AUDIO CODING: BASICS AND STATE OF THE ART

PACS REFERENCE: 43.75.CD

Brandenburg, Karlheinz
Fraunhofer Institut Integrierte Schaltungen, Arbeitsgruppe Elektronische Medientechnolgie
Am Helmholtzring 1
98603 Ilmenau
Germany
Tel: +49-3677-69-4340
Fax: +49-3677-69-4399
E-mail: bdg@emt.iis.fhg.de

ABSTRACT

High quality audio coding based on perceptual models has found its way to widespread application in broadcasting and Internet audio (e.g. mp3). After a brief presentation of the basic ideas, a short overview over current high quality perceptual coders will be given. Algorithms defined by the MPEG group (MPEG-1 Audio, e.g. MPEG Layer-3 (mp3), MPEG-2 Advanced Audio Coding, MPEG-4 Audio including its different functionalities) still define the state of the art. While there has been some saturation in the progress for highest quality (near transparent) audio coding, considerable improvements have been achieved for lower ("near CD-quality") coding. One technology able to push the bit-rates for good quality audio lower is Spectral Band Replication (SBR) as used in mp3PRO or improved AAC codes.

INTRODUCTION

High quality audio compression has found its way from research to widespread applications within a couple of years. Early research of 15 years ago was translated into standardization efforts of ISO/IEC and ITU-R 10 years ago. Since the finalization of MPEG-1 in 1992, many applications have been devised. In the last couple of years, Internet audio delivery has emerged as a powerful category of applications. These techniques made headline news in many parts of the world because of the potential to change the way of business for the music industry. Currently, among others the following applications employ low bit-rate audio coding techniques:

- Digital Audio Broadcasting (EUREKA DAB, WorldSpace, ISDB, DRM)
- ISDN transmission of high quality audio for broadcast contribution and distribution purposes
- Archival storage for broadcasting
- Accompanying audio for digital TV (DVB, ATSC, Video CD, ARIB)
- Internet streaming (RealAudio, Microsoft Netshow, Apple Quicktime and others)
- Portable audio (mpman, mplayer3, Rio, Lyra, YEPP and others)
- Storage and exchange of music files on computers

THE BASICS OF HIGH QUALITY AUDIO CODING

The basic task of a perceptual audio coding system is to compress the digital audio data in a way that

- the compression is as efficient as possible, i.e. the compressed file is as small as possible and
- the reconstructed (decoded) audio sounds exactly (or as close as possible) to the original audio before compression.

Other requirements for audio compression techniques include low complexity (to enable software decoders or inexpensive hardware decoders with low power consumption) and flexibility to cope with different application scenarios. The technique to do this is called perceptual encoding and uses knowledge from psychoacoustics to reach the target of efficient but inaudible compression. Perceptual encoding is a lossy compression technique, i.e. the decoded file is not a bit-exact replica of the original digital audio data.

Fig 1 shows the basic block diagram of a perceptual encoding system.

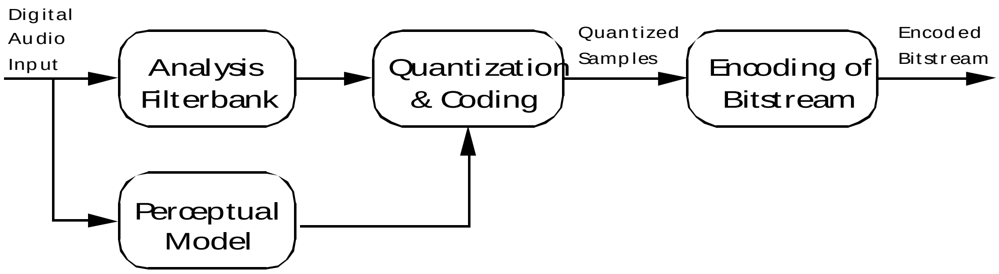

Figure 1: Block diagram of a perceptual encoding/decoding system.

