Psychoacoustics

The human hearing is well studied and there are good models for it.

This can be used when coding audio in order to place the distortion (quantization noise) so that a human listener will notice it as little as possible. All modern audio coding methods use psychoacoustics in the coding, which gives a significantly better subjective quality, compared to just maximizing SNR.

Sound pressure level

A sound is a pressure wave with pressure p.
The sound pressure level (SPL) is defined as

$$\mathrm{SPL} = 10 \cdot \log_{10}(\frac{p}{p_0})^2 \quad [\mathrm{dB}]$$

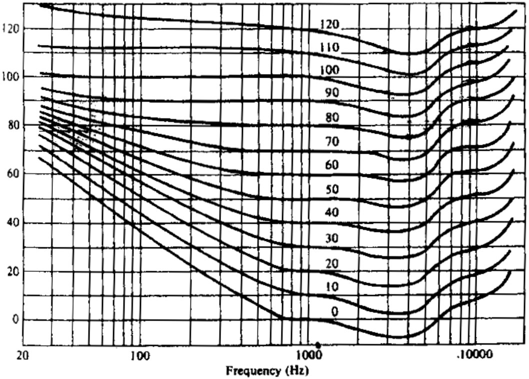
where $p_0 = 20\mu Pa$ is the smallest sound pressure that a human can hear. Often sound is described using the intensity I, which is the power per surface area (W/m^2) of the sound wave. I is proportional to p^2. Thus, the SPL can also be calculated as

$$\mathrm{SPL} = 10 \cdot \log_{10} \frac{I}{I_0} \quad [\mathrm{dB}]$$

where $I_0 = 10^{-12}\ W/m^2$ is the intensity of a sound wave of pressure p_0.

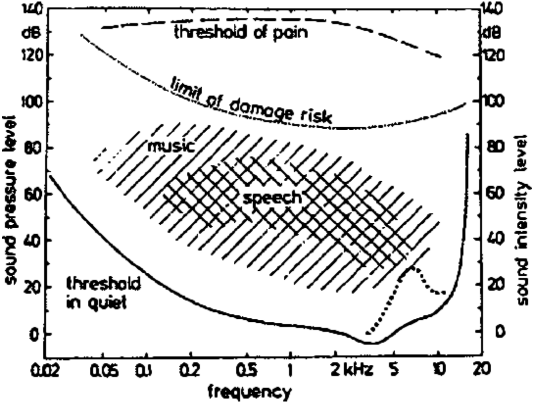
Loudness

The loudness of a sound is defined as the sound pressure level of a tone at 1 kHz that is perceived as being as loud as the sound. The loudness will depend on intensity, frequency content and duration. Loudness is a subjective measure.



The figure shows how the loudness varies with level and frequency for tone signals.

Hearing range



Loudness 0 dB marks the limit of how weak sounds that a human can perceieve, the so called hearing threshold. The figure shows the whole range of human hearing and typical values for speech and music.

Sampling frequency

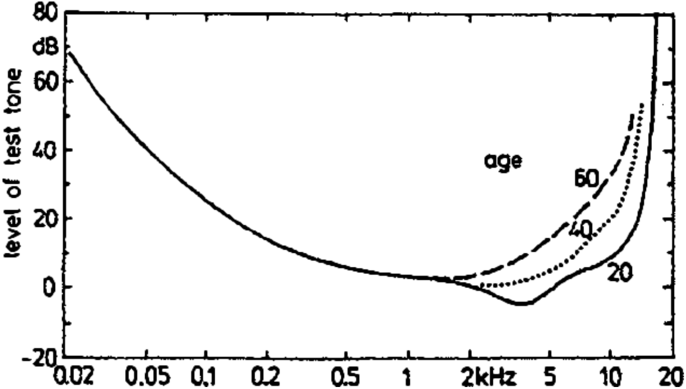
Humans have trouble hearing frequencies above 20 kHz, which will influence the choice of sampling frequency.

On CD:s the sampling frequency is 44.1 kHz.

Movie audio usually have a sampling frequency of 48 kHz.

Formats intended for high quality audio usually allow even higher sampling frequencies, such as 96 kHz or 192 kHz.

Hearing threshold



The figure show how the hearing threshold typically varies with age.

Hearing threshold

The hearing threshold is important in audio coding.

Frequency components of the input signal that are below the threshold can be removed without a listener noticing any difference.

