# MP3 AND AAC EXPLAINED

**KARLHEINZ BRANDENBURG**[1]

[1] *Fraunhofer Institute for Integrated Circuits FhG-IIS A, Erlangen, Germany*
bdg@iis.fhg.de

The last years have shown widespread proliferation of .mp3-files, both from legal and illegal sources. Yet most people using these audio files do not know much about audio compression and how to use it. The paper gives an introduction to audio compression for music file exchange. Beyond the basics the focus is on quality issues and the compression ratio / audio bandwidth / artifacts tradeoffs.

## MPEG AND INTERNET AUDIO

The proliferation of MPEG coded audio material on the Internet has shown an exponential growth since 1995, making ".mp3" the most searched for term in early 1999 (according to http://www.searchterms.com). "MP3" has been featured in numerous articles in newspapers and periodicals and on TV, mostly on the business pages because of the potential impact on the recording industry. While everybody is using MP3, not many (including some of the software authors writing MP3 encoders, decoders or associated tools) know the history and the details of MPEG audio coding. This paper explains the basic technology and some of the special features of MPEG-1/2 Layer-3 (aka MP3). It also sheds some light on the factors determining the quality of compressed audio and what can be done wrong in MPEG encoding and decoding.

### Why MPEG-1 Layer-3 ?

Looking for the reasons why MPEG-1/2 Layer-3 and not other compression technology has emerged as the main tool for Internet audio delivery, the following comes to mind:

- Open standard
  MPEG is defined as an open standard. The specification is available (for a fee) to everybody interested in implementing the standard. While there are a number of patents covering MPEG Audio encoding and decoding, all patent holders have declared that they will license the patents on fair and reasonable terms to everybody. No single company "owns" the standard. Public example source code is available to help implementers to avoid misunderstand the standards text. The format is well defined. With the exception of some incomplete implementations no problems with interoperability of equipment and software from different vendors have been reported.
- Availability of encoders and decoders
  Driven first by the demand of professional use for broadcasting, hardware (DSP) and software decoders have been available for a number of years.
- Supporting technologies
  While audio compression is viewed as a main enabling technology, the widespread use of computer soundcards, computers getting fast enough to do software audio decoding and even encoding, fast Internet access for universities and businesses as well as the spread of CD-ROM and CD-Audio writers all contributed to the ease of distributing music in MP3 format via computers.

In short, MPEG-1/2 Layer-3 was the right technology available at the right time.

### Newer audio compression technologies

MPEG-1 Layer-3 has been defined in 1991. Since then, research on perceptual audio coding has progressed and codecs with better compression efficiency became available. Of these, MPEG-2 Advanced Audio Coding (AAC) was developed as the successor for MPEG-1 Audio. Other, proprietary audio compression systems have been introduced with claims of higher performance. This paper will just give a short look to AAC to explain the improvements in technology.

## 1. 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 in-

expensive hardware decoders with low power consumption) and flexibility for 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. Perceptual coders for high quality audio coding have been a research topic since the late 70's, with most activity occuring since about 1986. For the purpose of this paper we will concentrate on the format mostly used for Internet audio and flash memory based portable audio devices, MPEG-1/2 Layer-3 (aka MP3), and the format the author believes will eventually be the successor of Layer-3, namely MPEG-2 Advanced Audio Coding (AAC).

### 1.1. A basic perceptual audio coder

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

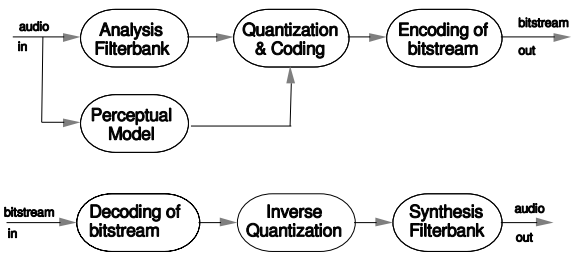


Figure 1: Block diagram of a perceptual encoding/decoding system.

It consists of the following building blocks:

- Filter bank:
  A filter bank is used to decompose the input signal into subsampled 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.

## 2. MPEG AUDIO CODING STANDARDS

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 coding for Digital Storage Media (DSM). The audio coding standard developed by this group has found its way into many different applications, including

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

The most widely used audio compression formats are MPEG-1/2 Audio Layers 2 and 3 (see below for the definition) and Dolby AC-3. A large number of systems currently under development plan to use MPEG-2 AAC.

### 2.1. MPEG-1

MPEG-1 is the name for the first phase of MPEG work, started in 1988, and was 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 [5] 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 is the highest complexity mode, optimised to provide the highest quality at low bit-rates (around 128 kbit/s for a stereo signal).

### 2.2. 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 [6] 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 multichannel extension contains some new tools for joint coding techniques.

#### *2.2.1. 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) ([7], see the description in [1]). AAC is a second generation audio coding scheme for generic coding of stereo and multichannel signals.

