PROCEEDINGS INTERNATIONAL COST 254 WORKSHOP ON INTELLIGENT COMMUNICATION TECHNOLOGIES AND APPLICATIONS, NEUCHATEL, SCHWEIZ, IN PRINT, 1999

SPEECH AND AUDIO CODING FOR MULTIMEDIA COMMUNICATIONS

Peter Noll

Technische Universität Berlin
Einsteinufer 25, 10587 Berlin, Germany
Phone: +49 30 3142 3326; Fax: +49 30 3142 2514
Email: noll@ee.tu-berlin.de

ABSTRACT

Keywords: *audio coding, speech coding, MPEG, noisy channels.*

We have seen rapid progress in high-quality compression of telephone speech and wideband speech signals. 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. 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.

The paper will explain basic approaches to such compressions, with concentration on existing and upcoming international standards. As typical signal classes we shall consider *telephone speech*, *wideband speech*, and *wideband audio signals* all of which differ in listener expectation of offered quality. The main motivations for low bit rate coding are outlined as well as basic and network-related requirements. It will become obvious that speech and audio coders must be both source-specific and hearing-specific to perform adequately at low bit rates.

1. INTRODUCTION

In current digital telephony and audio-only services, speech and audio signals are not always compressed. The conventional digital format for speech and audio signals is PCM, with sampling rates and amplitude resolutions (PCM bits per sample) as given in Table 1. Although high bit rate channels and networks become easier accessible, low bit rate coding of speech and audio signals has retained its importance. The main motivations for compression 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 or packet-oriented Internet/WWW networks. In this latter application the user is typically connected to the service provider via a, say, 28.8 kb/s modem or an 64 kb/s ISDN access. In addition, future digital communications services such as digital video broadcasting, storage, videotelephony, video- and audiographic conferencing, news gathering and interactive multimedia services (information, training, entertainment) will have a substantial speech or audio component.

	Frequency range in Hz	Sampling rate in kHz	PCM bits per sample	PCM bit rate in kb/s
Telephone speech	300 - 3,400	8	8	64
Wideband speech	50 - 7,000	16	8	128
Wideband audio (stereo)	10 - 20,000	44,1 [1]	2 x 16	2 x 705

Table 1: Basic parameters of telephone speech, wideband speech, and wideband audio signals

Even text, graphics, fax, still images, email documents, etc. will gain from voice annotation and audio clips. In all these cases, data compression will be of increasing importance to provide economically acceptable overall bit rates depending on network characteristics, potential applications, and coding goals such as high quality and low complexity.

First proposals to reduce wideband audio coding rates have followed those for speech coding [1]. Differences between audio and speech signals are manifold,

[1] CD format. Other sampling rates are 32 kHz , 48 kHz, and 96 kHz. Amplitude resolution goes up to 32 b/sample.

however: audio coding implies higher sampling rates, better amplitude resolution, higher dynamic range, larger variations in power density spectra, differences in human perception, stereo and multichannel audio signal presentations, and, finally, higher listener expectation of quality. Indeed, the high quality of the Compact Disc with its 16-b/sample 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. However, speech can be coded very efficiently because a speech production model is available, whereas nothing similar exists for audio signals (see Figure 1).

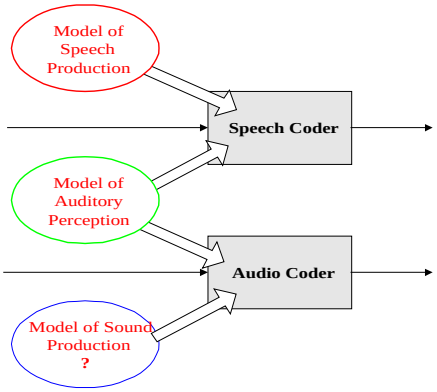

Fig. 1: Efficient speech and audio compression by employing models of perception and production

In the following, we will explain basic approaches to speech, wideband speech, and audio bit rate compressions [2 - 5]. We will concentrate on worldwide source coding standards because such standards are beneficial for consumers, service providers, and manufacturers. It will become obvious that the use of our knowledge of speech production and auditory perception helps minimizing perception of coding artifacts and leads to *source-specific* and *hearing-specific* coders.