If the quantization noise of the frequency components that are sent is below the threshold it will not be noticable.

Note however, that when we're coding audio we have no control of what volume the listener will use. Quantization errors that will not be audible at one volume setting might be noticable when the volume is increased. A common assumption when coding is that the quantization of the original data (eg 16 bits/sample for CD quality audio) is such that the quantization noise is just below the hearing threshold.

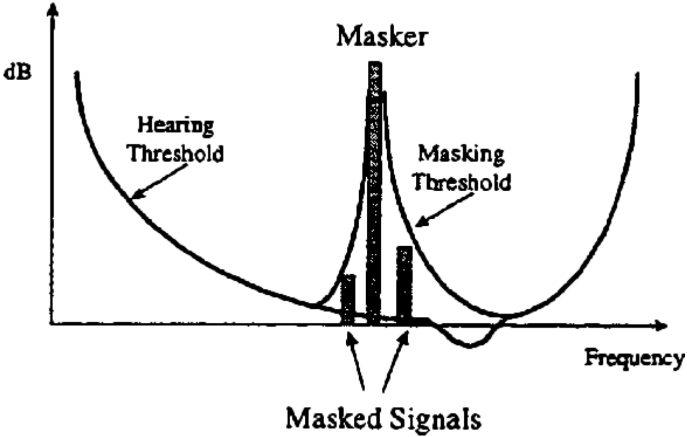
Masking

Besides the hearing threshold we can use a phenomenon called *masking* when coding audio. Masking means that strong sounds will mask weak sounds. That means that these weak sounds can either be quantized more coarsely or be removed completely from the signal.

Masking works both in time (temporal masking) and in frequency (frequency masking or spectral masking).

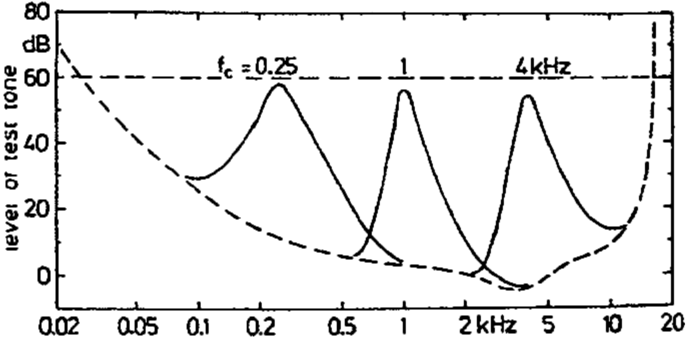
The difference in level for a signal component and the masking curve it gives rise to is called the *signal to mask ratio, SMR*. The higher the SMR is, the less masking we have.

Frequency masking



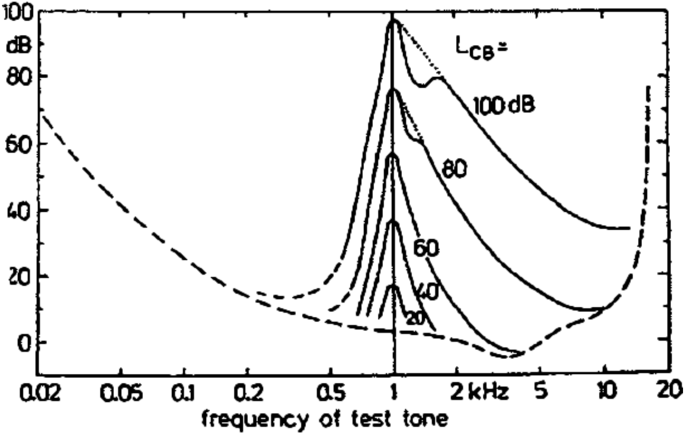
A strong sound masks weaker sounds at nearby frequencies.

Narrow band noise



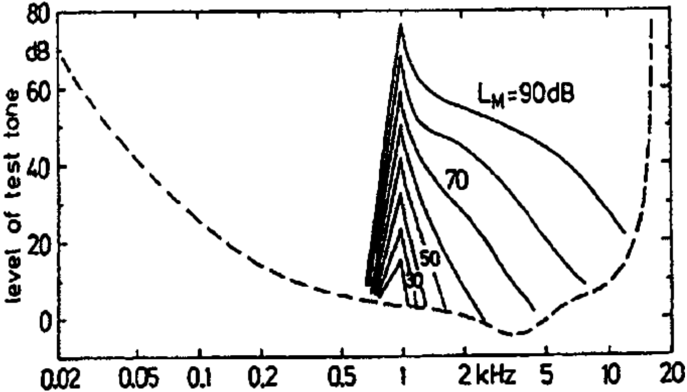
The figure shows how narrow band noise masks tone signals at different frequencies.

