ISO/MPEG AUDIO CODING: STATUS AND TRENDS

Peter NOLL

Technische Universität Berlin
Institut für Fernmeldetechnik · Sekr. FT 5 ·
Einsteinufer 25 · 10587 Berlin

Germany

Tel.: xx49 (030) 314-2 33 26 · Telefax: xx49 (030) 314-2 25 14
email: noll@ftsu00.ee.tu-berlin.de

ABSTRACT

The Moving Pictures Expert Group within the International Organization of Standardisation (ISO/MPEG) has provided a generic source coding method for storing audio-visual information on digital storage media /1/. The recent ISO/MPEG Phase 1 audio coding standard is the first international standard in the field of digital audio compression and can be used both for audio-visual and audio-only applications. It is about to become a universal standard in many application areas with totally different requirements in the fields of consumer electronics, professional audio processing, telecommunications, and broadcasting. The proposed coder provides a subjective quality that is equivalent to compact disc (CD) quality (16 bit PCM) at a rate of 128 kb/s per mono channel, and that is close to CD quality at 64 kb/s per mono channel.
The evolving ISO/MPEG Phase 2 standard extends the standard to lower sampling rates and bit rates as well as to multichannel and multilingual audio coding. Of highest interest for mobile radio applications are the activities of ISO/MPEG Phase 4 which address coding at very low bit rates.

1. SIGNALS AND SIGNAL DELIVERY

1.1 Wideband Speech and Audio

We witness an increasing interest in high-quality compression of telephone speech, wideband speech, and wideband audio signals /2/. These signal classes differ not only in bandwidth and dynamic range, but also in listener expectation of offered quality. The conventional digital format is PCM, with typical sampling rates and amplitude resolutions (PCM bits per sample) as given in Table 1. Lower bit rates are of high importance in the interest of an economic frequency use as in mobile communications and in indoor wireless communication systems, especially if a significant overhead capacity has to be provided for error protection.

	Frequency range in Hertz	Sampling rate in Hertz	PCM bits per sample	PCM bit rate
Telephone speech	300 - 3,400	8000	8	64 kb/s
Wideband speech	50 - 7,000	16000	14	224 kb/s
Wideband audio	10 -20,000	48000	16	768 kb/s

Table 1: Typical values of basic parameters of three classes of acoustic signals

Basic requirements in the design of low bit rate audio coders are

- high quality of the reconstructed signals with robustness to variations in spectra and levels
- robustness to random and bursty channel bit errors and to packet losses
- graceful degradation of quality with increasing bit error rates in mobile radio and broadcast applications.
- low complexity and power consumption.

Additional *network-related requirements* are

- low encoder/decoder delays
- transcodeability

All these partly conflicting factors have to be carefully considered in selecting a speech coding algorithm for a given application.

Wideband speech Higher bandwidths than that of the 300 - 3400 Hz telephone bandwidth result in major subjective improvements in represented speech quality. A bandwidth of 50 to 7000 Hz not only improves the intelligibility and naturalness of speech, but adds also a feeling of transparent communication, and eases speaker recognition. Applications of high relevance are loudspeaker telephony, ISDN and video conferencing systems, and the use of commentary channels for broadcasting. In 1986 CCITT has recommended a 64 kb/s wideband speech coder developed primarily for transmission over the ISDN basic rate (B) channel /3/. Current activities in wideband speech coding concentrate

on coding at 32 kb/s and below, with the 64 kb/s CCITT standard serving as reference.

Audio Typical application areas for *digital audio* with bandwidths up to 20 kHz are in the fields of audio production, distribution and program exchange, digital audio on PCs, digital audio broadcasting (DAB), digital storage (archives, studios, consumer electronics), interpersonal communications such as videoconferencing and multipoint interactive audivisual communications, and in the field of enhanced quality TV systems.

At 48 kHz sampling rate 16-bit PCM serves as an accepted audio representation standard although its bit rate of 1.536 Mb/s is rather high. Lower bit rates are mandatory if audio signals are to be transmitted over channels of limited capacity or are to be stored in storage media of limited capacity.