Basic requirements in the design of low bit rate speech or audio coders are firstly, to retain a high quality of the reconstructed signal with robustness to variations in spectra and levels. Secondly, robustness against random and bursty channel bit errors and packet losses and a graceful degradation of quality with increasing bit error rates in mobile radio and broadcast applications are required. Thirdly, low complexity and power consumption of the codecs are of high relevance. Additional network-related requirements are low encoder/decoder delays, robust cascading of codecs, and transcodability.

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 [1]. The 5-point ITU-R impairment scale of Table 2 is extremely useful if coders with only small impairments have to be graded.

The advantage of MOS values is that different impairment factors can be assessed simultaneously and that even small impairments can be graded. On the negative side, experience has shown that MOS values can vary with time and from listener panel to listener panel; it seems to be very difficult to duplicate test results at a different test site. In the case of audio signals, MOS values depend strongly on the selected test items; there are significant differences between MOS values obtained with "average" audio material and those obtained with more critical test items.

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

Table 2: 5-point MOS impairment scale

For all these reasons care is needed in comparing results between different experiments. On the other hand, it should be pointed out that the MPEG and ITU-R audio listening tests, carried out under very similar and carefully defined conditions with experienced listeners, have shown very similar and stable evaluation results.

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.

Simultaneous masking is a frequency domain phenomenon where a low-level signal (the maskee) can be made inaudible (masked) by a simultaneously oc-

curring 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, in the context of source coding also known as *threshold of just noticeable distortion* (JND) [7], 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 2: 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 2. 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.

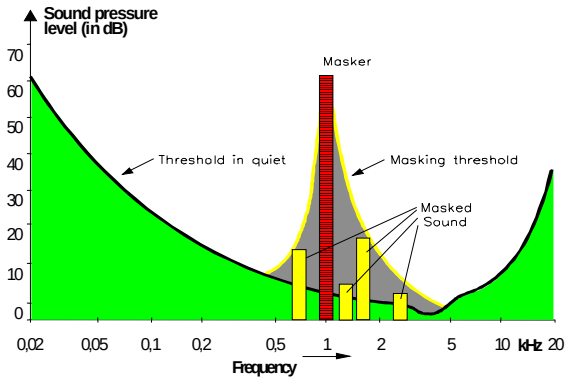

Fig. 2: Threshold in quiet and masking threshold
(Acoustical events in the gray areas will not be audible)

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 many cases the 8-b/sample PCM telecommunications standard and the Compact Disc's 16-b/sample PCM serve, respectively, as references for speech and audio.

In practical designs, we cannot go to the limits of just noticeable distortion, since postprocessing 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.

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 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) 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.

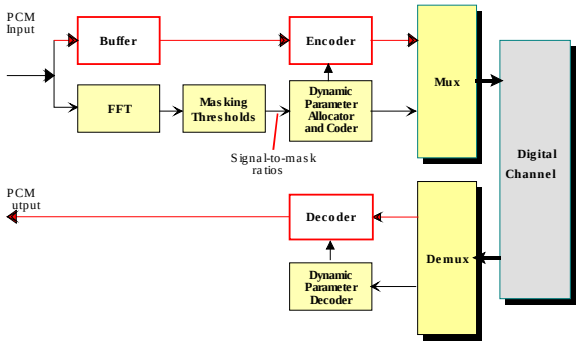

Fig. 3: Block diagram of perception-based coders

5. SPEECH CODING

Bit rates in the digital representation of 8-kHz sampled telephone-bandwidth speech (300 - 3400 Hz) can vary from, say, one kb/s (segment vocoders) to 128 kb/s (16 bit linear PCM), depending on the application and on end-user expectations of signal quality. Today's most widely used speech coding standards are 64-kb/s A-law and μ-law PCM [ITU-T G.711] and 32-kb/s adaptive differential pulse code modulation (ADPCM) [ITU-T G.726]. Both standards, which serve also as references in speech codec evaluation, achieve high quality with MOS scores above 4.2. Activities in the last years have concentrated on low bit rate speech coding for mobile radio and cordless systems, for voice mail systems, and for multipoint teleconferencing. Recently a speech coder has been standardized [ITU-T G.729] which offers almost transparent quality at 8 kb/s.