#### *2.2.2. MPEG-3*

The plan was to define the video coding for High Definition Television applications in a further phase of MPEG, to be called MPEG-3. However, early on it was decided that the tools developed for MPEG-2 video coding do contain everything needed for HDTV, so the development for MPEG-3 was rolled into MPEG-2. Sometimes MPEG-1/2 Layer-3 (MP3) is misnamed MPEG-3.

### 2.3. MPEG-4

MPEG-4, finished in late 1998 (version 1 work, an amendment is scheduled to be finished by end of 1999) intends to become the next major standard in the world of multimedia. Unlike MPEG-1 and MPEG-2, the emphasis in MPEG-4 is on new functionalities rather than better compression efficiency. Mobile as well as stationary user terminals, database access, communications and new types of interactive services will be major applications for MPEG-4. The new standard will facilitate the growing interaction and overlap between the hitherto separate worlds of computing, electronic mass media (TV and Radio) and telecommunications.
MPEG-4 audio consists of a family of audio coding algorithms spanning the range from low bit-rate speech coding (down to 2 kbit/s) up to high quality audio coding at 64 kbit/s per channel and above. Generic audio coding at medium to high bit-rates is done by AAC.

### 2.4. MPEG-7

Unlike MPEG-1/2/4, MPEG-7 does not define compression algorithms. MPEG-7 (to be approved by July, 2001) is a content representation standard for multimedia information search, filtering, management and processing.

## 3. MPEG LAYER-3 AUDIO ENCODING

The following description of Layer-3 encoding focuses on the basic functions and a number of details necessary to understand the implications of encoding options on the sound quality. It is not meant to be a complete description of how to build an MPEG-1 Layer-3 encoder.

### 3.1. Flexibility

In order to be applicable to a number of very different application scenarios, MPEG defined a data representation including a number of options.

- Operating mode
  MPEG-1 audio works for both mono and stereo signals. A technique called *joint stereo coding* can be used to do more efficient combined coding of the left and right channels of a stereophonic audio signal. Layer-3 allows both mid/side stereo coding and, for lower bit-rates, intensity stereo coding. Intensity stereo coding allows for lower bit-rates but brings the danger of a changing the sound image (like moving instruments). The operating modes are
  - Single channel
  - Dual channel (two independent channels, for example containing different language versions of the audio)
  - Stereo (no joint stereo coding)
  - Joint stereo
- Sampling Frequency
  MPEG audio compression works on a number of different sampling frequencies. MPEG-1 defines audio compression at 32 kHz, 44.1 kHz and 48 kHz. MPEG-2 extends this to half the rates, i.e. 16 kHz, 22.05 and 24 kHz. MPEG-2.5 is the name

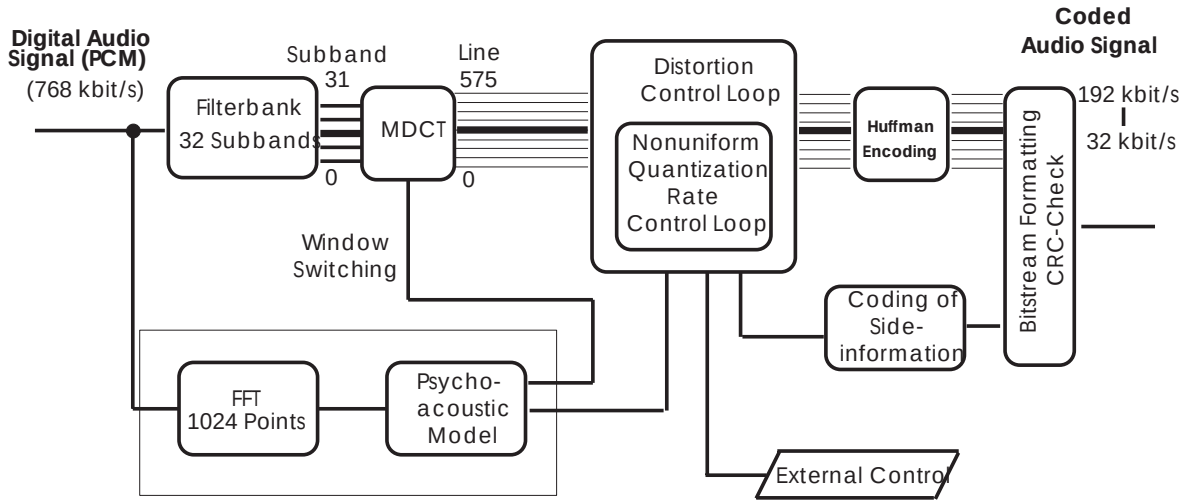


Figure 2: Block diagram of an MPEG-1 Layer-3 encoder.

of a proprietary Fraunhofer extension to MPEG-1/2 Layer-3 and works at 8 kHz, 11.05 and 12 kHz sampling frequencies.