Narrow band noise



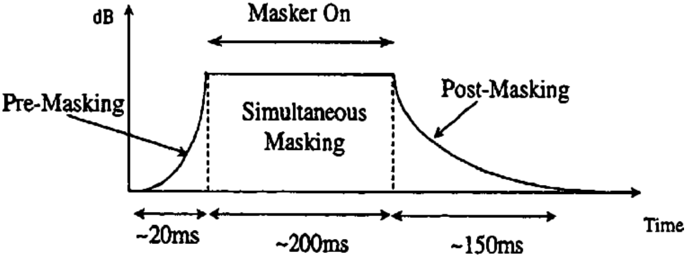
The figure shows how narrow band noise masks tone signals at different levels.

Tones



The figure shows how a tone at 1 kHz masks other tone signals.

Temporal masking



A strong sound masks weaker sounds both before (pre-masking) and afterwards (post-masking). The post-masking lasts for a longer time than the pre-masking.

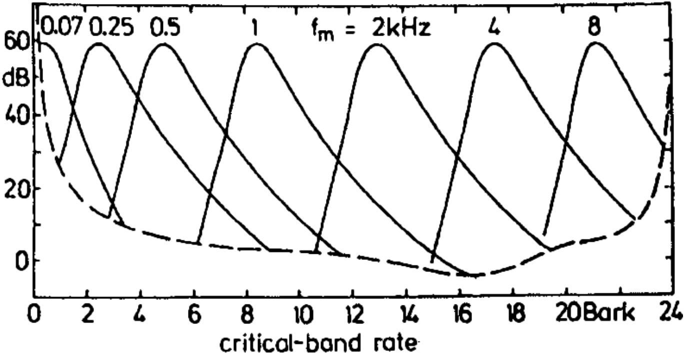
Critical bands

As seen in previous figures, the width of the masking will depend on the frequency of the masking signal. This is linked to how the hearing physically works. Often the signal is described in the Bark scale, where the frequencies are divided into bands corresponding to the frequency resolution of the human hearing. These bands are usually called *critical bands*.

The Bark scale

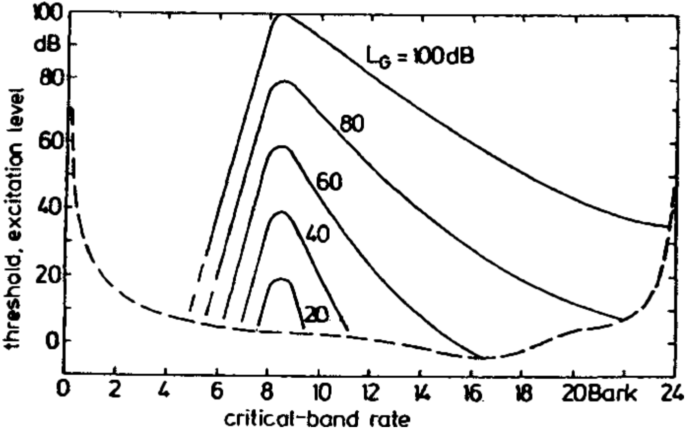
z	f_l	f_u	Δf	z	f_l	f_u	Δf
Bark	Hz	Hz	Hz	Bark	Hz	Hz	Hz
0	0	100	100	13	2000	2320	320
1	100	200	100	14	2320	2700	380
2	200	300	100	15	2700	3150	450
3	300	400	100	16	3150	3700	550
4	400	510	110	17	3700	4400	700
5	510	630	120	18	4400	5300	900
6	630	770	140	19	5300	6400	1100
7	770	920	150	20	6400	7700	1300
8	920	1080	160	21	7700	9500	1800
9	1080	1270	190	22	9500	12000	2500
10	1270	1480	210	23	12000	15500	3500
11	1480	1720	240	24	15500	22050	6550
12	1720	2000	280				

Narrow band noise



The figure shows how the masking looks for narrow band noise, described in the Bark scale. The masking curve is basically the same for all frequencies (compare to the previous figure).