It consists of the following building blocks:

- *Filter bank*: A filter bank is used to decompose the input signal into sub-sampled spectral components (time/frequency domain). Together with the corresponding filter bank in the decoder it forms an analysis/synthesis system.
- *Perceptual model*: Using either the time domain input signal and/or the output of the analysis filter bank, an estimate of the actual (time and frequency dependent) masking threshold is computed using rules known from psychoacoustics. This is called the perceptual model of the perceptual encoding system.
- *Quantization and coding*: The spectral components are quantized and coded with the aim of keeping the noise, which is introduced by quantizing, below the masking threshold. Depending on the algorithm, this step is done in very different ways, from simple block companding to analysis-by-synthesis systems using additional noiseless compression.
- *Encoding of bitstream*: A bitstream formatter is used to assemble the bitstream, which typically consists of the quantized and coded spectral coefficients and some side information, e.g. bit allocation information.

All current high quality low bit-rate audio coding systems follow the basic paradigm described above. They differ in the types of filter banks used, in the quantization and coding techniques and in the use of additional features.

STANDARDIZED CODECS

MPEG (formally known as ISO/IEC JTC1/SC29/ WG11, mostly known by its nickname, Moving Pictures Experts Group) has been set up by the ISO/IEC standardization body in 1988 to develop generic (to be used for different applications) standards for the coded representation of moving pictures, associated audio, and their combination. Since 1988 ISO/MPEG has been undertaking the standardization of compression techniques for video and audio. The original main topic of MPEG was video coding together with audio for Digital Storage Media (DSM). From the beginning, audio-only applications have been part of the charter of the MPEG audio subgroup. Since the finalization of the first standard in 1992, MPEG Audio in its different flavors (mostly Layer-2, Layer-3 and Advanced Audio Coding) has delivered on the promise to establish universally applicable standards.

<u>MPEG-1</u>

MPEG-1 is the name for the first phase of MPEG work, started in 1988, and finalized with the adoption of ISO/IEC IS 11172 in late 1992. The audio coding part of MPEG-1 (ISO/IEC IS 11172-3, see [1] describes a generic coding system, designed to fit the demands of many applications. MPEG-1 audio consists of three operating modes called layers with increasing complexity and performance from Layer-1 to Layer-3. Layer-3 (in recent years nicknamed *MP3* because of the use of .mp3 as a file extension for music files in Layer-3 format) is the highest complexity mode, optimized to provide the highest quality at low bit-rates (around 128 kbit/s for a stereo signal).

The following paragraphs describe the Layer-3 encoding algorithm along the basic blocks of a perceptual encoder. More details about Layer-3 can be found in [1] and [2].

Fig 2 shows the block diagram of a typical MPEG-1/2 Layer-3 encoder.

