Tutorial on Perceptual Audio Coding Algorithms

Markus Erne

Scopein Research
Sonnmattweg 6
5000 Aarau

Abstract

Digital audio devices, such as CD players or DAT-recorders, have become common during the past few years. A growing demand for high quality audio delivery and the increased requirements for the storage of digital audio data have motivated considerable research towards formulation of compression schemes which can satisfy both the conflicting demands of high compression ratios and transparent reproduction quality at the same time. This tutorial presents the principles of lossless and lossy audio compression schemes and will strongly focus on perceptual compression algorithms. Starting from the definition of entropy, lossless audio coding is discussed before perceptual coding will be addressed. Psychoacoustic principles, different quantizing and bit-allocation schemes, subband coding and transform coding will be presented in more detail in order to introduce different perceptual coding algorithms.

A short overview of the development of perceptual coding algorithms, including the different MPEG–Audio standards and the concept of different layers, used in MPEG-1/2 will be addressed before the following algorithms: MPEG-1, MPEG-AAC, Dolby AC3, ATRAC (Minidisc) and emerging standards, such as MPEG-4 and Wavelet-based coding schemes, are presented in more detail. In MPEG-4 audio, three different coder types are integrated into the standard: coders based on time-frequency mapping (General Audio-coders), Speech-coders and parametric coders each of which will be briefly presented.

1 Introduction

The central objective in audio coding or audio compression is to represent a digital audio signal with a minimum of bits per sample while keeping transparent (not distinguishable form the original under given listening conditions). Conventional digital audio devices (CD, DAT) typically sample audio signals at rates of 44.1 or 48 kHz, using linear PCM (Pulse Code Modulation) and a quantization of 16 Bits per sample. Therefore data rates in the range of 1,5 Mbit/s result for the storage or the transmission of a stereo signal. In contrast to lossless audio coding which is based on removing redundancy, lossy coding techniques make use of removing redundancy and irrelevancy based on perceptual criteria [1]. In archiving applications and for the high quality transmission or storage of audio signals, lossless coding is preferred in order to allow bit-identical reconstruction of the original signal at the decoder. For consumer and multimedia applications and for the transmission of audio signals using low bandwidth channels, a lossy compression scheme has to be used in order to guarantee a constant target bitrate. Perceptual coders are based on a psychoacoustic model and take advantage of the masking properties of the human auditory system. Most audio compression algorithms typically segment the input signal into blocks of 2ms up to 50ms duration. A time-frequency analysis then decomposes each analysis block in the encoder. This transformation or subband filtering scheme compacts the energy into a few transform coefficients and therefore de-correlates successive samples. These coefficients, subband samples or parameters are quantized and encoded according to perceptual criteria. Depending on the system objectives, the time-frequency analysis section might contain:

- Unitary transform (MDCT, FFT)
- Polyphase filterbank with uniform bandpass filters
- Time-varying, critically sampled bank of non-uniform bandpass filters
- Hybrid transform/filterbank scheme
- Harmonic/sinusoidal signal analyzer
- Source system analysis (LPC)

The time-frequency analysis approach always involves a fundamental tradeoff between time and frequency resolution requirements. The choice of the time-frequency analysis method additionally determines the amount of coding delay introduced, a parameter which may become important in duplex broadcast and live-events applications. The variety of existing musical instruments, such as castanets, harpsichord or pitch-pipe, create various coding requirements due to their completely different temporal and spectral fine –structure and suggest to use a filterbank with a flexible time-frequency tiling.

2 Principles of lossless coding

The source entropy [2] of a discrete source is defined as the average information (bit/symbol) generated per symbol if all M symbols are statistically independent and identically distributed:

$$H(A) = -\sum_{j=1}^{M} P(a_j)\log P(a_j)$$

With H(A): entropy of the signal
$P(a_j)$: probability distribution

Any information can be subdivided into 4 categories:

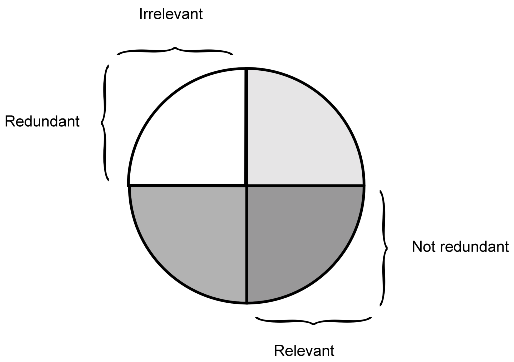


Fig 1: Partitioning into relevant vs. irrelevant parts and into redundant vs. non-redundant parts.

