Applications of Lossless Audio Coding

Examples for the use of lossless audio coding in general and MPEG-4 ALS in particular include both professional and consumer applications:

- Archival systems (broadcasting, studios, record labels, libraries)
- Studio operations (storage, collaborative working, digital transfer)
- High-resolution disc formats
- Internet distribution of audio files
- Online music stores (download)
- Portable music players

In the case online music stores, downloads of the latest CD releases will no longer be restricted to lossy formats such as MP3, AAC, or WMA. Instead, the consumer can purchase all tracks in full quality of the original CD, but still receive the corresponding files at reduced data rates.

3 Lossy but subjectively transparent audio coding

3.1 Speech and Audio Coding

First proposals to reduce wideband audio coding rates have followed those for speech coding [5]. Speech and audio coding are similar in that in both cases quality is based on the properties of human auditory perception. However, speech can be coded very efficiently because a speech production model is available, whereas nothing similar exists for audio signals (see Figure 5).

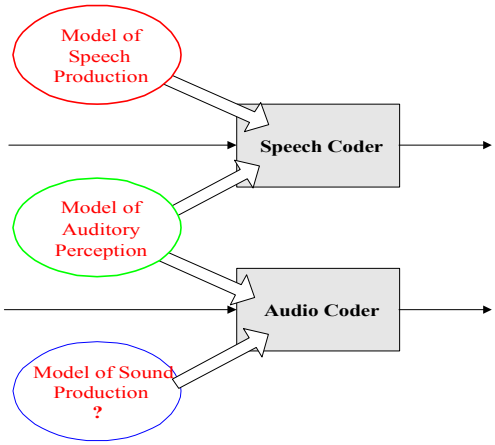

Fig. 5: Efficient speech and audio compression by employing models of perception and production

3.2 Quality Measures

As a measure of quality, the most popular subjective assessment method is the mean opinion scoring where subjects classify the quality of coders on an N-point quality scale. The final result of such tests is an averaged judgement called the *mean opinion score (MOS)*. Two 5-point adjectival grading scales are in use, one for signal *quality*, and the other one for signal *impairment*, and an associated numbering [5]. The 5-point ITU-R impairment scale of Table 3 is extremely useful if coders with only small impairments have to be graded.

Mean opinion score	Impairment scale
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 3: 5-point MOS impairment scale

3.3 Auditory Perception

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 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.

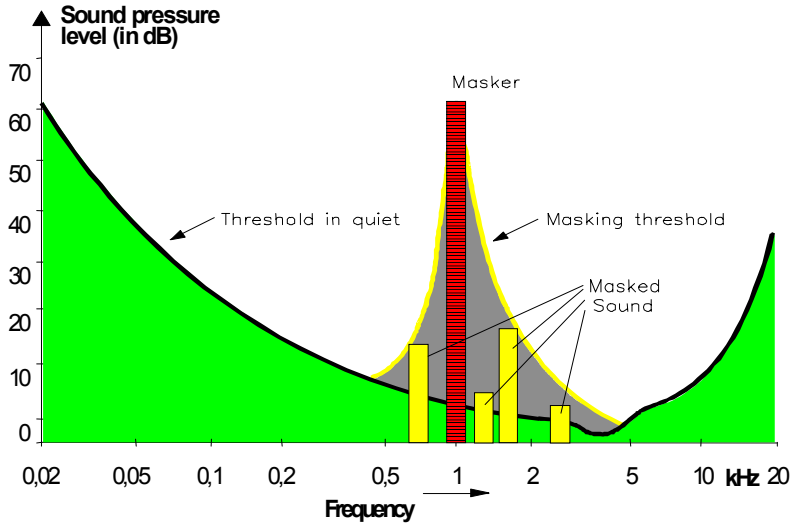

Fig. 6: Threshold in quiet and masking threshold
(Acoustical events in the gray areas will not be audible)

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 [6]. 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, of quantization noise, aliasing distortion, or of transmission errors. The masking threshold varies with time. It depends on the sound pressure level (SPL), the frequency of the masker, and on the characteristics of masker and maskee. Take the example of the masking threshold for the SPL = 60 dB narrowband masker in Figure 6: around 1 kHz the five maskees (one of which is hidden behind the masker) 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 easier 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. 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 6. The distance between the level of the masker and the masking threshold is called *signal-to-mask ratio* (SMR). Within a critical band, coding noise will not be audible as long as its signal-to-noise ratio SNR(m), the signal-to-noise ratio resulting from an m-bit quantization, is higher than its SMR.