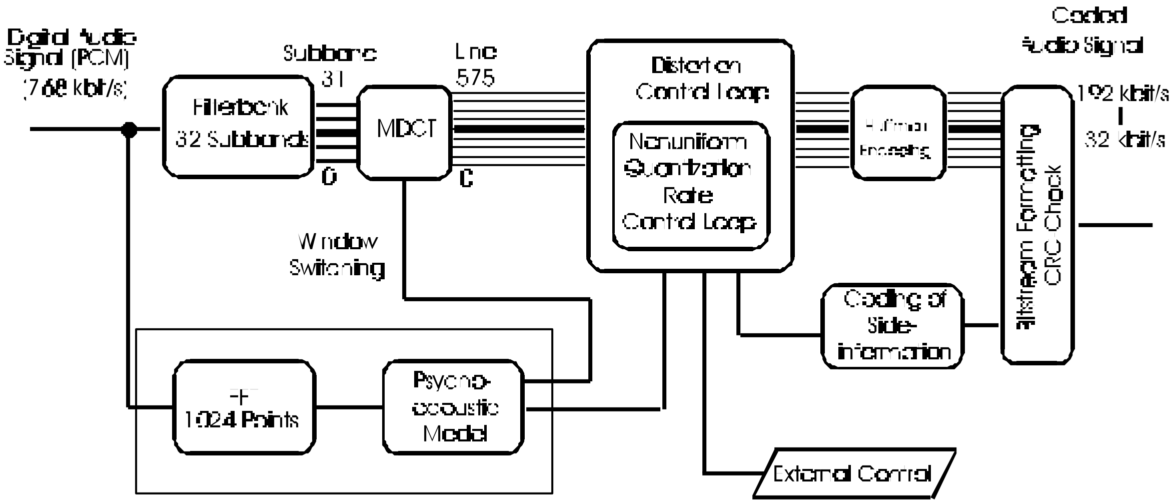

Fig. 2: Block diagram of MPEG Layer-3 (MP3) encoding

<u>Filterbank</u>

The filterbank used in MPEG-1 Layer-3 belongs to the class of hybrid filterbanks. It is built by cascading two different kinds of filterbank: First a polyphase filterbank (as used in Layer-1 and Layer2) and then an additional Modified Discrete Cosine Transform (MDCT). The polyphase filterbank has the purpose of making Layer-3 more similar to Layer-1 and Layer-2. The subdivision of each polyphase frequency band into 18 finer subbands increases the potential for redundancy removal, leading to better coding efficiency for tonal signals. Another positive result of better frequency resolution is the fact that the error signal can be controlled to allow a finer tracking of the masking threshold. The filter bank can be switched to less frequency resolution to avoid preechoes.

<u>Perceptual Model</u>

The perceptual model is mainly determining the quality of a given encoder implementation. A lot of additional work has gone into this part of an encoder since the original informative part in [1] has been written. The perceptual model either uses a separate filterbank as described in [1] or combines the calculation of energy values (for the masking calculations) and the main filterbank. The output of the perceptual model consists of values for the masking threshold or allowed noise for each coder partition. In Layer-3, these coder partitions are roughly equivalent to the critical bands of human hearing. If the quantization noise can be kept below the masking threshold for each coder partition, then the compression result should be indistinguishable from the original signal.

Quantization and Coding

A system of two nested iteration loops is the common solution for quantization and coding in a Layer-3 encoder. Quantization is done via a power-law quantizer. In this way, larger values are automatically coded with less accuracy and some noise shaping is already built into the quantization process. The quantized values are coded by Huffman coding. To adapt the coding process to different local statistics of the music signals the optimum Huffman table is selected from a number of choices. The Huffman coding works on pairs or quadruples. To get even better adaption to signal statistics, different Huffman code tables can be selected for different parts of the spectrum. Since Huffman coding is basically a variable code length method and noise shaping has to be done to keep the quantization noise below the masking threshold, a global gain value (determining the quantization step size) and scalefactors (determining noise shaping factors for each scalefactor band) are applied before actual quantization. The process to find the optimum gain and scalefactors for a given block, bit-rate and output from the perceptual model is usually done by two nested iteration loops in an analysis-by-synthesis way:

MPEG-2

MPEG-2 denotes the second phase of MPEG. It introduced a lot of new concepts into MPEG video coding including support for interlaced video signals. The main application area for MPEG-2 is digital television. The original (finalized in 1994) MPEG-2 Audio standard [3] just consists of two extensions to MPEG-1:

- Backwards compatible multichannel coding adds the option of forward and backwards compatible coding of multichannel signals including the 5.1 channel configuration known from cinema sound.
- Coding at lower sampling frequencies adds sampling frequencies of 16 kHz, 22.05 kHz and 24 kHz to the sampling frequencies supported by MPEG-1. This adds coding efficiency at very low bit-rates.

Both extensions do not introduce new coding algorithms over MPEG-1 Audio. The multi-channel extension contains some new tools for joint coding techniques.