Narrow band noise



Masking for narrow band noise at 1 kHz for different levels.

Psychoacoustic models

When coding audio, the signal is typically coded in blocks of on the order of 1000 samples.

To use psychoacoustics we measure the frequency content of each block (preferrably in the Bark scale). We usually try to classify the different bands as either tones or noise. Prototype masking curves for the different bands are combined with the hearing threshold to give a total SMR curve.

Bits are allocated to the different bands so that the SNR (if possible) is larger than the SMR for all bands. If we don't have enough bits to allocate, we allocate the bits so that the total difference between SNR and SMR is maximized.

The difference between SNR and SMR is called the *mask to noise ratio*, MNR.

MDCT

A popular transform in audio coding is MDCT (modified DCT) which is an extension of the normal DCT to overlapping blocks. Assume that $x_t(k)$ are samples in the signal domain and $X_t(m)$ samples in the transform domain for block t.

$$X_t(m) = \sum_{k=0}^{n-1} w(k) \cdot x_t(k) \cdot \cos(\frac{\pi}{2n}(2k + 1 + \frac{n}{2})(2m + 1))$$

for $m = 0, \ldots, \frac{n}{2} - 1$

$w(k)$ is a windowing function that need to fulfill certain demands:

$$w(k) = w(n - 1 - k), \quad w(k)^2 + w(k + \frac{n}{2})^2 = 1$$

The transform thus takes n samples and transform them to $\frac{n}{2}$ transform components. For the next block we take the last $\frac{n}{2}$ samples from the previous block and $\frac{n}{2}$ new samples.

MDCT

Inverse transform:

$$y_t(p) = w(p) \cdot \frac{4}{n} \sum_{m=0}^{\frac{n}{2}-1} X_t(m) \cdot \cos(\frac{\pi}{2n}(2p+1+\frac{n}{2})(2m+1))$$

for $p = 0, \ldots, n-1$

$$x_t(q) = y_{t-1}(q+\frac{n}{2}) + y_t(q)$$

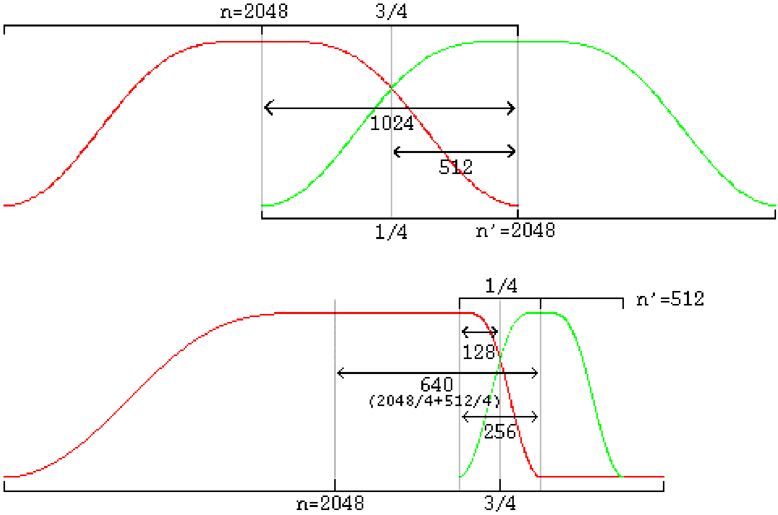
for $q = 0, \ldots, \frac{n}{2}-1$

Examples of windowing functions:

$$w(k) = \sin(\frac{\pi}{n}(k+\frac{1}{2}))$$

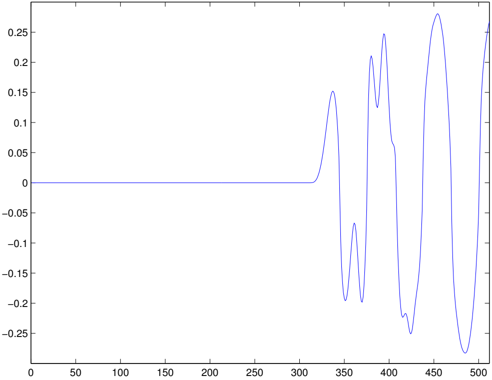
$$w(k) = \sin(\frac{\pi}{2}\sin^2(\frac{\pi}{n}(k+\frac{1}{2})))$$