Multichannel Stereophony A logical further step in digital audio is the definition of a universal loudspeaker reproduction standard to provide an improved stereophonic image not only for audio-only applications including teleconferencing, but also for enhanced television systems, in particular for High Definition Television systems such as HDTV-T, HD-SAT, ATV, HD-MAC, and for digital storage media. Loudspeaker arrangements, 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), offer an improved realism of auditory ambience. In particular, the three front loudspeakers ensure a sufficient directional stability of the frontal sound image, i.e., a stable middle and an enlarged listening area.

Examples of digital multichannel surround systems are the upcoming ISO/MPEG 3/2-stereo coding standard and Dolby's Stereo SR•D system based on its AC-3 audio coding algorithm. Both systems offer an additional optional *low frequency effect (subwoofer) channel*, to reproduce frequencies below around 120 Hz with one or more loudspeakers which can be positioned freely in the listening room. The overall bitrate for a 3/2-stereo system will possibly fit into the 384-kb/s H0 channel of the ISDN hierarchy (see below).

1.2 Signal Delivery and Storage

Delivery of digital speech and audio signals is possible over terrestrial and satellite-based digital broadcast and transmission systems such as subscriber lines, program exchange links, cellular mobile radio networks, cable-TV networks, etc.

ISDN In Integrated Services Digital Networks (ISDN) customers have physical access to a number of communications channels. The basic-rate interface consists of two 64 kb/s B channels and one D channel (which supports signaling at 16 kb/s but can also carry user information). The primary-rate interface is either a 23 B + D configuration (North America and Japan) or a 30 B + D configuration (Europe); the D channels operate at 64 kb/s.. In both cases other configurations are possible, e.g., 4 H0 or 3 H0 + D (North America and Japan), and 5 H0 or 5 H0 + D (Europe) where a H0 channel supports a bit rate of 384 kb/s. From these numbers it is clear that ISDN offers useful channels for a practical distribution of stereophonic and multichannel audio signals.

IVDLAN The Integrated Voice/Data Local Area Network (IVDLAN; see IEEE 802.9) can cope with real time constraints. It provides a high bandwidth packet service (P channel) and a number of full-duplex isochronous digital channels (B, C, and D channels), similar to ISDN channels. There are two 64-kb/s B channels, a 16-kb/s or 64 kb/s packet channel, and a m x 64 - kb/s broadband channel, similar to ISDN H channels.

Digital Audio Broadcasting (DAB) and Mobile Radio

Satellite-based or terrestrial digital broadcasting is a complex task, in particular, if listeners use mobile and portable receivers. Multipath interference and selective fading are the main impairments to be expected. A broadcast network chain can include sections with different quality requirements ranging from production quality where editing, cutting, postprocessing, etc. has to be taken into account, to program connection quality and emission quality. A program connection must be able to deliver audio of the highest quality to the DAB emitter which implies that the contribution links (which support exchange of programs) and the distribution links (which transmit the sound to the emitters) must perform adequately.

The CCIR has recommended that ISO/MPEG Layer II audio coding algorithms should be used for digital audio broadcasting at a bit rate of 128 kb/s. In addition, the ISO/MPEG Layer III coder is recommended for commentary links (both bit rates are for the mono channel).

Digital Storage The ISO/MPEG activities were originally aiming at coding of audiovisual information for digital storage media, such as magneto-optical disks (MOD), digital audio tape (DAT), read-only and interactive CD, etc. The ISO/MPEG bitstream supports both audio-only signals and multiplexed audio-visual signals A first consumer product in the audio field, Philips *Digital Compact Cassette (DCC)*, makes use of Layer I of the ISO/MPEG coder. One example of a non-standard algorithm is Sony's magneto-optical *MiniDisc (MD)*.

2. PERCEPTION AND PERCEPTUAL CODING

2.1 Auditory Masking

Auditory masking describes the effect that a low level signal (the maskee) can become inaudible when a louder signal (the masker) occurs simultaneously. Without a masker, a signal is inaudible if its sound pressure level (SPL) is below the *threshold in quiet* which depends on frequency and covers a dynamic range of almost 80 dB as shown in the lower curve of Fig.1.