MPEG-2 Advanced Audio Coding

In verification tests in early 1994 it was shown that introducing new coding algorithms and giving up backwards compatibility to MPEG-1 promised a significant improvement in coding efficiency (for the five channel case). As a result, a new work item was defined and led to the definition of MPEG-2 Advanced Audio Coding (AAC) ([4], see the description in [5]). AAC is a second-generation audio coding scheme for generic coding of stereo and multi-channel signals.
Figure 3 shows a generic block diagram of a typical AAC encoder. Comparing this to Layer-3, the most visible difference is the addition of a number of new blocks. AAC follows the same basic paradigm as Layer-3. AAC encoders often use the same double iteration loop structure as described for Layer-3. The difference is in a number of details and in the addition of more flexibility and more coding tools.

Tools to enhance coding efficiency

The following changes compared to Layer-3 help to get the same quality at lower bit-rates:

- Higher frequency resolution: The number of frequency lines in AAC is up to 1024 compared to 576 for Layer-3:
- Improved joint stereo coding: Compared to Layer-3, both the mid/side coding and the intensity coding are more flexible, allowing applying them to reduce the bit-rate more frequently.

- Improved Huffman coding: In AAC, coding by quadruples of frequency lines is applied more often. In addition, the assignment of Huffman code tables to coder partitions allows for many more options.

Tools to enhance audio quality

There are other improvements in AAC, which help to retain high quality for classes of very difficult signals.

- Enhanced block switching: Instead of the hybrid (cascaded) filter bank in Layer-3, AAC uses a standard switched MDCT (Modified Discrete Cosine Transform) filter bank with an impulse response (for short blocks) of 5.3 ms at 48 kHz sampling frequency. This compares favorably with Layer-3 at 18.6 ms and reduces the amount of pre-echo artifacts (see below for an explanation).
- Temporal Noise Shaping, TNS: This technique does noise shaping in time domain by doing an open loop prediction in the frequency domain. TNS is a new technique that proves to be especially successful for the improvement of speech quality at low bit-rates.

With the sum of many small improvements, AAC reaches on average the same quality as Layer-3 at about 70 % of the bit-rate.