- Bit-rate
  MPEG audio does not work just at a fixed compression ratio. The selection of the bit-rate of the compressed audio is, within some limits, completely left to the implementer or operator of an MPEG audio coder. The standard defines a range of bit-rates from 32 kbit/s (in the case of MPEG-1) or 8 kbit/s (in the case of the MPEG-2 Low Sampling Frequencies extension (LSF)) up to 320 kbit/s (resp. 160 kbit/s for LSF). In the case of MPEG-1/2 Layer-3, the switching of bit-rates from audio frame to audio frame has to be supported by decoders. This, together with the bit reservoir technology, enables both variable bit-rate coding and coding at any fixed bit-rate between the limits set by the standard.

### 3.2. Normative versus informative

One, perhaps the most important property of MPEG standards is the principle of minimizing the amount of normative elements in the standard. In the case of MPEG audio this leads to the fact that only the data representation (format of the compressed audio) and the decoder are normative. Even the decoder is not specified in a bit-exact fashion but by giving formulas for most parts of the algorithm and defining compliance by a maximum deviation of the decoded signal from a reference decoder implementing the formulas with double precision arithmetic accuracy. This enables decoders running both on floating point and fixed point architectures. Depending on the skills of the implementers, fully compliant high accuracy decoders can be done with down to 20 bit (in the case of Layer-3) arithmetic wordlength without using double precision calculations.
Encoding of MPEG audio is completely left to the implementer of the standard. ISO/IEC IS 11172-3 (and MPEG-2 audio, ISO/IEC 13818-3) contain the description of example encoders. While these example descriptions have been derived from the original encoders used for verification tests, a lot of experience and knowledge is necessary to implement good quality MPEG audio encoders. The amount of investment necessary to engineer a high quality MPEG audio encoder has kept the number of independently developed encoder implementations very low.

### 3.3. Algorithm description

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

#### *3.3.1. Filterbank*

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 the 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 (see below).

### *3.3.2. Perceptual Model*

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 [5] has been written.
The perceptual model either uses a separate filterbank as described in [5] 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.

### *3.3.3. 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 and, only in the case of very small numbers to be coded, 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:

- Inner iteration loop (rate loop)
  The Huffman code tables assign shorter code words to (more frequent) smaller quantized values. If the number of bits resulting from the coding operation exceeds the number of bits available to code a given block of data, this can be corrected by adjusting the global gain to result in a larger quantization step size, leading to smaller quantized values. This operation is repeated with different quantization step sizes until the resulting bit demand for Huffman coding is small enough. The loop is called *rate loop* because it modifies the overall coder rate until it is small enough.
- Outer iteration loop (noise control loop)
  To shape the quantization noise according to the masking threshold, scalefactors are applied to each scalefactor band. The systems starts with a default factor of 1.0 for each band. If the quantization noise in a given band is found to exceed the masking threshold (*allowed noise*) as supplied by the perceptual model, the scalefactor for this band is adjusted to reduce the quantization noise. Since achieving a smaller quantization noise requires a larger number of quantization steps and thus a higher bit-rate, the rate adjustment loop has to be repeated every time new scalefactors are used. In other words, the rate loop is nested within the noise control loop. The outer (noise control) loop is executed until the actual noise (computed from the difference of the original spectral values minus the quantized spectral values) is below the masking threshold for every scalefactor band (i.e. critical band).

While the inner iteration loop always converges (if necessary, by setting the quantization step size large enough to zero out all spectral values), this is not true for the combination of both iteration loops. If the perceptual model requires quantization step sizes so small that the rate loop always has to increase them to enable coding at the required bit-rate, both can go on forever. To avoid this situation, several conditions to stop the iterations early can be checked. However, for fast encoding and good coding results this condition should be avoided. This is one reason why an MPEG Layer-3 encoder (the same is true for AAC) usually needs tuning of perceptual model parameter sets for each bit-rate.

## 4. MPEG-2 ADVANCED AUDIO CODING

Figure 3 shows a block diagram of an MPEG-2 AAC encoder.
AAC follows the same basic coding paradigm as Layer-3 (high frequency resolution filterbank, non-uniform quantization, Huffman coding, iteration loop structure using analysis-by-synthesis), but improves on Layer-3 in a lot of details and uses new coding tools for improved quality at low bit-rates.