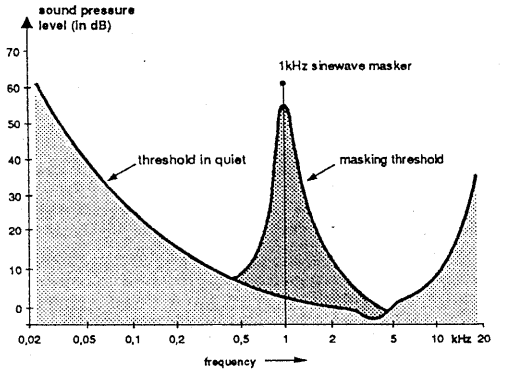

Figure 1: Threshold in quiet (lower curve) and masking thresholds (tone-masks-noise).

In the presence of a masker a *masking threshold* can be measured below which any signal will not be audible. The masking threshold 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 noise masker in Figure 1: around 1 kHz the SPL of the pure tone maskee can be surprisingly high, it will be masked as long as its SPL is only a few dB below the masking threshold!

This phenomenon can be exploited in speech and audio coding by an appropriate noise shaping in the encoding process. Such noise shaping can provide high coding quality without providing a high signal-to-noise ratio. The masking depends on the spectral distribution of masker and maskee and on their variation with time. The masked signal may consist either of quantization noise or even of components of the source signal. In the latter case source signal components below the masking threshold need not to be coded and transmitted since they will be masked by the masker. An efficient source coding will try to remove all signal components, which are irrelevant to the ear. If all noise contributions can be masked the coder is *perceptually transparent.*

2.2 Noise Shaping and Perceptual Coding

The dependence of human auditory perception *on frequency* and the accompanying perceptual tolerance to errors can (and should) directly influence encoder designs; *dynamic noise-shaping* techniques can shift coding noise to frequency bands where it is not of perceptual importance. The noise shifting must be related to the actual short-term input spectrum and can be done in different ways. As one example, *quantization noise feedback* can be used in predictive schemes. Various configurations exist, they have in common a feedback of filtered quantization noise to the input of the quantizer. In frequency domain coding a *fixed or dynamic allocation of bits* (and hence of quantization noise contributions) to subbands or transform coefficients offers the easiest way to take into account properties of the auditory system. As a third example, *perceptual weighting filters* can easily be employed in analysis-by-synthesis coders, since the coding errors are actually determined in the encoding process of these coders. Note that in all examples the noise shaping is located in the encoder only, so it does not contribute to the bit rate.

Noise shaping is the basis for an efficient, namely perception-based coding of an speech or audio signal. Fig. 2 depicts the structure of a perceptual coder (buffering is not included). The encoding process is controlled by the signal-to-mask ratio vs. frequency curve that is calculated via an FFT-based spectral analysis of the audio segment to be coded. Principally, any coding scheme can be used that can be controlled by such perceptual information. If the necessary bit rate for a complete masking of distortions is available the coding scheme will be *transparent*, i.e. the decoded signal is indistinguishable from the source signal, e.g. the compact disc reference. In practical design of perceptual coding we cannot go to the limits of masking or just noticeable distortion, since postprocessing of the acoustic signal (e.g., filtering in equalizers) by the end-user and imultiple encoding/decoding processes have to be considered. In addition 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 coders.

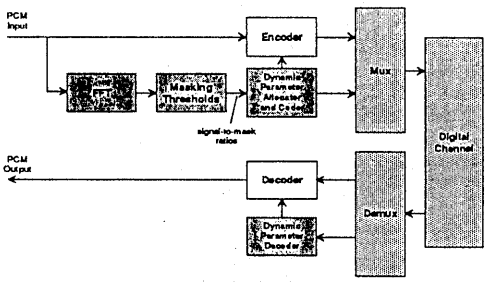

Figure 2: Block diagram of perceptual coding

3. THE ISO/MPEG PHASE 1 AUDIO CODING STANDARD(s)

3.1 The Standardization Process

Since 1988 the ISO Moving Pictures Expert Group (MPEG) has undertaken the task of developing a standard for audio-visual coding and digital storage media at bit rates up to about 1.5 Mb/s. The coded audio signal should occupy only a very small fraction of that capacity. The ISO/MPEG Audio Subgroup had the responsibility to develop an audio codec with sampling rates of 32, 44.1 and 48 kHz and with bit rates per monophonic channel between 32 and 192 kb/s, and per stereo channel between 128 and 384 kb/s.