Most low bit rate coding schemes in the interesting range between 2 to 16 kb/s are based on various forms of analysis-by-synthesis coding. They make use of a speech generation model where speech is represented by an excitation signal feeding an (typically frame-wise) adaptive filter, the so-called *LPC synthesis filter*. The differences are in the ways of including pitch periodicity, of representing the residual signal, and of describing the spectral envelope.

5.1 ADPCM - the ITU-T and DECT Standard

In adaptive differential pulse code modulation schemes, it is the difference between each input sample and its adaptively predicted value that is used as excitation in its quantized form. Since the prediction error signal (also called the residual) has much less power and much less correlation than the source signal, it can be quantized and transmitted more easily.

The ITU-T G.726 32-kb/s ADPCM coder is well-established; it is used both in ITU-T networks and in digital cordless telephony applications. The Digital European Cordless Telephony (DECT) is an important example. ADPCM supports also bit rates of 16, 24, and 40 kb/s to be employed in systems with variable bit rates:

- Firstly, ADPCM can be used in (TASI-like) digital circuit multiplication equipment to increase the number of simultaneous voice calls transmitted over a link of given capacity by exploiting silences in two-way telephone connections. The coder is allowed to reduce its number of bits per sample in order to be able to open new channels during overload conditions.
- Secondly, a modification, called *Embedded ADPCM* supports speech transmission over packet-oriented transmission networks. The network is allowed to strip off least significant bits to avoid or to shorten momentary overloads in network nodes which would otherwise cause packet losses.

5.2 Analysis-by-Synthesis Coding

A very powerful tool in linear predictive coding at low bit rates in the interesting range of 1 to 16 kb/s is the analysis-by-synthesis approach (see Fig.4). Analysis-by-synthesis predictive coders have as a common characteristic that it is not, as in ADPCM, the quantized residual signal that is transmitted to the decoder.

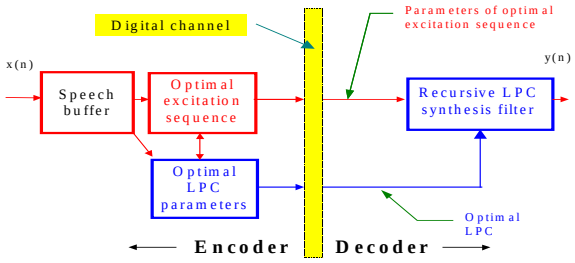

Fig. 4: Block diagram of analysis-by-synthesis coders

Instead, a number of candidate excitation signals are available in the encoder, e.g., organized in a codebook. The excitation signals can be modeled in different ways. Each excitation can be (i) a small number of optimally placed impulses (multipulse-excitation), (ii) a small number of regularly spaced impulses (regular pulse excitation), or (iii) a complete sequence of impulses (code- or vector-excitation). Combinations are possible. Today, the codebook-based approach is commonly known as *code-excited linear predictive coding* (CELP).

A feedback loop allows all possible candidate excitation signals (also called vector excitations) to locally decode corresponding candidate output signals

by feeding them through an encoder-based LPC synthesis filter. Since the coding approach requires synthesis during error analysis, it is referred to as *analysis-by-synthesis* technique; its principle structure is shown in Figure 4. Each output signal is compared with the buffered speech input sequence. The best excitation is defined to be that candidate excitation signal that minimizes the coding error, the difference between the buffered speech input sequence and its reconstructed version. Since the coding errors are actually determined, we have control over distortions in the reconstructed speech, and it is easy to incorporate models of human auditory perception to compute perceptually meaningful distortion measures.

The address of the optimum candidate excitation signal is transmitted to the receiver, which has a copy of the codebook, so that the best reconstruction sequence can be synthesized. Assume that a codebook contains 1024 candidate excitation signals, each consisting of 40 samples (5 ms of speech). A 10-bit address informs the decoder, the resulting bit rate is only 0.25 b/sample.

The LPC synthesis filter and a codebook scaling factor are updated for each buffered speech block of, say, 160 samples (20 ms of speech). Less than 40 b/block are necessary to transmit update information about the filter coefficients to the receiver. This side information amounts to 2000 b/s.