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

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 or from another reference.

In practical designs, we cannot go to the limits of just noticeable distortion, since post-processing of the audio signal (e.g., filtering in equalizers) 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.

3.4 Perceptual Coding

Digital coding at high bit rates is dominantly waveform-preserving, i.e., the amplitude-versus-time waveform of the decoded signal approximates that of the input signal. However, at lower bit rates, facts about the production and perception of speech and 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 (i) remove redundant components of the source signal by exploiting correlations between its samples and (ii) remove components which are irrelevant to the ear. Irrelevancy manifests itself as unnecessary amplitude or frequency resolution; portions of the source signal which are masked need not 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 shift coding noise to frequency bands where that noise is not of perceptual importance. The noise shifting must be dynamically adapted to the actual short-term input spectrum in accordance with the signal-to-mask ratio and 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 7 depicts the structure of a *perception-based coder* that exploits auditory masking. The encoding process is controlled by the signal-to-mask ratio (SMR) vs. frequency curve from which the necessary amplitude resolution (and hence the bit allocation and rate) in each critical band is derived. The SMR is the ratio of the sound pressures of signal and its masking threshold within a given frequency band. It is 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 allows for a dynamic control by such perceptual information. Frequency domain coders are of particular interest since they offer a direct method for noise shaping.

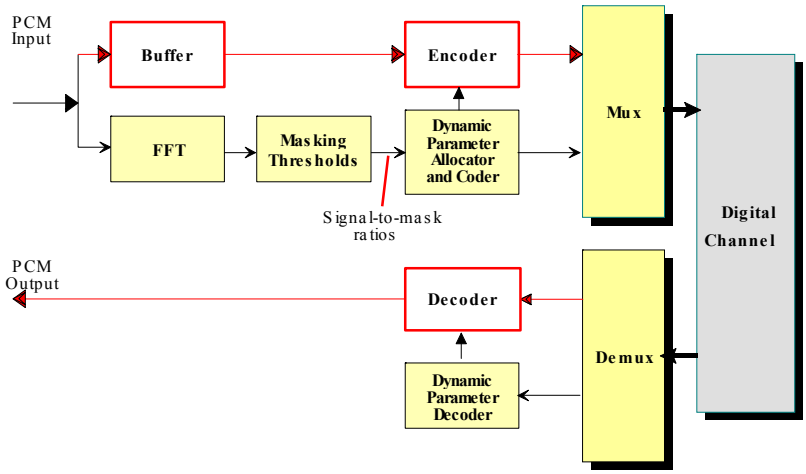

Fig. 7: Block diagram of perception-based coders

3.5 ISO/MPEG Audio Coding

The MPEG-1 audio coding standard [7][8] 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 at and above 128 – 256 kbit/s for many types of music.

The structure of MPEG coders follows that of perception-based coders. 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 *psychoacoustical model.* 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.

The Layer III hybrid filterbank approach has become quite popular, in particular in Internet applications (MP3). The structure of the switched hybrid filterbank is given in Figure 8. This filterbank achieves a higher frequency resolution closer to critical band partitions by subdividing the 32 subband signals 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.

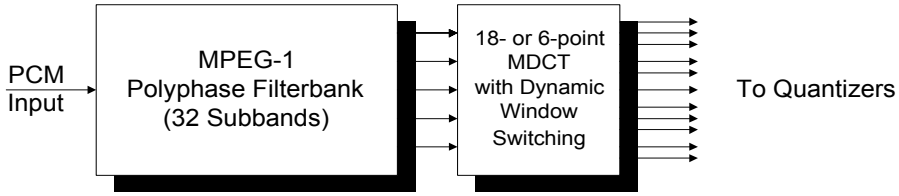

Fig. 8: Hybrid filterbank of MPEG-1 Layer III encoder

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.

4 MPEG Advanced Audio Coding

The MPEG-2/MPEG-4 AAC standard employs high resolution filter banks, prediction techniques, and noiseless coding [9][10]. 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 an 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. The filterbank is a 1024-point modified discrete cosine transform (due to a 50% overlap, the transform is taken over 2048 windowed samples), the perceptual model is taken from MPEG-1.