16-bit PCM is an accepted audio representation standard, also used often as reference in subjective assessments. In the *ISO/MPEG Phase 1 Audio Coding standard* reductions of the PCM rates by a factor of about 6 to 12 are achieved by using frequency-domain coding, which offers an easy way for noise shaping in accordance with psychoacoustically based perceptual coding. In the first stage the audio signal is converted into spectral subbands components via an analysis filterbank. Each spectral component is quantized whereby the number of quantizer levels for each component is obtained from an dynamic bit allocation rule that is controlled by a psycho-acoustic model that makes use of the known properties of auditory masking and that obtains its information about the short-term spectrum by running an FFT in parallel to the encoder. The model is only needed in the encoder which makes the decoder less complex, a desirable feature for audio playback and audio broadcasting applications.

The standard consists of three layers I, II, and III of increasing complexity and subjective performance.

ISO/MPEG Layer I

Layer I contains the basic mapping of the digital audio input into 32 subbands via equally spaced bandpass filters (Fig.3), fixed segmentation to format the data into blocks (8 ms at 48 kHz sampling rate) and quantization with block companding. At 48 kHz sampling rate each band has a width of 750 Hz.

A polyphase filter structure is used for the frequency mapping; its filters are of order 511 which implies an impulse response of 5.33 ms length (at 48 kHz). In each subband blocks of 12 decimated samples are formed and for each block *signal-to-mask ratios* are calculated via an 512-point FFT. For each subband the bit allocation selects a quantizer (out of a set of 15) such that both the bit rate requirement and the masking requirement are being met.

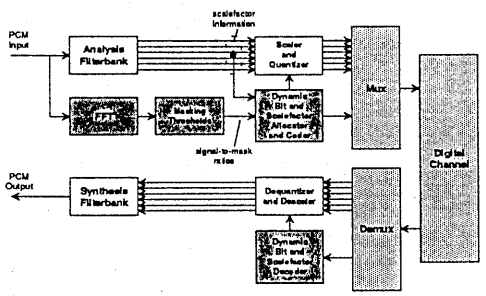

Fig. 3: Structure of ISO/MPEG Phase 1 Audio encoder and decoder, Layers I and II

The *decoding* is straightforward: the subband sequences are reconstructed on the basis of 12-sample subband blocks 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. *At a rate of 192 kb/s per monophonic channel, the average Mean Opinion Score (MOS) value of the Layer I coder, measured over ten test items, was around 4.7.*

ISO/MPEG Layer II

The ISO/MPEG AUDIO Layer II coder is basically similar to the Layer I coder but achieves a better performance through three modifications. Firstly the input to the psychoacoustic models is a 1024-point FFT leading to a finer frequency resolution for the calculation of the global masking threshold. Secondly the overall scalefactor side information is reduced by exploiting redundancies between three adjacent 12-sample blocks. Thirdly, a finer quantization with a maximum of 16 bit amplitude resolution is provided (which reduces the coding noise). *At a rate of 128 kb/s per monophonic channel the average MOS value, measured over ten test items, was around 4.8.*

ISO/MPEG Layer III

This layer introduces many new features which are not part of the Layer I and Layer II codecs (see Fig. 4). It achieves its improvement mainly by an improved frequency mapping, an analysis-by-synthesis approach for the noise allocation, an advanced pre-echo control, and finally by nonuniform quantization with entropy coding. 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 a 6-point or 18-point modified DCT (MDCT) with 50% overlap to each of the subbands. The maximum number of frequency components is 32 x 18 = 576 each representing a bandwidth of 24000/575 = 41.67 Hz. The term *hybrid filterbank* is used to describe such a cascade of polyphase filterbank and MDCT transform. The decoding follows that of the encoding process. *The average MOS values (determined over ten test items) for Layer III was around 3.7 at a rate of 64 kb/s per monophonic channel*; seventy percent of the test items had a MOS value of 4 and above.

3.2 Evolution of the Phase 1 Standard

The ISO/MPEG Phase 1 standard is the first standard in audio coding (besides the quasi-standard of the CD). It is worthwile to note that its normative part describes the decoder and the meaning of the encoded bitstream, but that the encoder is not defined thus leaving room for an evolutionary improvement of the encoder. In particular, different psychoacoustic models can be used ranging from very simple to very complex ones based on quality and implementability requirements (the standard gives two *examples* of such models). Therefore we will see different solutions for encoding including proposals for a better use of the joint stereo mode provided by the standard.