5.3 Speech Coding Standards

The following tables list current speech coding standards. Coders operating at 16 kb/s and below are all based on the CELP principle. The differences are mainly in the way the vector excitations are defined and updated. A careful design of vector excitations can significantly reduce the computational load for finding the optimum excitation signal for a given speech frame bit rates. Note that all bit rates given in the tables do not include the overhead bit rate for channel coding. We finally note that present research focuses on transparent transmission of speech at a rate of 4 kb/s.

Type of Coder	Standard or Product	Bit Rates in kb/s
LPC-10E	FS 1015	2.4
CELP	FS 1016	4.8
MELP	FS 1017	2.4

MELP = mixed excitation LP coding

Table 3: Speech coding for US Fed. secure voice

Type of Coder	ITU-T Standard	Bit Rates in kb/s
PCM	G.711	64
ADPCM	G.726	16, 24, 32, 40
Embedded ADPCM	G.727	16, 24, 32, 40
Low-Delay CELP	G.728	16
ACELP	G.729	8
ACELP/ MP-LPC	G.723.1	5.3 / 6.3
Not defined yet	ITU-T G.xxx	4

Table 4: Speech coding for telecommunications

Type of Coder	Standard/ Product	Bit Rates in kb/s	Region
RPE-LTP	GSM 06.10	13	Europe
ACELP	GSM EFR, PCS 1900 EFR	12.2	Europe, USA
ACELP	EFR IS-641	7.4	USA
QCELP	IS-96*	≤8.5	USA
VSELP	IS-54*	7.95	USA
VSELP	GSM Half Rate 06.20	5.6	Europe
VSELP	PDC	6.7	Japan
PSI-CELP	PDC Half Rate	3.45	Japan

RPE = regular pulse excitation; LTP = long-term prediction; ACELP = algebraic CELP; EFR = enhanced full rate; SELP = vector sum excited LP coding; PSI = pitch-synchronous innovation CELP
*These coders will be replaced by the EFR coder IS-641

Table 5: Speech coding for mobile radio

Coding scheme	Bit Rate in kb/s	MOS value
FS 1015 LPC-10E	2.4	2.6
FS 1017 MELP	2.4	3.3
GSM Enhanced Half Rate	5.6	3,3
ACELP	8.0	4.0
GSM Enhanced Full Rate	12.2	3,8
GSM (RPE-LTP)	13.0	3.5
Low Delay CELP	16.0	4.1
ADPCM (DECT)	32.0	4.1
PCM (ISDN)	64.0	4.3

Table 6: MOS values of various speech coding schemes

6. NOISY CHANNELS

We have already stated, that one basic requirement in speech and audio coding is robustness against random and bursty channel bit errors and packet losses. Graceful degradations of quality with increasing bit error rates are required in mobile radio and broadcast applications. Major impairments in mobile and cell-oriented communications come from additive noise, from multipath reception, and from cell losses. In addition, there are a great many other dimensions to speech coder quality that need to be investigated. Examples are

- Robustness to background noise

- multiple encodings
- different coders in tandem
- ability to carry signaling tones, voiceband modem and fax signals
- insensitivity to automatic speech recognition and/or speaker verification systems.

Errors affect audio and speech in various ways. As a result of errors squeaks and squawks, chirps, clicks or pops or other strange artifacts may occur. Error robustness can be obtained by algebraic channel coding; source-dependent ("unequal") error protection (UEP) can improve further overall performance. Even then, we have to cope with breakdowns despite algebraic error protection, and concealment techniques have to be employed. As an example, we consider cell losses in PCM transmission of speech. In case of lost cells, we need a substitution signal. This can be white or colored noise, a simple cell repetition, or a pitch-synchronous signal extrapolation strategy. Figure 5 shows corresponding signal waveforms.

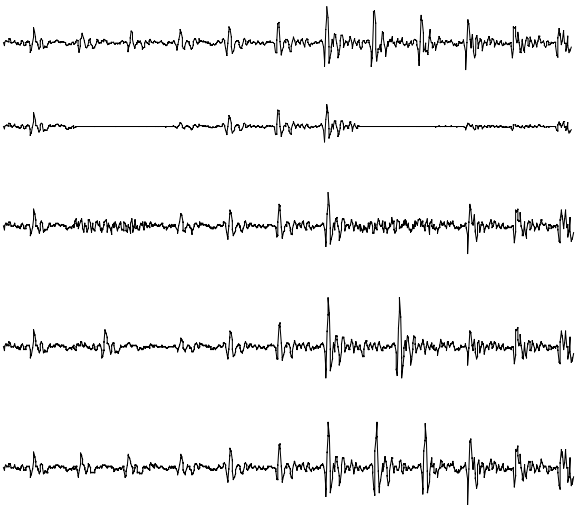

Fig. 5: Substitution of lost cells. a) no loss b) zeroing c) noise substitution d) block repetition e) pitch-synchronous substitution