### 4.1. 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

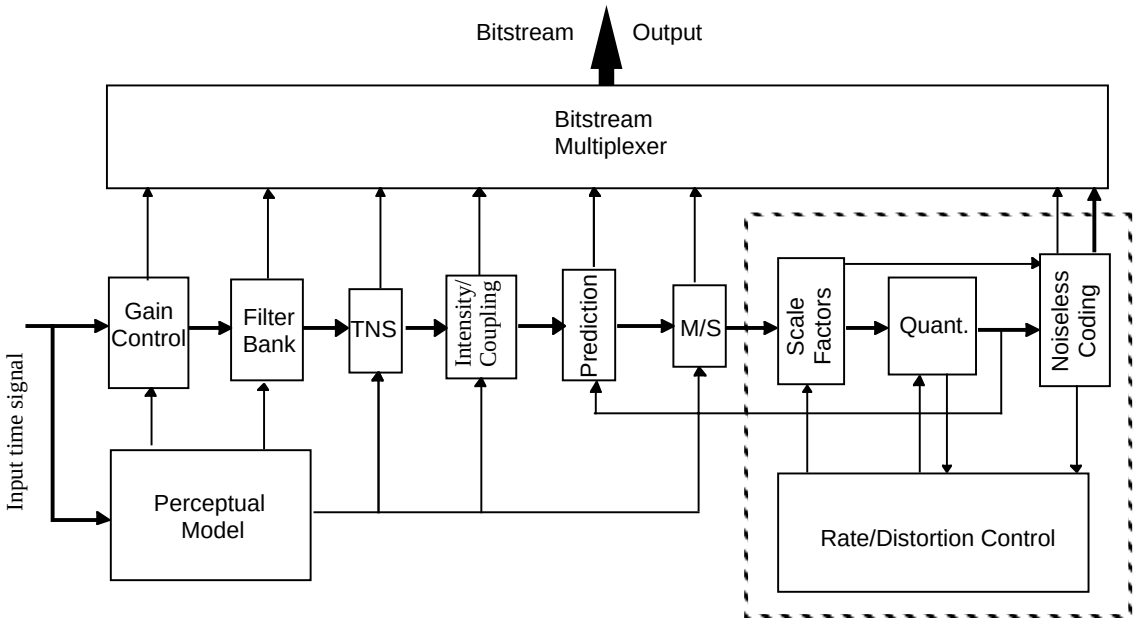


Figure 3: Block diagram of an MPEG-2 AAC encoder.

1024 compared to 576 for Layer-3

- Prediction
  An optional backward prediction, computed line by line, achieves better coding efficiency especially for very tone-like signals (e.g. pitchpipe). This feature is only available within the rarely used main profile.

- Improved joint stereo coding
  Compared to Layer-3, both the mid/side coding and the intensity coding are more flexible, allowing to apply 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 can be much more flexible.

### 4.2. 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) filterbank in Layer-3, AAC uses a standard switched MDCT (Modified Discrete Cosine Transform) filterbank with an impulse response (for short blocks) of 5.3 ms at 48 kHz sampling frequency. This compares favourably 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 which 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.

## 5. QUALITY CONSIDERATIONS

As explained above, the pure compliance of an encoder with an MPEG audio standard does not guarantee any quality of the compressed music. Audio quality differs between different items, depending on basic parameters including, of course, the bit-rate of the compressed audio and the sophistication of different encoders even if they work with the same set of basic parameters. To get more insight about the level of quality possible with MP3 and AAC, let us first have a look at typical artifacts associated with perceptual audio coders.

### 5.1. Common types of artifacts

Unlike analog hi-fi equipment or FM broadcasting, perceptual encoders when run at too low bit-rates or with the wrong parameters exhibit sound deficiencies which are in most cases different from the noise or distortion characteristics we all are used to. The reason for this is the

process generating differences in sound: The error introduced by a high frequency resolution perceptual coder is usually best modeled by a time-varying (in the rhythm of the basic block or frame length) error at certain frequencies, but not constrained to the harmonics of the music signal.
So the signal may be sounding

- distorted, but not like harmonic distortions.
- noisy, but with the noise introduced only in a certain frequency range.
- rough, with the roughness often being very objectionable because the error is changing characteristics about every 20 ms.

#### *5.1.1. Loss of bandwidth*

If an encoder runs out of bits, i.e. it does not find a way to encode a block of music data with the required fidelity (e.g. allowed noise per critical band) within the bounds of available bit-rate, some frequency lines might get set to zero (deleted). The most common case is that some high frequency content is lost. If the loss of bandwidth is not constant, but changing from frame to frame (e.g. every 24 ms) the effect becomes more objectionable than in the case of a constant bandwidth reduction.

#### *5.1.2. Preechoes*

Preechoes are a very common and famous possible artifact for high frequency resolution perceptual coding systems. The name *preecho* is somewhat misleading: The basic coding artifact is noise spread out over some time even before the music event causing the noise. To understand preechoes, let us have a look at the decoder of a perceptual coding system (see Fig 1). The reconstructed frequency lines are combined in the synthesis filterbank. This filterbank consists of a modulation matrix and a synthesis window. The quantization error in the coder can be seen as a signal added to the original frequency line. The length (in time) of such a signal is equal to the length of the synthesis window. Thus, reconstruction errors are spread over the full window length. If the music signal contains a sudden increase in signal energy (like a castanet attack), the quantization error is increased, too. As explained above, this quantization error (noise) signal is spread over the length of the synthesis window. If the attack occured well within the analysis window, this error will precede the actual cause for its existence in time. If this prenoise extends beyond the premasking (backwards masking) period, it becomes audible and is called preecho. There are a number of techniques to avoid audible preechoes (including variable bit-rate coding, local increase in the bit-rate locally to reduce the amplitude of the preecho), but overall this type of artifact belongs to the most difficult to avoid.