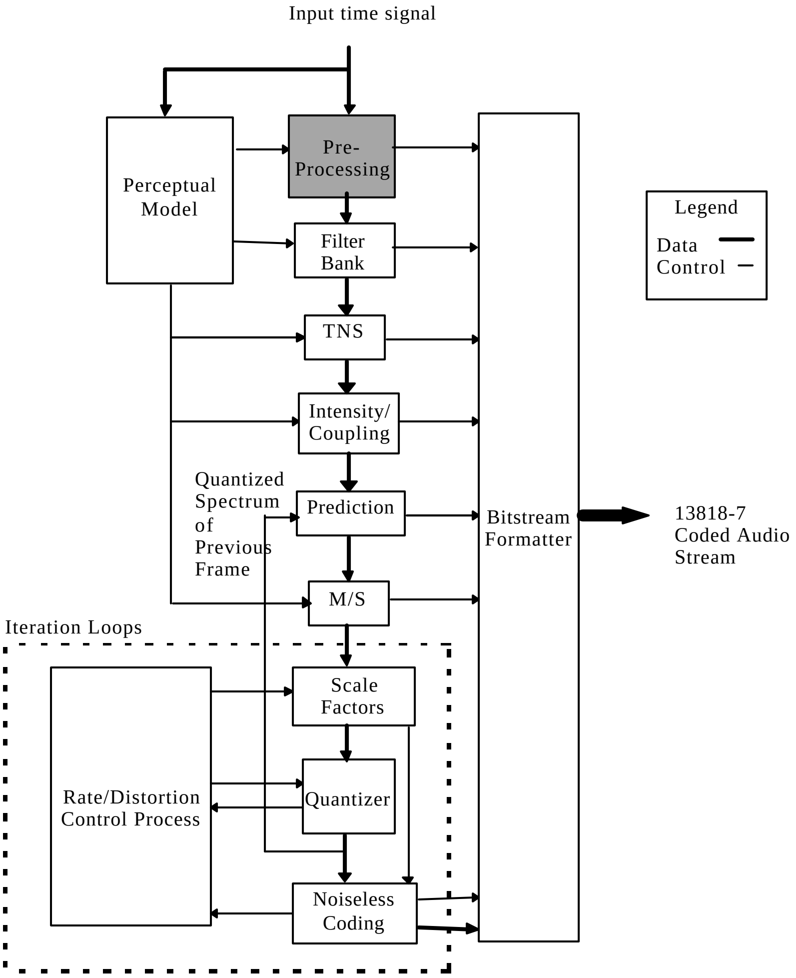

Fig 3: Block diagram of MPEG-2 Advanced Audio Coding

NEW FEATURES

The development of recent new audio codecs like MPEG-4 audio [6] has put less emphasis on more compression efficiency (there seems to be some satu-ration) and more on additional functionalities. In MPEG-4 audio standardization the emphasis on new functionalities led to a complete toolbox for com-pressing audio and speech signals at very different bit-rates, to render the signals using postprocessing such as audio effects (e.g. reverb) and even synthe-size audio instead of just compressing natural audio signals. The bitrate requirements for such systems range from very low bit-rates (e.g. audio synthesis or the bandwidth extension technology mentioned be-low) to high bandwidth (e.g. transmission of a large number of high resolution audio channels).

Scalability

Current systems are mostly either targeted for a cer-tain bit-rate (like most speech codecs) or flexible within a certain range of bit-rates. Scalability or em-bedded coding is a powerful tool to postpone the choice of bit-rates and quality to the time the com-pressed audio needs transfer over a low bit-rate con-nection or to storage media with limited memory capacity. With scalable audio coding like the fea-tures designed into MPEG-4 audio and other codecs, the original compression can be done at the highest quality level envisioned while downloading or trans-fer to a portable device can be tailored to the actually available resources.

Bandwidth extension technologies

The newest addition to the toolkit of perceptual en-coding techniques are bandwidth extension tech-niques as known under the names of mp3PRO [7] or AAC+. These techniques add high frequency signals at the decoder in a way that the transmitted signal is still compliant with the standard. This allows a smooth transition: Old decoders will still get a good quality signal (but not the advantage of the new technology), new decoders are capable of decoding stan-dard, non-enhanced signals.

CONCLUSIONS

High quality audio coding is after 14 years of standardization still an active research area. New systems will use new paradigms, probably of parametric coding similar to techniques already in some use in MPEG-4 audio and the Spectral Band Replication (SBR) technology. For high qualities near to the limit of detectability of audible differences, the state of the art is defined by MPEG-2 Advanced Audio Coding and it's close relatives such as MPEG-4 general audio coding. It will be interesting to see whether more knowledge about psychoacoustics will enable further progress in this area.

REFERENCES:

[1] ISO/IEC IS 11172-3, Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s, Part 3: Audio, 1993

[2] K. Brandenburg, G. Stoll: ISO-MPEG-1 Audio: A Generic Standard for Coding of High Quality Digital Audio, in: Collected Papers on Digital Audio Bit-Rate Reduktion, N. Gilchrist and Chr. Grewin, ed., New York 1996, pp. 31 - 42

[3] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard IS 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio".

[4] ISO/IEC IS 13818-7, Information technology -- Generic coding of moving pictures and associated audio information Part 7: Advanced Audio Coding (AAC), 1997

[5] K. Brandenburg and Marina Bosi, Overview of MPEG audio: Current and Future Standards for Low Bit-Rate Audio Coding, Journal of the Audio Engineering Society, Vol. 45, Jan/Feb 1997, pp. 4 – 21

[6] K. Brandenburg, O. Kunz, A. Sugiyama: "MPEG-4 Natural Audio Coding", Image Communications Journal, 1999

[7] http://www.codingtechnologies.de/