At the TU Berlin, we have developed a very robust substitution technique, that uses the last received cell to compute an estimate of the power spectral estimate of the following lost cell (see Fig. 6). In case of a lost cell, a residual signal is derived from the last received cell in order to synthesize the speech output. Note, that it is not the speech waveform itself that is substituted, but the residual signal.

Substitution techniques are more complex in case of source-coded speech. As an example, the new ITU-T G.729 8 -kb/s ACELP coder reconstructs lost 10 ms – frames on the basis of information from earlier frames. The substitution depends on the identification of the last frame either as voiced or unvoiced (Fig. 7).

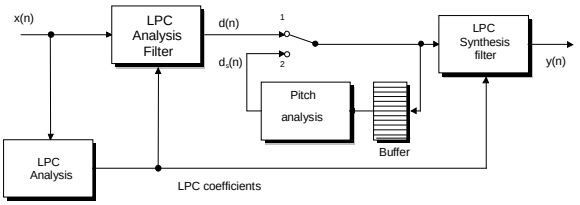

Fig. 6: Asynchronous Transfer Mode (ATM): Pitchsynchronous speech substitution via residual signal.

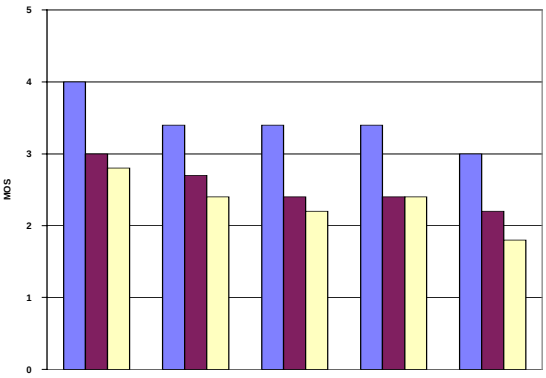

Fig. 7: ITU-T G.729 8 kb/s-coder:
MOS performance; Frame losses: 0 – 5%
(From left to right: no losses;
3% frame loss, isolated; 3% frame loss, bursty;
5% frame loss, isolated; 5% frame loss, bursty)
Left column: clean speech; center column:vehicle noise; right column: road noise.

7. WIDEBAND SPEECH CODING

Speech coding with a bandwidth wider than that offered in telephony results in major improvements in represented speech quality, it is particularly useful in loudspeaker telephony and videoconferencing. The [ITU-T G.722] wideband speech coding algorithm supports a bandwidth of 7 kHz with a sampling rate of 16 kHz and bit rates of 64, 56, and 48 kb/s. The codec has an overall delay of around 3 ms, small enough to cause no echo problems in telecommunication networks [8,9]. In this ITU-T wideband speech coder, a subband splitting, based on two identical quadrature mirror (bandpass) filters (QMF), divides the 16 kHz sampled 14-bit PCM representation of the wideband input signal into two critically subsampled (8-kHz-sampled) components, called low subband and high subband.Medium to high quality coding with the G.722 wideband speech coder is provided by ADPCM encoding of the two subband sequences, where the low and high subband ADPCM coders use 6-b/sample and 2-b/sample quantizers, respectively. In the low subband the signal resembles the narrowband speech signal in most of its properties. A reduction of the quantizer resolution to 5 or 4 b/sample is possible to support transmission at lower overall rates or to transmit auxiliary data at rates of 8 or 16 kb/s.

The uncoded source has the same MOS score of around 4.2 as the 64-kb/s G.722 wideband speech

coder, implying no measurable difference in subjective quality, and we notice a graceful degradation of subjective quality with decreasing bit rate to a MOS score of 3.8 at 48 kb/s. The coder provides also a good to fair quality for coding of audio signals. Table 7 compares MOS values for speech and audio signals at two bit rates. The resulting subjective quality is partly limited by the restriction on the bandwidth. The coder is also error-robust; its subjective performance at a bit error rate of 0.1% has still a *fair* rating, close to the score of 64-kb/s PCM.