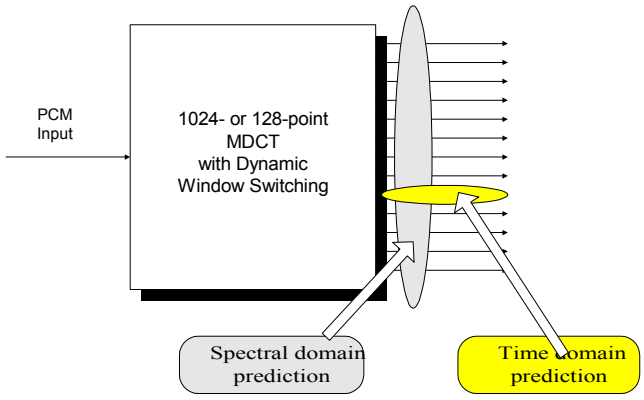

Fig. 9: Spectral and time domain prediction in MPEG-2/4 Advanced Audio Coding (AAC)

The temporal noise shaping (TNS) tool plays an important role in improving the overall performance of the coder (see Figure 9). 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 a high predictability in the spectral domain. (Recall that “peaky” spectra lead to a high predictability in the time domain). Therefore the TNS tool controls the time dependence of the quantization noise. 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 9 again). Second-order backward-adaptive predictors are used for this purpose. Finally, for quantization an iterative method is employed so as to keep the quantization noise in all critical bands below the global masking threshold.

The MPEG-2/4 AAC standard offers high quality at lowest possible bit rates, it will therefore find many applications, both for consumer and professional use. Figure 10 shows MOS differences, with diffscore = 0 for the compact disc reference. For example, the AAC coder operating at 128 kbit/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 III coder (MP3 128) has a diffscore near −1. Note also, that, at a rate of 96 kbit/s, the AAC main coder performs better than the MPEG-1 Layer II coder at twice the rate (MP2 192).

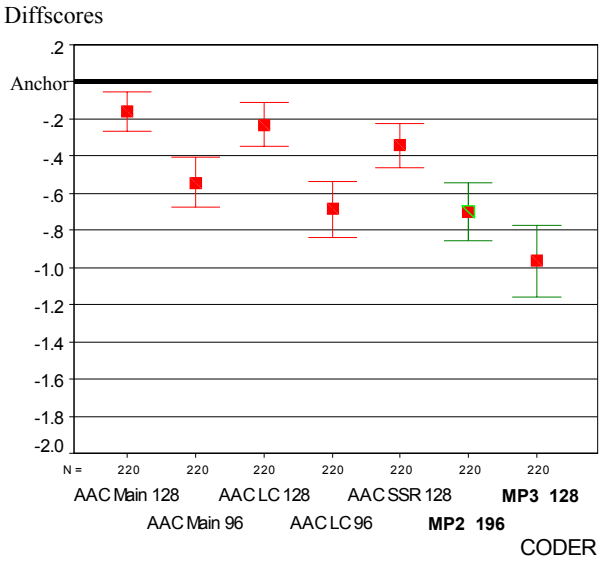

Fig.10: Subjective quality of AAC and MPEG-1 audio coders [11].

5 MPEG-4 Audio Coding

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 has 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 supports new forms of communications, in particular for Internet and Multimedia applications and in Mobile Communications.

MPEG-1 and MPEG-2 have some main disadvantages: they offer only a very limited interaction and control over the presentation and configuration of the system. In addition, an integration of natural and synthetic content is difficult, and an 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 with objects with scene is possible. For example, it will be possible to associate a Web address to a person in a scene.

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 11 describes an audiovisual

scene with a number of audio „objects“: the noise of an incoming train, an announcement, a conversation, and background music.

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.

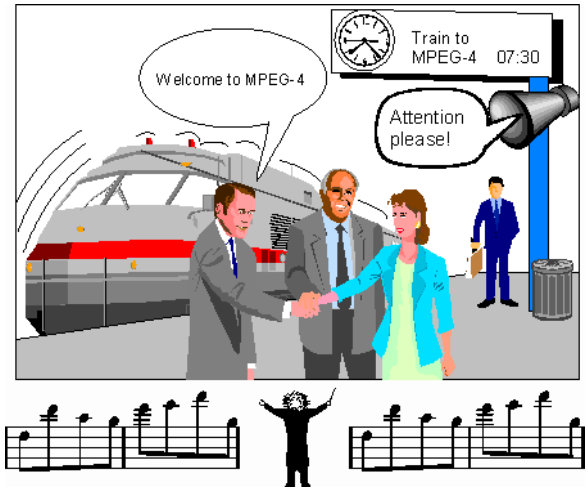

Fig. 11: Audiovisual scene [12]

In order to represent, integrate and exchange pieces of audio-visual information, MPEG-4 offers tools which can be combined to satisfy specific user requirements. A number of such configurations have been 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 kbit/s)
- a transform-based coding scheme for higher bit rates (up to 64 kbit/s).