Undoubtedly non-standard algorithms serving different goals will also be developed and used.

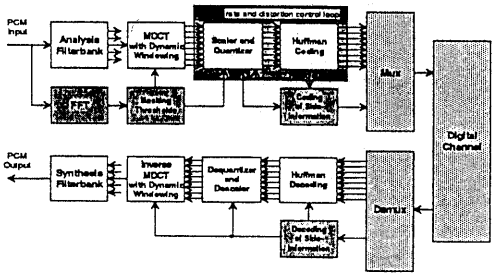

Fig. 4: Structure of ISO/MPEG Phase 1 Audio encoder and decoder, Layer III

4. THE ISO/MPEG PHASE 2 AUDIO CODING STANDARD

In the current ISO/MPEG Phase 2 emphasis in the audio coding part of the new activity is on multi-channel audio and on an extension of the existing standard to lower sampling frequencies and bit rates.

4.1 Multichannel Stereophony

ISO/MPEG Phase 2 aims at a universal loudspeaker reproduction standard to provide a better stereophonic image both for audio-only applications and for enhanced television systems, in particular for High Definition Television systems ("home-theatre sounds"). Phase 2/Audio will provide a hierarchical, multi-channel digital audio system which supports stereophonic presentation with three to five channels and/or a number of multilingual channels:

- 3 channels 3/0-stereo: left, right, centre
- 4 channels: 3/1-stereo left, right, centre, surround
- 5 channels: 3/2-stereo: left, right, centre, surround left, surround right.

In addition an optional low frequency (subwoofer) channel with a bandwidth of less than 120 Hz dedicated to special effects may be provided. In digital multichannel surround systems redundancies and irrelevancies, such as interchannel correlations and interchannel masking effects, respectively, are exploited to reduce the overall bit rate. In addition, stereo-irrelevant components of the multichannel signal may be identified and reproduced in a monophonic format to bring the bit rates further down.

4.2 Lower Sampling Frequencies

Phase 2 includes an extension that adds sampling frequencies of 24, 22.05 and 16 kHz. It is expected that these sampling frequencies will be useful for the transmission of wideband speech and medium quality audio signals. Only small modifications of ISO/MPEG Phase 1 are needed for this extension, the bit stream syntax still complies with the standard and causes an MPEG Phase 1 decoder to mute.

5. ISO/MPEG PHASE 4: CODING AT VERY LOW BIT RATES

Although ISO/MPEG Phase 2 supports lower bit rates, its potential is basically limited to coding at 64 kb/s and above. The scope of ISO/MPEG Phase 4 is the development of standards for generic high quality audio coding systems operating below 64 kb/s, perhaps much lower than that rate. A very attractive application is the transmission of a stereo audio signal over an 64 kb/s ISDN basic rate channel. An even more ambitious goal is the transmission of high quality audio signals over the existing analog telephone network. For example, the CCITT *fast modem project* will support such digital transmissions at bit rates of up to 24 kb/s. Transmission of audio signals from and to remote data bases, interactive multimedia and interchange media such as Smart Cards and PCMCIA Memory Cards could benefit significantly from MPEG Phase 4 /4/.

3rd Generation Systems A very broad range of applications can be foreseen in the area of mobile radio, where services are moving to portable and handheld devices and where new networks are evolving. The end of this century will see third generation mobile communications, in specific the CCIR Task Group 8/1 defined *Future Public Land Mobile Telecommunications Systems* (FPLMTS). Its projected European standard will be the *Universal Mobile Telecommunications System* (UMTS) for persons, cars, busses, trains, ships and aircrafts. Future *Personal Communication Networks* must

- be of high quality
- be spectrally efficient (high capacity)
- support various services including those provided by other networks
- provide full area coverage
- cover indoor/outdoor applications
- offer security
- be economic (in terminal prices and usage charges).

Coders for these networks must not only operate at low bit rates but must be error-robust in burst-error and packet-loss environments. Error concealment techniques will play a significant role, since due to the lack of available bandwidth, traditional channel coding techniques cannot sufficiently improve the reliablity of the channel.