Bit rate in kb/s	MOS value for speech signals	MOS value for audio signals
48	3.8	3.2
64	4.2	3.9

Table 7: MOS performance of ITU-R G.722 coder (7 kHz bandwidth) [8]

Bit rates lower than 48 kb/s, the minimum bit rate of the G.722 standard, are highly desirable in applications that use narrowband ISDN. ITU-T studies a number of approaches for coding of wideband audio at bit rates of down to 8 kb/s.

8. ISO/MPEG AUDIO CODING

The MPEG-1 audio coding standard [5, 10, 11] 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 – 192 kb/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 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. 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 structure of *Layer I and Layer II* coders is shown in the following Figure 8. 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. Again, we will skip details of this coder which has become quite popular, in particular in Internet applications (MP3). The structure of the switched hybrid filterbank is given in Figure 9. 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, resp., 12 or 36 subband samples.

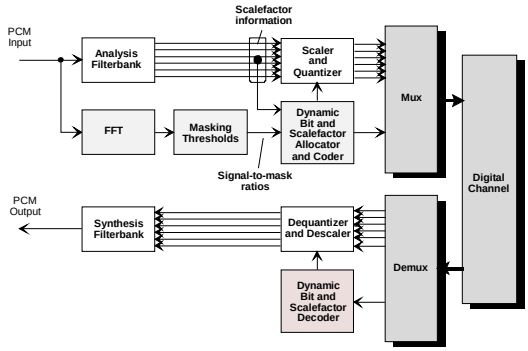

Fig. 8: Structure of MPEG-1/Audio encoder and decoder, layers I and II

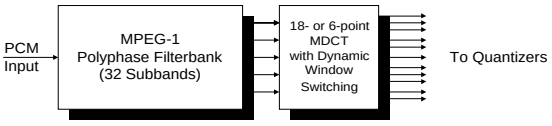

Fig. 9: Hybrid filterbank of MPEG-1 Layer III encoder

8.1 Example: Digital Audio Broadcast (DAB)

The Digital Audio Broadcast (DAB) standard employs the MPEG-1 Layer II audio coder with its range of bit rates between 32 and 384 kb/s; it supports also Layer III coding. The transmission scheme was especially designed to allow a mobile reception under additive noise and multipath propagation conditions. Poor reception conditions exist not only at the borders of service coverage areas, but also within the coverage area due to shadowing effects etc. DAB is a multicarrier system employing an efficient unequal bit error protection strategy. Therefore DAB provides an

excellent subjective 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 Fig. 10).

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 ¼, i.e., 300% redundancy is added to the bitstream. Less protection is obtained by puncturing code bits of the rate ¼ mother code. The punctured code bits are not transmitted and they are treated as erased bits in the decoding process. Therefore the decoder needs not to be modified in case of puncturing. The DAB specification provides 24 code rates from 8/9 to 8/32.

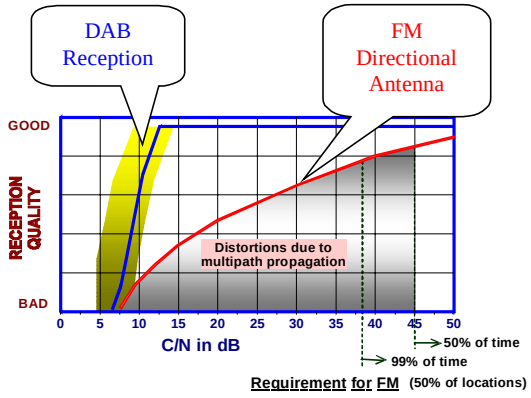

Fig. 10: Comparison of analog FM and Digital Audio Broadcast [12]
[Raleigh channel, rural I model, speed of mobile 50 km/h].

Fig. 11 shows the *residual* BER vs. C/N - performance for code rates between 2/3 and 1/3 [12]. These curves have been obtained from simulations employing Rayleigh 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 ¾ 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. 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, 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.