The notion of redundancy is connected to the entropy of a signal:
The entropy of a signal denotes the least amount of information in order to represent the signal in binary (or any other) number-format form. The redundancy is the difference between the amount of symbols which is currently used to represent the signal and the lower bound, the entropy. Therefore, the most efficient coding scheme will represent the signal with the least amount of symbols, equal to the amount of bits, given by the entropy and in such case, redundancy will be zero. Irrelevancy, denotes these parts of the signal which cannot be perceived by a human-being.

According to these categories, a digital compression scheme targets at removing either the irrelevant or redundant information part in a signal or both. A lossless coding scheme is designed to remove redundant parts in a signal and therefore will come as close as possible to the Shannon-bound of entropy. In practice, however, average compression ratios between 2 and 3,5 can be achieved for material in CD format (16bits, 44.1 kHz sampling rate) while the compression ratio will vary with the content of the audio signal, resulting in signal-dependent requirements in terms of transmission-bandwidth and storage-size of the media.
Lossless audio compression mostly is realized by using a combination of linear prediction or a transformation followed by entropy coding. The linear predictor attempts to minimize the variance of the difference signal between the predicted sample value and its actual value. The entropy coder (Huffmann, LZW) allocates short codewords to samples with high probability of occurance and longer codewords to samples with lower probability and in this way reduces the average bit consumption.
Lossless coding schemes allow to reconstruct the coded (original) signal on a bit by bit basis and therefore, by definition, cannot degrade the signal quality. Lossless audio coding is used in applications, such as archiving, multichannel sound storage, transmission of audio signals and forensic applications. Most lossless audio coders tend to have similar complexity for the encoder as well as for the decoder in terms of DSP instructions needed for realtime computation. Additionally, for the transmission and storage of losslessly compressed audio bitstreams, some additional error protection is required because error concealment on decorrelated samples becomes a difficult issue.

3 Principles of lossy coding

In contrast to a lossless coding system, a lossy compression schemes not only exploits the statistical redundancies but also the perceptual irrelevancies of the signal, as they result from the properties of the human auditory system.

3.1 The human auditory system

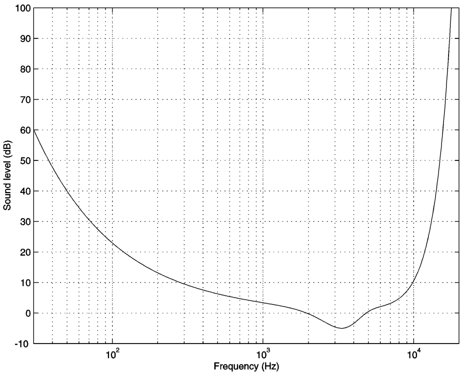
Fletcher [3] reported on test results for a large range of listeners, showing their absolute threshold of hearing, depending on the stimulus frequency. This *threshold in quiet* is well approximated by the non-linear function:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \text{ [dB-SPL]}$$

The absolute threshold of hearing is used in some coding schemes in order to scale the tolerable injected quantization noise but care has to be applied when using this threshold in perceptual audio coding since the final level of sound reproduction is usually not known at the time of the encoding of the signal.

In the cochlea, a frequency-to-place transformation takes place which leads to the notion of critical bands [4].

Figure 2: Absolute Threshold of hearing



Critical bandwidth can be considered as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrow band noise at constant sound pressure level remains constant within a critical band and then, begins to increase, once the bandwidth of the stimulus is increased beyond one critical band [5]. There is an almost infinite number of detectors on the cochlea with vastly overlapping frequency responses, each one having its own critical band width. The so called bark scale results if one puts together critical band widths such that their borders line up and in the literature about 25 bark bands can be found. .

It turns out that the critical bands at low frequencies show a constant bandwidth of around 100 Hz up to center frequencies of 500 Hz and at higher frequencies tend to turn into a more “constant-Q-type filter”.