MPEG-4 not only offers simple means of manipulation of coded data such as time scale control, pitch change, but also a flexible access to coded data and subsets thereof, i.e. scalability of bit rate, bandwidth, complexity, and of 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 offers also an interface between TTS and facial animation for synthetic face models to be driven from speech (“*Talking Heads*”).

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. The following Figure 12 indicates the range of bit rates offered by the new standard.

Transform-based audio coders show a very good performance at bit rates down to 16 kb/s, whereas speech coder perform clearly better at rates between 2.4 kb/s and 16 kb/s. Currently a number of speech coders is available with good performance in that range of bit rates. Both coder classes, however, do not offer solutions for audio coding at 4 - 16 kb/s.

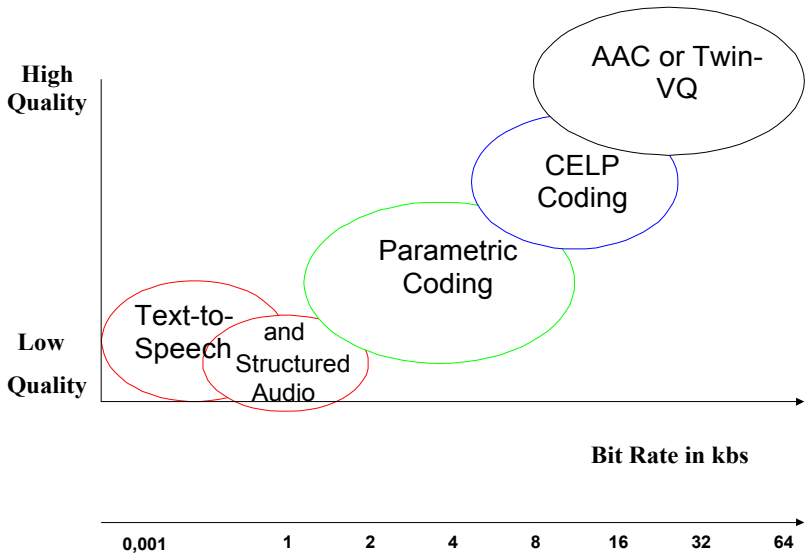

Fig. 12: Range of audio quality and bit rates in MPEG-4.

6 References

[1] ISO/IEC 14496-3:2001, “Information technology - Coding of audio-visual objects - Part 3: Audio”, International Standard, 2001.

[2] ISO/IEC JTC1/SC29/WG11 N7016, “Text of 14496-3:2001/FPDAM 4, Audio Lossless Coding (ALS), new audio profiles and BSAC extensions”, 71st MPEG Meeting, Hong Kong, China, January 2005.

[3] T. Liebchen, Y. Reznik: “Improved Forward-Adaptive Prediction for MPEG-4 Audio Lossless Coding”, 118th AES Convention, Barcelona, 2005.

[4] T. Liebchen, Y. Reznik, T. Moriya, D. Yang: “MPEG-4 Audio Lossless Coding”, 116th AES Convention, Berlin, 2004.

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

[6] E. Zwicker and R. Feldtkeller, „Das Ohr als Nachrichtenempfänger“. Stuttgart: S. Hirzel Verlag, 1967.

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

[8] P. Noll, “MPEG Audio Coding Standards“, IEEE Signal Processing Magazine, Sept. 1997.

[9] ISO/IEC JTCl/SC29, “Information Technology - Generic Coding of Moving Pictures and Associated Audio Information - IS 13818 (Part 7, Audio)” , 1997.

[10] M. Bosi et al, “ISO/IEC MPEG-2 Advanced Audio Coding“, J. Audio Eng. Soc., Vol 45, No. 10, S. 789 - 814, 1997.

[11] D. Meares, K. Watanabe, and E. Scheirer, “Report on the MPEG-2 AAC Stereo Verification Tests”, MPEG Document N2006 (Febr. 98).

[12] ISO/IEC/JTC1/SC29, MPEG Document N2431 (Oct. 98).