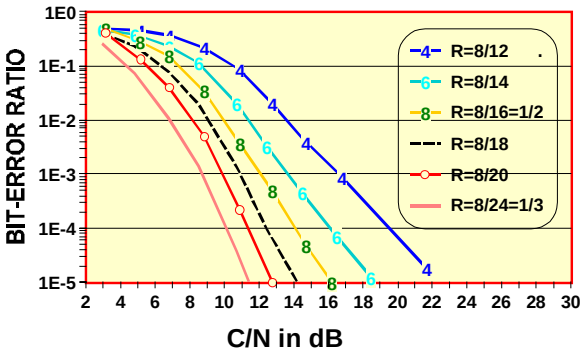

Fig. 11: Residual bit error rates [12]
[Raleigh channel, rural 1 model, speed of mobile: 50 km/h]

8.2 MPEG Advanced Audio Coding

The MPEG-2 AAC standard employs high resolution filter banks, prediction techniques, and noiseless coding [13, 14]. 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[1], the perceptual model is taken from MPEG-1.

The temporal noise shaping (TNS) tool plays an important role in improving the overall performance of the coder (see Figure 12). 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 predictibility in the spectral domain. (Recall that „peaky“ spectra lead to a high predictibility 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 12). Second-order backward-adaptive predictors are used for this purpose. Finally, for quantization an iterative method is

[1] Due to a 50% overlap, the transform is taken over 2048 windowed samples.

employed so as to keep the quantization noise in all critical bands below the global masking threshold.

The MPEG-2 AAC standard offers high quality at lowest possible bit rates, it will therefore find many applications, both for 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).

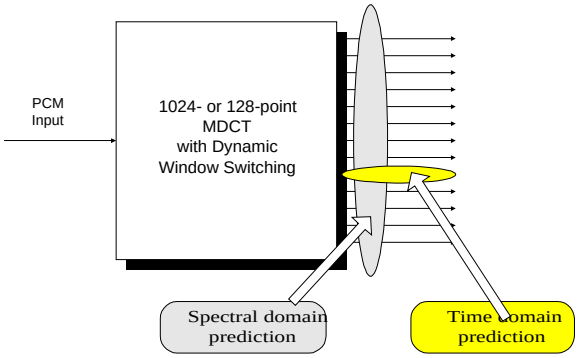

Fig. 12: Spectral and time domain prediction in MPEG-2 Advanced Audio Coding (AAC)

8.3 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

- Internet
- Multimedia
- Mobile Communications.

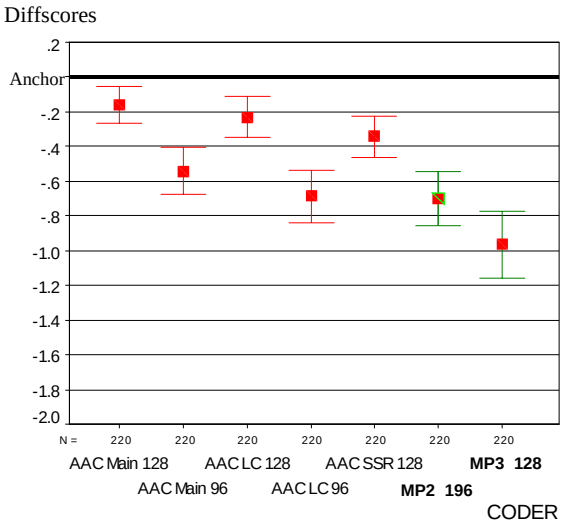

Fig.13: Subjective quality of AAC and MPEG-1 audio coders [15].

Indeed, 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: Fig.14 describes an audiovisual scene with a number of audio „objects“: the noise of an incoming train, an announcement, and a conversation. Fig. 15 lists the corresponding objects in detail.

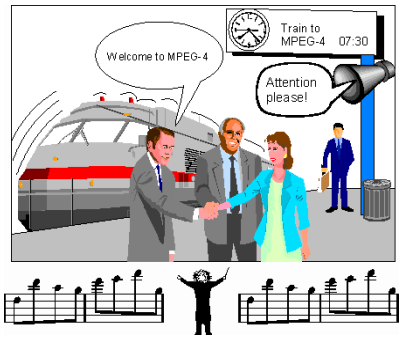

Fig. 14: Audiovisual scene [16]

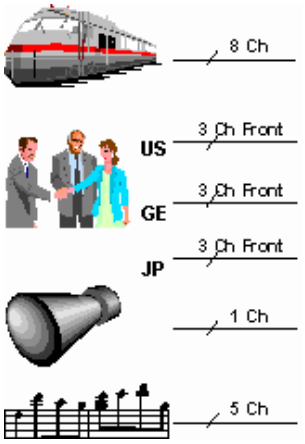

Fig.15: Audio channels for the audiovisual scene of Fig. 14 [16]

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 audio-visual 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.