Bark Scale	Lower edge f_l	Upper edge f_u	Band-width	Center Frequency
1	0	100	100	50
2	100	200	100	150
3	200	300	100	250
4	400	510	110	350
5	510	630	120	450
6	630	770	140	570
7	770	920	150	700
8	920	1080	160	1000
9	1080	1270	190	1170
10	1270	1480	210	1370
11	1480	1720	240	1600
12	1720	2000	280	1850
13	2000	2320	320	2150
14	2320	2700	380	2500
15	2700	3150	450	2900
16	3150	3700	550	3400
17	3700	4400	700	4000
18	4400	5300	900	4800
19	5300	6400	1100	5800
20	6400	7700	1300	7000
21	7700	9500	1800	8500
22	9500	12000	2500	10500
23	12000	15500	3500	13500
24	15500			

Figure 3: Critical band derived Bark scale

Perceptual coding algorithms profit from the masking properties of the human auditory system. Masking is a process where one sound (maskee) becomes inaudible in the presence of another sound (masker). Masking can occur in the time domain (temporal masking) or in the frequency domain (frequency domain masking). For the implementation of a psychoacoustic model in a perceptual coder we have to address the tone-masking-noise and the noise-masking-tone situation separately because their influence on the human auditory system is different [5].

3.1.1 Frequency domain masking

Frequency domain masking can be observed within critical bands (intra band masking) or across critical bands (inter band masking).

In the presence of a masker, the absolute threshold of hearing will be modified to become the masking threshold. All signals which are below the masking threshold are not perceived by the human auditory system and therefore the quantization noise in every subband can be as high as the masking threshold permits while maintaining subjectively perfect sound quality.

Some algorithms use a subband decomposition with frequency bands approximating the critical bands in order take advantage of frequency domain masking.

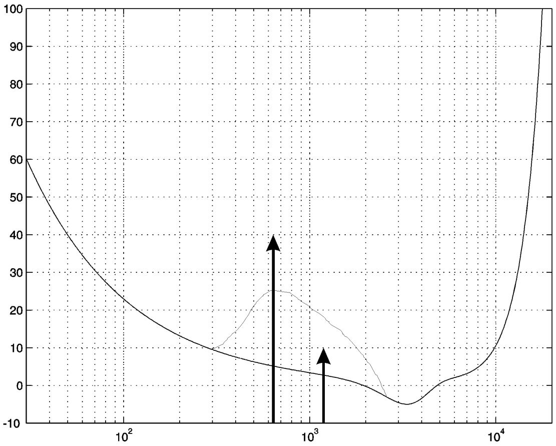


Figure 4: Masking effect, highlighted by the modified masking threshold in the presence of a masker and a maskee

The presence of a masker and a maskee generates an excitation in the inner ear which can be characterized by the masking threshold and the spreading function of the masking curve.

Assuming that the masker is uniformly quantized at m bits, then the Signal-to-Mask-Ratio (SMR) denotes the log distance of the masker to the masking threshold in this particular critical band and the Noise-to-Mask-Ratio (NMR) denotes the log distance of the quantization noise level and the masking threshold.

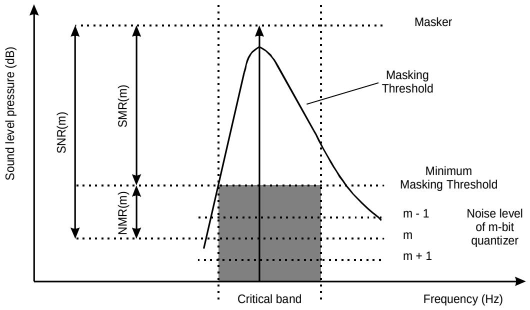

Masker
Masking Threshold
Minimum Masking Threshold
Noise level of m-bit quantizer
m - 1
m
m + 1
Sound level pressure (dB)
SNR(m)
SMR(m)
NMR(m)
Critical band
Frequency (Hz)

Figure 5: Frequency domain masking showing the signal to mask ratio and the noise to mask ratio

3.1.2 Temporal masking

Masking also occurs in the time domain. In the presence of abrupt signal transients, a listener will not perceive signals below the audibility threshold in the pre- and post-masking regions.