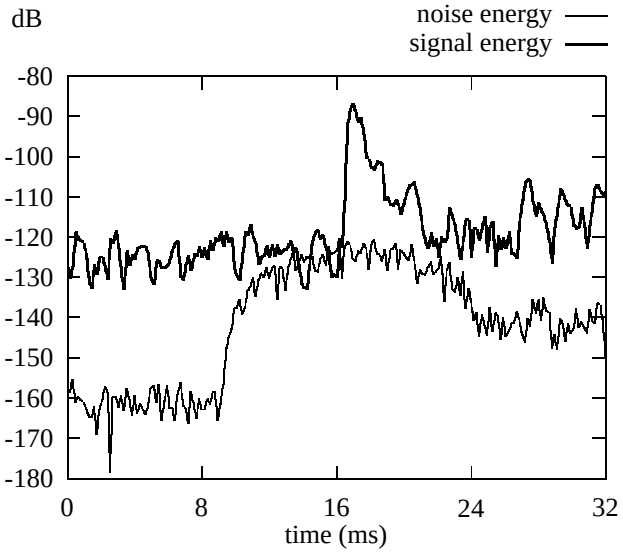


Figure 4: Example of a pre-echo.
The lower curve (energy of the noise signal) shows the form of the analysis window

#### *5.1.3. Roughness, double-speak*

Especially at low bit-rates and lower sampling frequencies there is a mismatch between time resolution of the coder (at least in its more efficient ”normal block” mode) and the requirements to follow the time structure of some signals. This effect is most noticeable for speech signals and for listening via headphones. The effect is sometimes called *double-speak* because a single voice tends to sound like the same recording was done twice (with the same speaker) and overlayed. AAC containes a technique called *Temporal Noise Shaping* (TNS) which provides noise shaping in time domain and thus virtually enhances the time resolution of the filterbank.

### 5.2. Dynamic range and frequency response

There are some kinds of deficiencies of standard audio equipment which cannot be found in properly designed Layer-3 and AAC codecs. They are listed here to mention the fact that it does not make sense to test for them. Most noticable are

- Dynamic range
  MP3 and AAC both contain a global gain adjustment parameter for every block of music data. According to the word length and resolution of this parameter, the dynamic range of both MP3 and AAC is well beyond the equivalent of a 24 bit D/A resolution. In short, MP3 and AAC represent the

music in a way that the dynamic range of every known audio source is perfectly retained.

- Frequency response
  Something similar is true for the frequency response of an MP3 or AAC encoder/decoder pair. Due to the nature of the coding process, the frequency response (on average) is perfectly even (0 dB deviation). There may be short term variations, but there is no coloration of the signal due to coding artifacts (of course with the exception of parts of the spectrum which are not transmitted at all).

### 5.3. Decoder differences

The decoders for both Layer-3 and AAC are fully specified in the relevant ISO standards. The conformance part of the standard gives the choice between standard "MPEG-1 audio decoders" and "high accuracy MPEG-1 audio decoders", but up to now all implementations followed the rules for high accuracy decoders. While the description is not fully accurate to the bit (some parts are defined via formulas allowing different methods of implementation), it is done in a way that there can be no audible differences between compliant decoders. This makes listening tests of MP3 or AAC decoders a moot exercise: The only question is whether a decoder is compliant to the standard or not. One example for non-standard decoders are implementations which do not allow switching of the bit-rate within decoding a compressed bitstream. The MPEG-1 standard specifically requires a decoder to adapt to changing bit-rates. If a decoder is not able to decode variable rate bitstreams, the decoder does not perform properly. Another feature required in the standard but not implemented in some (widely distributed) Layer-3 decoders is the support of intensity stereo coding.

### 5.4. Not all encoders are created equal

The MPEG standards leave the implementation of an encoder completely open. In the extreme case, one could completely avoid implementing the perceptual model, decide not to use the scalefactors (and therefore the outer iteration loop) and do a very simple inner iteration loop. Such an encoder would be very fast (potentially much faster than any current encoder product), compliant with the standard, produce nice audio quality for some signals (where the build-in noise shaping of the non-uniform quantizer is sufficient) but sound very bad for a large selection of music. While this project is easy, it is much more difficult to build an encoder with very high audio quality across all types of music and even for the most exotic test items. In MPEG, testing had always aimed to verify sufficient encoder performance in worst case scenarios. Nonetheless, the MP3 encoders around differ quite a bit in their ability to produce, in a persistent way, the best sounding compressed audio at low bit-rates.