Windowing in MDCT



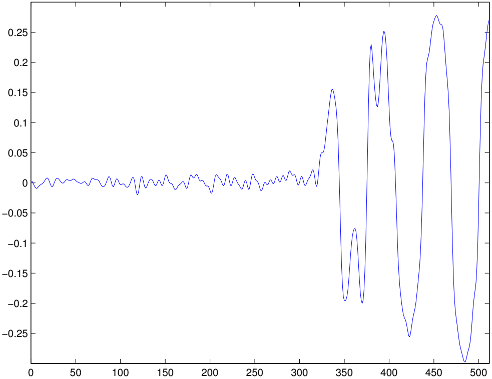
Block size vs quantization noise

A block of 512 samples from a music signal. The signal is quiet in the beginning and roughly halfway into the signal the music starts.



Block size vs quantization noise

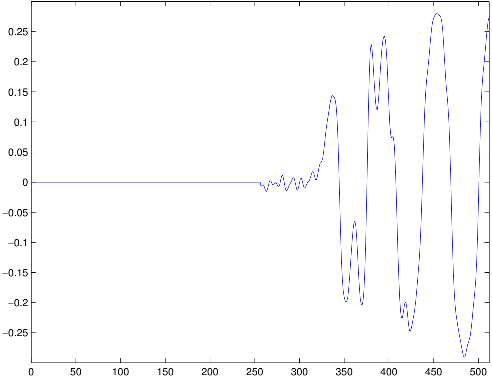
The block is transformed (DCT), quantized and inverse transformed.



The quantization of the transform components makes it so that the part of the signal that was quiet now contains noise. If the block size is too big the temporal masking might not be enough, so that we can hear some noise before the actual music starts (*pre-echo*).

Block size vs quantization noise

If we instead split the block into two blocks of 256 samples and transform and quantize we get



The noise before the music is now much shorter and is more easily masked. In audio coders it is common to be able to choose between different blocksizes to be able to adapt the coding to the signal.

MPEG-1 audio

The audio coding part of the video coding standard MPEG-1.

Supports sampling frequencies of 32, 44.1 and 48 kHz.

One or two channels:

- Mono
- Two mono channels (Dual channel)
- Stereo
  - Simple stereo: One right channel and one left channel.
  - Joint stereo: Utilize the dependency between the channels.

Rates of 32 to 224 kbit/s per channel.

Joint stereo

The dependency between the two channels can be used in two different ways.

- Middle/Side stereo:
  Instead of coding left and right channels we code the sum and difference between the channels. The difference channel often contains little information and is therefore easy to code.
- Intensity stereo:
  For high frequencies the stereo information is given more by the time envelope than by the actual frequency content. For frequencies over some value only one channel of frequency content is coded, but we send different scaling factors for the right and the left channel.

Middle/Side and Intensity stereo can be combined.

MPEG-1 layers

Three different levels (layers) of compression, with different complexity and compression ratios.

- Layer I
  Simple, suitable for $>$ 128 kbit/s per channel (DCC)
- Layer II
  More complex, suitable for rates around 128 kbit/s per channel (DAB)
- Layer III
  Largest complexity, gives the most compression, acceptable quality at 64 kbit/s per channel (mp3)

All three layers are based on a subband coder using 32 equally wide frequency bands.

MPEG-1, layer I

Coded in *frames* of 384 sampel (12 samples from each subband).

A psychoacoustic model is used to allocate bits to the different subbands. Each subband is given between 0 and 15 bits and a scale factor used to scale the signal before quantization. The quantization is uniform.

The scale factors and the bit allocation is sent as side information in the bitstream.

MPEG-1, layer II

Coded in frames of 1152 samples ($3 \cdot 12$ samples from each subband).

A psychoacoustic model is used to allocate bits to the different subbands. Each subband is given between 0 and 15 bits and 0-3 scale factors (each group of 12 samples can have its own scale factor, or they can share scale factors). The quantization is uniform.

The scale factors and bit allocation is sent as side information in the bitstream.

MPEG-1, layer III

1152 samples/frame. First the same subband transform as in layer I and II is used. Thereafter a MDCT is used in each subband, with blocksize 12 or 36, to get a finer frequency resolution which can be used to describe the signal in the Bark scale.

Unlike layer I and II a type of compander quantization is used. Scale factors are given to different frequency bands, corresponding to critical bands.

Static Huffman coding (fixed codewords) are used to get an even lower rate.