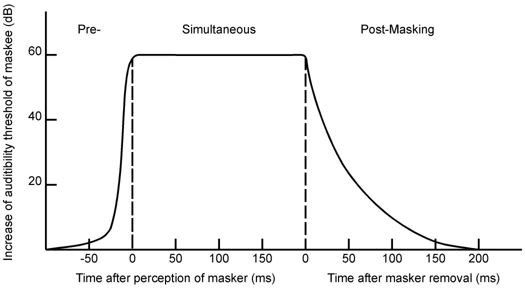


Figure 6: Temporal masking

Forward or post stimulatory masking depicts the masking effect that start at the end of the masking sound and lasts about two tenths of a second. This effect can physically be explained by the recovery time required by the hair cells after stimulation. Backward or pre-stimulatory masking refers to the phenomenon whereby the threshold of audibility is already raised before the onset of the masking sound. It is assumed that backward masking is caused by the interference of masker with the yet not incompleted auditory processing of the sound. Despite the fact that premasking only lasts about 5 ms, it can be used to compensate for pre-echo distortion in transform based coders. Postmasking could be integrated into the design of the filterbank but it is rarely implemented in current audio compression algorithms.

Although pre-echoes might be masked by the pre-masking effect, this is only true for small block sizes. Pre-echoes are a known problem of transform coders where an attack of a percussive sound (castanets) starts near the end of an analysis block. The inverse distortion preceding the signal attack.

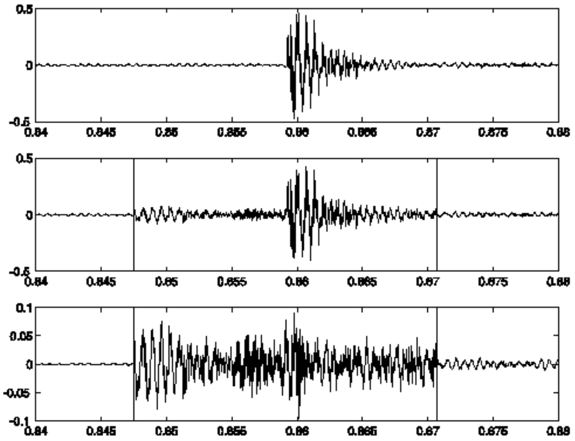


Figure 7: The effect of pre-echo distortion, demonstrated with a castanet signal, the coded version of it showing pre-echo-distortion, the difference signal between the original and the coded version.

4 Subband Decomposition

The filterbank is certainly one of the most important parts of a perceptual coding scheme. Despite the fact that there exist audio coding schemes using signal transforms (MDCT, Lapped Transform, Wavelet Transform) and others using filterbanks (Polyphase filters, Time varying, Quadrature Mirror Filter), they are mathematically almost equal and can be interpreted as filterbanks. Transform coders usually process the signals in blocks of samples whereas filterbanks based on convolutions may operate on a sample by sample basis (but can also perform subsampling). There are several options for the type of filterbank which directly will affect the computational complexity:

Uniform vs. Non-uniform filterbanks: uniform filterbanks are rather easy to implement and they are widely used in the ISO/MPEG Audio standard for the

Layer-1, Layer-2 and Layer-3 coders. Non-uniform filterbank often mimic the response of the human auditory system (e.g. critical band filterbanks) and they therefore better approximate the critical bands. Nevertheless, they may not necessarily be superior in terms of coding gain compared to a uniform filterbank.

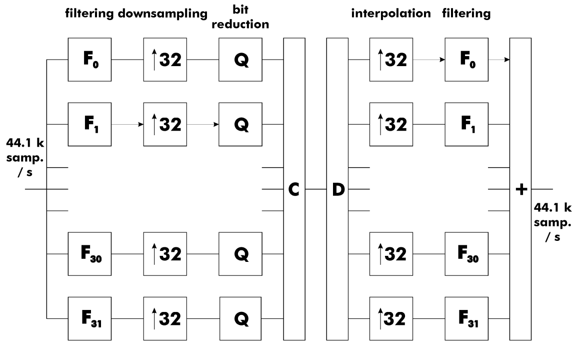
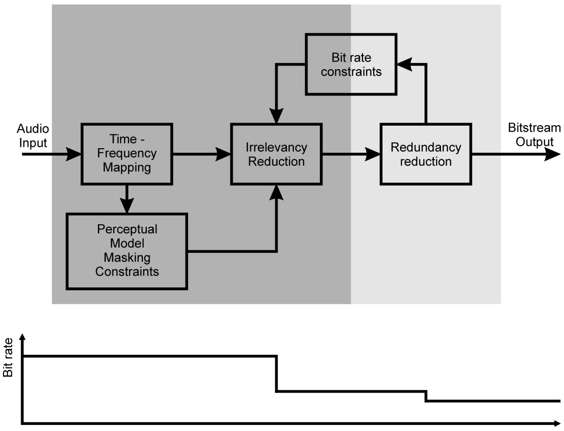