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 |17|. A number of such configurations has 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 kb/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 16 indicates the range of bit rates offered by the new standard.

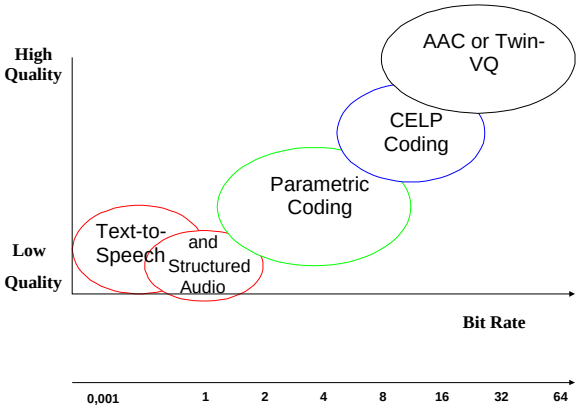

Fig. 16: Range of audio quality and bit rates in MPEG-4.

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. The MPEG-4 standard contains a concept for coding music of low complexity content [16]. 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. Fig. 17 shows the block diagram of such object-based audio encoder.

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 [16].

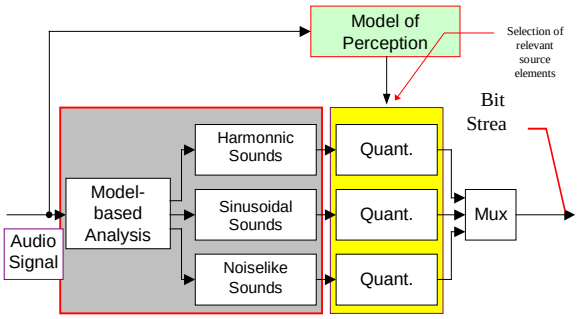

Fig. 17: Object-based audio encoder (After [17]).

9. CONCLUSION

A wide range of standardized coders is available for coding speech, wideband speech, and wideband audio signals with applications in mobile radio systems, satellite-based communications, cordless digital telephone systems, Internet telephony, etc. Important attributes of these coders are quality, bit rate, delay, and complexity. The application engineer determines which of these attributes are most important. It has been shown that such coders can be also robust against channel errors such as bursty errors in mobile radio, and cell losses in cell-oriented transmission such as ATM systems and Internet/WWW services. Error concealment techniques will play a significant role, since, due to the lack of available bandwidth, traditional channel coding techniques may not be able to sufficiently improve the reliability of the channel.

Current coders are controlled by psychoacoustical models which may be improved thus leaving room for an evolutionary improvement of codecs. In particular, different psychoacoustical models can be used ranging from very simple ones (including none at all) to very complex ones based on quality and implementability requirements. The MPEG-2 Advanced Audio Coding (AAC) standard performs remarkably well and may become popular both for stereophonic and for multichannel audio coding.

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 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.

10. 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. 94.

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

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

[7] 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.

[8] P. Mermelstein, "G.722, A New ITU-T Coding Standard for Digital Transmission of Wideband Audio Signals," *IEEE Commun. Mag.*, pp. 8-15, Jan. 1988.

[9] N.S. Jayant, J.D. Johnston and Y. Shoham, "Coding of wideband speech," *Speech Communication* 11, pp. 127 - 138,1992.

[10] 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.

[11] 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.

[12] C. Weck “The Error Protection of DAB,“ AES Conference “DAB - The Future of Radio“, London,1995

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

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

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

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

[17] H. Purnhagen, B. Edler, C. Ferekidis, “Object-based aAnalysis/Synthesis Audio Coder for Very Low Bit Rates“, 104th Audio Engineering Society Convention (AES), preprint, Amsterdam, 1998.