#### *5.4.1. Reference encoders*

There are two sources for reference encoders:

- The MPEG committee's own software implementation (technical report)
  This encoder is the result of collaborative work of a large number of individuals. The basic goal of the effort was to provide a source for MPEG audio bitstreams with correct syntax and to give an implementation example helping people to understand the syntax. High audio quality was not a goal of the joint implementation effort. It can even be said that some of the companies participating in this effort are not interested to see a publicly available high quality implementation of Layer-3 or AAC encoding. Therefore, encoders based on the public source but without additional work on encoding strategies and perceptual model usually sound bad and should not be seen as examples for "MP3 sound quality".

- Encoders used for the verification tests
  In MPEG-1, verification tests used DSP based encoders. MPEG-2 AAC development was based on software simulation encoders. The source for the encoders in all cases was the same development lab which took the input from the MPEG Audio subgroup and built the hardware or software encoders for the tests. Only these encoders can be called reference encoders in terms of encoding quality. However, in all cases encoder development continues well after the verifications tests leading to the availability of improved encoders both from this lab and others.

There are a number of tradeoffs in the design of an MP3 or AAC encoder. Up to now, the best quality has been achieved using carefully tuned double iteration loop type of encoders. This follows the paradigm as described for the example encoder description in the MPEG standard. These encoders are very slow. Faster, but maybe somewhat lower quality encoders can be built by changing the iteration strategy. Other differences between encoders concern the psychoacoustic model, the strategy for switching between the short and long window encoding mode as well as the use of joint stereo coding. All this, combined with the difficulties of tuning the encoder to a given bit-rate and the choice of encoder bandwidth at low bit-rates (see below) leads to quite some variation between different encoders.

### 5.5. How to measure codec quality

To measure codec quality of high quality audio codecs has, over the last ten years, developed to an art of its own.

There are basically three measurement methods: Listening tests, simple objective measurement methods and perceptual measurement techniques.

#### *5.5.1. Listening tests*

To this date, large scale and well-controlled listening tests are still the only method available to compare the performance of different coding algorithms and different encoders. The ITU-R (International Telecommunications Union, Radiocommunications sector) with input from a number of broadcasters and the MPEG audio group, has developed a very elaborate set of rules for listening tests. The aim of these tests is to stress the encoders under worst case conditions , i.e. to find the material which is most difficult to encode and to exhibit the performance of the encoders under test for this material. This follows the observation that in a lot of cases coding artifacts become audible and even objectionable only after an extensive training period. Since the use of equipment based on audio compression technology (like memory based portable audio players) itself constitutes extensive training, we can expect that over time everybody becomes an expert listener. Therefore, from the beginning, encoders should better be tuned to satisfy the quality requirements of expert listeners.
The ITU-R test procedure requires a relatively large number of test subjects and the test to be done in a double blind fashion. More details about this type of testing can be found in [9].

#### *5.5.2. Simple objective measurement techniques*

Over and over again people have tried to get a measure of encoder quality by looking at units like Signal-to-Noise-Ratio or bandwidth of the decoded signal. Since the basic paradigm of perceptual coders relies on improving the subjective quality by shaping the quantization noise over frequency (and time), leading to an SNR lower than possible without noise shaping, these measurements defy the whole purpose of perceptual coding. The authors still using these measurements just show that they have not understood what they are doing. As explained below, to rely on the bandwidth of the encoded signal does not show much better understanding of the subject.
Another approach is to look at the codec result for certain test signal inputs (transients, multi-tone signals). While the results of such a test can tell a lot of information about the codec to the expert, it is very dangerous to rely solely on the results of such tests.

#### *5.5.3. Perceptual measurement techniques*

Beginning 15 years ago, there was a lot of research to apply psychoacoustic modeling to the prediction of codec quality and the audibility of certain artifacts. While the state of the art is not yet sufficient to make large scale and well-prepared listening tests obsolete, perceptual measurement techniques have progressed to the point where they are a very useful supplement to listening tests and can replace them in some cases. ITU-R Task Group 10/4 worked for a number of years on the standardization of perceptual measurement techniques and produced a recommendation on a system called PEAQ (Perceptual Evaluation of Audio Quality). The recommendation defines a multi-mode system based on the collaborative effort of all the leading laboratories working on perceptual measurement techniques. For a more detailed description of PEAQ, look for [4] in this conference proceedings.

### 5.6. Bit-rate versus quality

MPEG audio coding does not work with a fixed compression rate. The user can choose the bit-rate and this way the compression factor. Lower bit-rates will lead to higher compression factors, but lower quality of the compressed audio. Higher bit-rates lead to a lower probability of signals with any audible artifacts. However, different encoding algorithms do have ”sweet spots” where they work best. At bit-rates much larger than this target bit-rate the audio quality improves only very slowly with bit-rate, at much lower bit-rates the quality decreases very fast. The ”sweet spot” depends on codec characteristics like the Huffman codebooks, so it is common to express it in terms of bit per audio sample. For Layer-3 this target bit-rate is around 1.33 bit/sample (i.e. 128 kbit/s for a stereo signal at 48 kHz), for AAC it is around 1 bit/sample (i.e. 96 kbit/s for a stereo signal at 48 kHz). Due to the more flexible Huffman coding, AAC can keep the basic coding efficiency up to higher bit-rates enabling higher qualities. Multichannel coding, due to the joint stereo coding techniques employed, is somewhat more efficient per sample than stereo and again than mono coding. A good overview of the tradeoff between bit-rates and achievable quality for a number of coding algorithms (including AAC and MP3) can be found in [8].