Figure 8: Audio Coder using a uniform polyphase filterbank

Static or Time-varying filterbank: Ideally the filterbank should adapt to the signal statistics in order to optimize the time-frequency tiling. Not all coding algorithms use time-varying filterbanks and window switching and block switching can be considered as a first, but also limited approach to a time-varying filterbank. Block-switching is used in order to avoid pre-echoes. In case of a transient, large block-sizes will create disturbing artifacts due to smearing of the quantization noise at the decoding throughout the whole block. Therefore a transient-detector will monitor the signal and in case of a transient, the block-size will be switched to typically 128 or even less samples. From this principle, it becomes pretty obvious these coding schemes have to use a “look-ahead” principle which might further increase the overall coding delay. In MLT based [10] algorithms, audio samples are windowed using overlapping of 50% or more. In case of a transient and a switching of the blocksize, perfect reconstruction of the filterbank only can be guaranteed when the size and the shape of the window are adapted as well. This process is called “window-switching” and is shown in Figure 8.



5 Block diagram

Figure 9: Block diagram of a basic perceptual audio coder

Combining the individual building blocks, we can now derive a block diagram of a basic audio coder. The time-frequency mapping (filterbank) decomposes the audio signal into subbands and therefore decorrelates successive samples. In the perceptual model, the masking threshold is estimated based on the presence of frequency domain and temporal masking present in the current block of audio data. The reduction in irrelevance is performed in the quantizer. The quantizer can be a uniform quantizer, a non-uniform quantizer or an adaptive quantizer. The quantizer is controlled according to the estimates of the masking threshold calculated by the perceptual model in order to guarantee that the quantization noise in each subband is below the masking threshold. After quantization, the quantized subband samples can be grouped in an efficient manner in the redundancy reduction block using entropy coding techniques. Figure 8 shows the reduction in bitrate resulting from the quantization, and the redundancy reduction.

6 Implementations

There exist numerous perceptual audio coding algorithms and most of them implement the principles explained earlier [6].

6.1 Application specific algorithms

Some audio compression algorithms have been developed for a specific application. ATRAC (Adaptive Transform Acoustic Coding) [ref] based on a combination of QMF filters and MDCT is used in Minidisc-recorders. AC-2 [ref] and AC-3 [ref], developed by Dolby Laboratories, are perceptual coding algorithms which are widely used in satellite links and surround sound applications for cinemas and the home theatre.

6.2 MPEG-1 and MPEG-2

MPEG-1 [7] was standardized in 1992 and has been developed for an overall target bitrate of 1,5 Mbit/s (audio & video).

The audio part of the MPEG-1 standard defines 3 different layers, each having different target bitrates, coding delays and complexity.

Layer-1, the implementation with the lowest complexity, initially has been developed for DCC-applications and was optimized for a bitrate of 192 kbit/s per channel.

Layer-2, a derivative of the MUSICAM-algorithm [1] has a target bitrate of 128 kbit/s per channel and is widely used in broadcast applications and DAB (Digital Audio Broadcasting [ref]).
Layer-3, the most complex algorithm within this family, targets for a high audio quality at bitrates down to 64 kbit/s or even less per channel. Layer-3 therefore can be used for low bitrate applications (satellite transmission, ISDN-applications, solid state memory recorders etc.). More recently, this coder has become extremely popular in the area of Internet audio and its nick name "MP3".

Finalized in 1994, MPEG-2 Audio provides two types of extensions to the MPEG-1 standard. The first extension allows to use lower sampling frequencies (16 kHz, 22.05 kHz and 24 kHz) for low bitrate applications. Additionally, multi-channel coding for surround sound applications (including the popular 5.1. configuration) has been added to the MPEG-2 standard in a way which is backward ("BC") compatible to MPEG-1 coding.
In order to enable better multi-channel compression performance than possible with the backward compatibility to MPEG-1 Audio, MPEG-2 Advanced Audio Coding (AAC) was added in 1997, initially named MPEG-2 NBC (Non Backward Compatible) coding. This addendum defines a new coding scheme, offering some advanced features which cannot be integrated into the MPEG-1 standard. Among these features AAC includes tools, such as TNS (Temporal Noise Shaping), Intensity Coupling, M/S-Stereo Coding and Gain Control. MPEG-2 NBC can represent up to 48 channels of compressed audio.