The standard also allows surrounding frames to use the available bits if they require a higher rate than the current frame.

Vorbis

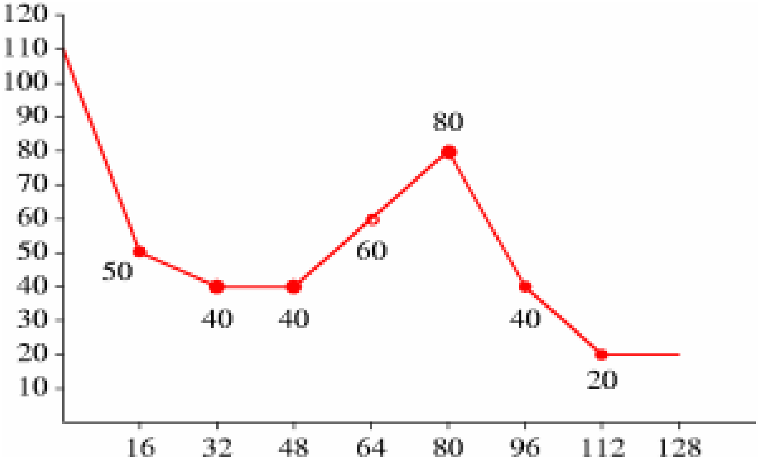
Open source project for an audio coder that contain no patented parts.

MDCT using block sizes between 64 and 8192 (powers of two).

The envelope of the transform data is calculated and sent compactly as either the frequency response of a linear filter or as a piecewise linear (in dB scale) function.

The “residual”, ie the transform data divided by the envelope is vector quantized and Huffman coded.

Envelope type 1



Quantization and Huffman coding

A number of different vector quantizers are sent as side information. The number of dimensions is maximum 65535. In practice only small dimensions are used.

The quantizers are either general (explicit values of all vectors in the codebook), or lattice quantizers (the reconstruction vectors are placed in a regular pattern).

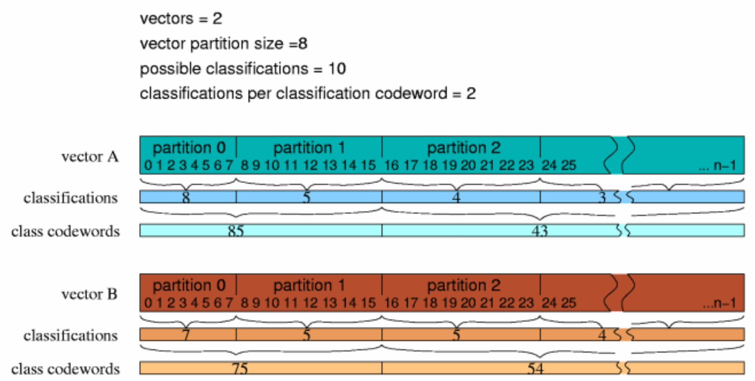
For each quantizer a Huffman code is also sent.

Residual coding

Each channel is split into a number of equally sized partitions. For each partition is specified which quantizers that are used. Either only one quantizer is used, or several passes of quantization (multi-stage VQ).

There is also a possibility to interleave several channels into one channel before quantization.