### 5.7. The bandwidth myth

Reports about encoder testing often include the mention of the bandwidth of the compressed audio signal. In a lot of cases this is due to misunderstandings about human hearing on one hand and encoding strategies on the other hand.

#### *5.7.1. Hearing at high frequencies*

It is certainly true that a large number of (especially young) subjects are perfectly able to hear single sounds at frequencies up to and sometimes well above 20 kHz. However, contrary to popular belief, the author is not aware of any scientific experiment which showed beyond doubt that there is any listener (trained or not) able to detect the difference between a (complex) musical signal

with content up to 20 kHz and the same signal, but bandlimited to around 16 kHz. To make it clear, there are some hints to the fact that there are listeners with such capabilities, but the full scientific proof has not yet been given. As a corollary to this (for a lot of people unexpected) theorem, it is a good encoding strategy to limit the frequency response of an MP3 or AAC encoder to 16 kHz (or below if necessary). This is possible because of the brick-wall characteristic of the filters in the encoder/decoder filterbank. The generalization of this observation to other types of audio equipment (in particular analog) is not correct: Usually the frequency response of the system is changed well below the cutoff point. Since any deviation from the ideal straight line in frequency response *is* very audible, normal audio equipment has to support much higher frequencies in order to have the required perfectly flat frequency response up to 16 kHz.

#### *5.7.2. Encoding strategies*

While loss of bandwidth below the frequency given by the limits of human hearing is a coding artifact, it is not necessarily the case that an encoder producing higher bandwidth compressed audio sounds better. There is a basic tradeoff where to spent the bits available for encoding. If they are used to improve frequency response, they are no longer available to produce a clean sound at lower frequencies. To leave this tradeoff to the encoder algorithm often produces a bad sounding audio signal with the high frequency cutoff point varying from block to block. According to the current state of the art, it is best to introduce a fixed bandwidth limitation if the encoding is done at a bit-rate where no consistent clean reproduction of the full bandwidth signal is possible. Technically, both MP3 and AAC can reproduce signal content up to the limit given by the actual sampling frequency. If there are encoders with a fixed limited frequency response (at a given bit-rate) compared to another encoder with much larger bandwidth (at the same bit-rate), experience tells that in most cases the encoder with the lower bandwidth produces better sounding compressed audio. However, there is a limit to this statement: At low bit-rates (64 kbit/s for stereo and lower) the question of the best tradeoff in terms of bandwidth versus cleanness is a hotly contested question of taste. We have found that even trained listeners sometimes completely disagree about the bandwidth a given encoder should be run at.

### 5.8. Tuning for different bit-rates

As explained above, the double iteration loops do not converge if there is a mismatch between the coding requirements as given by the perceptual model and the bit-rate available to code a block of music. To avoid this situation, it is wise to set the parameters in the psychoacoustic model in a way that the iteration loops will normally converge. This may require settings which lead to audible differences, but the final coding result is still better than the one from a perceptual model set to avoid any audible difference combined with coding loops which do not converge in a sensible way. To achieve this balance between requirements from the perceptual model and bit-rate, the coding parameters have to be readjusted if the encoder is run at different bit-rates. This tuning procedure can be responsible for a large share of the development effort for an MP3 or AAC encoder.

## 6. FILE FORMATS

The MPEG standards define the representation of audio data. Above this, MPEG as well defines how to put the coded audio into a bitstream with synchronization and header info sufficient to do the proper decoding without any additional information given to the decoder.

### 6.1. MPEG-1/2 Layer-3 header format

MPEG-1/2 defines a mandatory header format which is contained in every frame (every 24 ms at 48 kHz sampling frequency). It contains, among others the following data:

- Sync word
  Unlike in other standards, the sync word *can* occur within the audio data, too. Therefore a proper synchronization routine should check for the occurance of more than one sync word in the right distance and should completely resync only if no more sync words are found at the right distance as given by the bit-rate and sampling frequency.

- Bit-rate
  The bit-rate is always given for the complete audio stream and not per channel. In the case of Layer-3, it is specifically allowed to switch the bit-rate on the fly, leading to variable bit-rate encoding.

- Sampling frequency
  This will switch the decoder hardware (or software) to different sampling frequencies, like 32 kHz, 44.1 kHz or 48 kHz in the case of MPEG-1.

- Layer
  The header contains information whether this is a Layer-1, Layer-2 or Layer-3 bitstream (all share the same header structure) and whether this is MPEG-1 or MPEG-2 low sampling frequency encoding.