A preliminary ISO/MPEG Requirements List indicates that sampling rates down to 8 kHz, codecs with various coding delays, and both fixed and variable bit rates will be supported by Phase 4 coders. The basic audio quality will be more important than compatibility with the existing or upcoming ISO/MPEG standard. This opens the

door for completely new solutions, which will be necessary to meet the goals.

Approaches to audio coding at very low bit rates include

- enhanced perceptual models (including excitation-based models)
- enhanced adaptation of time/frequency resolution /4/
- vector quantization
- entropy coding
- analysis-by-synthesis coding /5/.

Analysis-by-synthesis coding

We describe now a coding technique that has proven to be a very powerful tool in linear predictive coding of *speech* signals at low bit rates. Analysis-by-synthesis predictive coders have as a common characteristic that, unlike differential PCM, it is not the quantized residual signal that is transmitted to the decoder. Instead a number of candidate excitation signals is available in the encoder, e.g., organized in a codebook. A feedback loop allows all possible candidate excitation signals (usually called *excitation vectors*) to locally decode corresponding candidate output vectors by feeding them through a synthesis filter. The approach requires synthesis during error analysis and is therefore refered to as *analysis-by-synthesis technique*; its principle structure is shown in Fig. 5. The optimal excitation is defined to be that candidate excitation signal that minimizes the difference between the buffered speech input vector 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 excitation signals can be modeled in different ways: The excitation is a small number of optimally placed impulses in multipulse-excited MPE-LP coding, a small number of regularly spaced impulses in regular-pulse-excited RPE-LP coding, and a complete sequence of impulses in code-excited CE-LP coding. In the latter case the stored excitation candidates are either of stochastic nature (the term stochastic coding is often used in this context), or they are characteristic for the signal to be encoded and have been derived from a training procedure based on an appropriate ensemble of that signal. A refinement is reached by adding an adaptive codebook vector quantizer that consists of contributions from the most recent excitation vectors and is updated with each coding frame. In the exhaustive two-stage search first the optimum excitation vector from the adaptive codebook is selected, and then a stochastic codebook is searched to match the remaining signal, which will be more random in nature. The total excitation is a weighted sum of the contributions of both codebooks (see Fig. 5).

By using such techniques borrowed from speech coding it may be possible to reduce audio bit rates significantly below 64 kb/s.

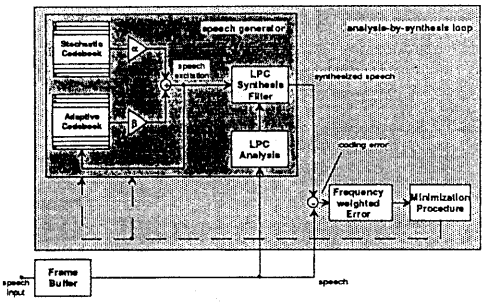

Fig.5: Principle of code-excited encoder with two codebooks*.

6. Conclusion

Digital audio is on its way to be applied in many different fields, such as consumer electronics, professional audio processing, telecommunications and broadcasting, with very different applications. Perceptual coding, either based on subband coding or transform coding algorithms, has paved the way to high quality coding at low bit rates. In the field of digital audio compression the recent ISO/MPEG Phase 1 audio coding algorithm with its three layers is the first international standard for digital storage media. Other coding schemes serving different goals and meeting other requirements will evolve, and application of audio coding to mobile radio will become an important issue. Emerging activities of the ISO/MPEG expert group aim at a proposal for audio coding at very low bit rates. Phase 4 coders must be designed for mobile radio applications which will play a significant role in future communications.

7. References

/1/ ISO/IEC JTC1/SC29, "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s - IS 11172-3 (Part 3, Audio)", 1993-05-20.

/2/ P. Noll, "Wideband Speech and Audio Coding," IEEE Commun. Mag., Vol. 31, No.11, pp.34 - 45, Nov.1993.

/3/ /1/ P. Mermelstein, "G.722, A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals," IEEE Commun. Mag., pp. 8-15, Jan. 1988.

/4/ R. van der Waal, K. Brandenburg, G.Stoll, "Current and Future Standardization of High-Quality Digital Audio Coding in MPEG", 1993 Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 17-20, 1993.

/5/ P. Noll, "Speech Coding for Communications", Eurospeech '93, Proc. pp. 479 - 488, Berlin.

* the algorithm is that of the US Federal Standard 4.8 kb/s CE-LP speech coder