Quantization pass



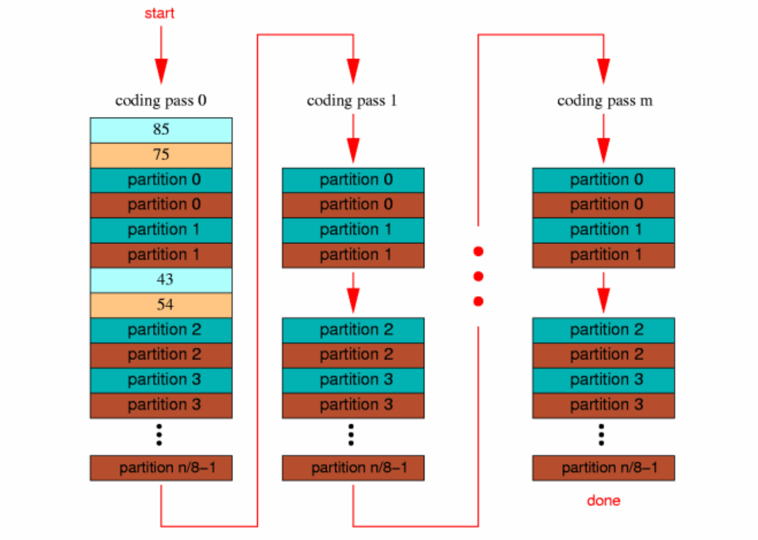
vectors = 2

vector partition size =8

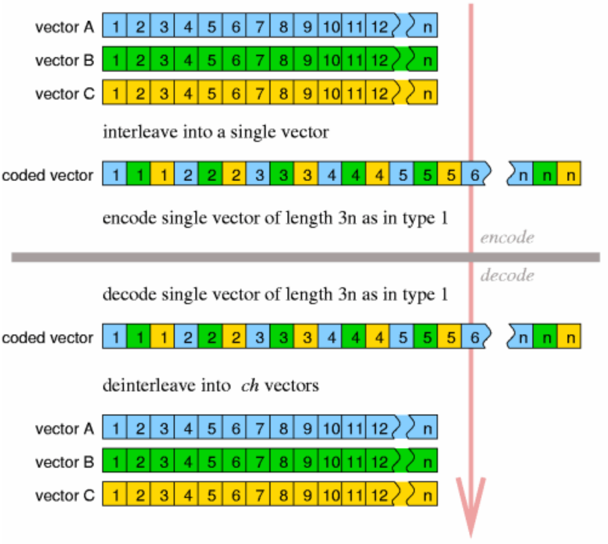
possible classifications = 10

classifications per classification codeword = 2

Quantization pass



Interleaving


vector A
1 2 3 4 5 6 7 8 9 10 11 12 n
vector B
1 2 3 4 5 6 7 8 9 10 11 12 n
vector C
1 2 3 4 5 6 7 8 9 10 11 12 n
interleave into a single vector
coded vector
1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 n n n
encode single vector of length 3n as in type 1
encode
decode
decode single vector of length 3n as in type 1
coded vector
1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 n n n
deinterleave into *ch* vectors
vector A
1 2 3 4 5 6 7 8 9 10 11 12 n
vector B
1 2 3 4 5 6 7 8 9 10 11 12 n
vector C
1 2 3 4 5 6 7 8 9 10 11 12 n

Channel correlation

The correlation between the channels (for stereo the right and left channels) can be utilized in different ways. By using interleaving the dependency can by used by the vector quantizer.

Another option is to code the largest value and the difference between the largest and the smallest value. The difference signal can then be quantized harder than the max signal. In the extreme case it is removed completely. The only stereo information is then given by the different envelopes of the channels (compare to intensity stereo in mp3)

Dolby Digital

Supports sampling frequencies of 32, 44.1 and 48 kHz.

Up to 5+1 channels.

Rate between 32 and 640 kbit/s.

Audio coded in frames consisting of 1536 samples. MDCT, mainly using block size $n = 512$. If a block contains a transient from low to high amplitudes, MDCT with block size $n = 256$ is used instead.

Dolby Digital, cont.

Channels can be combined so that only one channel of frequency content plus scale factors for all channels are sent (compare to intensity stereo in mp3).

For pure stereo signals we can code either a left and a right channel, or a sum and a difference channel (compare middle/side stereo in mp3).

The transform components are coded using an exponent and a normalized mantissa.

The bit allocation is not sent explicitly. Instead a SMR curve is sent and the decoder makes the bit allcation of the mantissas according to the curve and the decoded exponents. Naturally the coder should perform exactly the same bit allocation.

AAC

Advanced Audio Coding

Coding method first standardized in MPEG-2 and later extended in MPEG-4.

MDCT using blocksizes $n = 2048$ or $n = 256$.

Compander quantization.

"Huffman coding" of quantized coefficients (fixed codes to chooose among, 2 or 4 coefficients at a time).

In MPEG-4 arithmetic coding and vector quantization was added as options.

Spectral Band Replication (SBR)

A method where the high frequency content of the signal is removed before coding. At decoding the high frequency content is regenerated using the low frequency content of the signal. Some extra information (envelope) is added to the coded data to aid in the recreation of the high frequency content.

SBR can be used together with any other coding method. Combined with mp3 it is called mp3PRO. Combined with AAC it is called HE-AAC.

Claims to give 25-50 percents reduction in rate with no subjective loss of quality.

HE-AAC is used in DRM (Digital Radio Mondiale), which is used for digital radio via the shortwave and middlewave radio bands, and in DAB+.