- Coding mode
  Again, as a fixed parameter this allows to differentiate between mono, dual mono, stereo or joint stereo coding.

- Copy protection
  Each header carries the two bits for the SCMS (Serial Copy Management Scheme). However, since the ease of manipulating these bits via software, the practical importance of this way of copy protection is relatively minor.

Due to the repetition of all information necessary to do a successful decode in every frame, MPEG-1/2 bitstreams are self sufficient and allow to start decoding at any point in time. A properly built decoder even can read over other information attached to the begin of an audio file (like RIFF/WAV headers or metadata describing the content) and then just start decoding the audio.

### 6.2. MPEG-2 AAC audio transport formats

While in MPEG-1 the basic audio format and the transport syntax for synchronization and coding parameters are tied together in an unseparable way, MPEG-2 AAC defines both, but leaves the actual choice of audio transport syntax to the application. The standard defines two examples for the transport of audio data:

- ADIF
  The "Audio Data Interchange Format" puts all data controlling the decoder (like sampling frequency, mode etc.) into a single header preceeding the actual audio stream. Thus it is useful for file exchange, but does not allow for break-in or start of decoding at any point in time like the MPEG-1 format.

- ADTS
  The example "Audio Data Transport Stream" format packs AAC data into frames with headers very similar to the MPEG-1/2 header format. AAC is signaled as the (otherwise non-existant) "Layer-4" of MPEG Audio. Unlike Layer-3, the frame rate is variable, containing always the audio data for a complete frame between two occurances of the sync word. ADTS again allows start of decoding in the middle of an audio bitstream. The ADTS format has emerged as the de-facto standard for a number of applications using AAC.

## 7. CONCLUSIONS

Using an encoder with good performance, both Layer-3 and MPEG-2 Advanced Audio Coding can compress music while still maintaining near-CD or CD quality. Among the two systems, Layer-3 at somewhat lower complexity is the system of choice for current near-CD quality applications. AAC is its designated successor, providing near-CD quality at larger compression rates (increasing the playing time of flash memory based devices by nearly 50 % while maintaining the same quality compared to Layer-3) and enabling higher quality encoding and playback up to high definition audio (at 96 kHz sampling rate). AAC, together with copyright protection systems as defined by the Secure Digital Music Initiative (SDMI), will be the compression system of choice for future Electronic Music Distribution (EMD) and thus follow in the footsteps of worldwide adoption of other MPEG defined compression algorithms like MPEG Audio Layer-2, MPEG Audio Layer-3 or MPEG Video.

## ACKNOWLEDGEMENTS

The author would like to thank all collegues at Fraunhofer IIS-A and MPEG Audio for all the wonderful collaborative work over the last 11 years since the start of MPEG. Special thanks are due to Jürgen Herre, Harald Popp, Martin Dietz, Oliver Kunz and Jürgen Koller for helpful suggestions. Part of the audio coding work at Fraunhofer IIS-A has been supported by the Bavarian Ministry for Economy, Transportation and Technology and the European Commission (within the RACE and ACTS programmes).

## REFERENCES

[1] M. Bosi, K. Brandenburg, Sch. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Yoshiaki Oikawa. ISO/IEC MPEG-2 Advanced Audio Coding. In *Proc. of the 101st AES-Convention*, 1996. Preprint 4382.

[2] K. Brandenburg and Marina Bosi. Overview of MPEG audio: Current and future standards for low bit-rate audio coding. *J. Audio Eng. Soc.*, 45(1/2):4 – 21, January/February 1997.

[3] K. Brandenburg and G. Stoll. ISO-MPEG-1 Audio: a generic standard for coding of high quality digital audio. In N. Gilchrist and Ch. Grewin, editors, *Collected Papers on Digial Audio Bit-Rate Reduction*, pages 31 – 42. AES, 1996.

[4] C. Colomes, C. Schmidmer, and W.C. Treurniet. Perceptual-quality assessment for digital audio: Peaq – the proposed itu standard for objective measurement of perceived audio quality. In *Proceedings of the AES 17th. International Conference*, 1999.

[5] MPEG. Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s, part 3: Audio. International Standard IS 11172-3, ISO/IEC JTC1/SC29 WG11, 1992.

[6] MPEG. Information technology — generic coding of moving pictures and associated audio, part 3: Audio. International Standard IS 13818–3, ISO/IEC JTC1/SC29 WG11, 1994.

[7] MPEG. MPEG–2 advanced audio coding, AAC. International Standard IS 13818–7, ISO/IEC JTC1/SC29 WG11, 1997.

[8] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault. Subjective evaluation of state-of-the-art 2-channel audio codecs. *J. Audio Eng. Soc.*, 46(3):164 – 176, March 1998.

[9] G. A. Soulodre and M. Lavoie. Subjective evaluation of large and small impairments in audio codecs. In *Proceedings of the AES 17th. International Conference*, 1999.