6.3 MPEG-4

The upcoming MPEG-4 standard has become an International Standard in 1999. In contrast to MPEG-1 and MPEG-2, MPEG-4 is a universal framework for the integration of tools, profiles and levels. MPEG-4 therefore not only includes a bitstream syntax and a set of compression algorithms but also offers a complete framework for synthesis, rendering, transport, compression and the integration of audio (speech, music etc.) and visual data. MPEG-4 audio is subdivided into natural and synthetic coding whereas natural audio integrates the following coding tools:

- HVXC low rate clean speech coder
- CELP wideband speech coder
- GA General Audio Coding (AAC-based)
- Twin VQ Additional Coding Tools to increase the coding efficiency at very low bitrates
-

The target bitrates of MPEG-4 range from 2 kbit/s up to 64 kbit/s per channel and, depending on the application, generic audio and speech coding or a combination of coding and synthesis may be used. MPEG-4 AAC [8] includes new tools, such as PNS (Perceptual Noise Substitution [14]) which allows to save transmission bandwidth for noise-like signals. Instead of coding the noise-like signals as spectral coefficients, a "noise flag" and the total noise power are transmitted and noise is re-synthesized at the decoder during the period of interest.
An additional feature of MPEG-4 is the concept of scalability. Certain subsets of the entire bitstream are sufficient for the decoding of the audio signal at a lower quality. This feature allows to trade bitrate versus quality and is extremely useful for applications where a certain bandwidth cannot be guaranteed (e.g. Audio on the Internet). In MPEG-4, scalability can be achieved in large steps of several kbit/s or by the use of a fine granularity mode. In the context of scalability, MPEG-4 can make use of a speech core coder and additional enhancement layers.

7 The next generation of audio coding algorithms

In the last few years, many high quality audio compression algorithms have been developed but most of them use a perceptual distortion measure and operate at different but fixed bitrates. A variety of these algorithms are based on uniform polyphase filterbanks, modified discrete cosine transforms [1], using window switching or, alternatively, lapped orthogonal transforms [10]. Many proposals for wavelet based audio coding [11] schemes have been published recently. Uniform polyphase filterbanks can be implemented efficiently, but they do not approximate the human auditory system well and they do not offer large coding gains in a rate-distortion metric. Among filterbanks or transforms which mathematically can be considered as being equal, Wavelet based filterbanks offer an interesting alternative to lapped orthogonal transforms. Wavelet filterbanks are known for a flexible time-frequency tiling [12] but most wavelet-based audio coding algorithms are focussed to mimic the response of the human auditory system.
Some new proposals have been published and they target to optimize the rate-distortion function based on perceptual criteria [13]. Nevertheless it is fair so say that audio coders using wavelet-packet transforms may exhibit other problems. The short support of the underlying basis functions (wavelets) creates filters which have limited stopband-attenuation. In an iterated filterbank, sidelobes appart from the center frequency may appear and may be filled with quantization noise which at the sidelobes position may become unmasked.
An interesting alternative to subband coding system are parametric coding schemes which are based on a analysis-synthesis approach. The signal is

decomposed in sinusoids and noise-like parts in the encoder and then at the decoder, these sinusoids and the noise-part are re-synthesized based on the transmitted control parameters. Parametric coders not only offer to transmit at extremely low bitrates down to a few kBit/s but due to the synthesis in the decoder offer interesting rendering of the audio signal at the decoder (volume control, pitch-change, tempo-change etc.)

8 Conclusions

A lot of effort has been put into the development of high quality low bitrate audio compression algorithms during the past few years. Audio coding still is a young field of research and the progress which could be achieved, especially thanks to the MPEG-standardization, is remarkable. Current research topics include the development of signal adaptive filterbanks, a more detailed understanding of the psychoacoustic principles for advanced perceptual models and the combination of coding and synthesis methods.

Future audio coding algorithms might be based on advanced rendering techniques which allow the decoder to render the audio based on a provided set of parameters. With increasing transmission bandwidth over the Internet, it is possible that the customer at home will have the possibility to perform his own mix at home, using a multi-channel bitstream or he can choose between different mixes.

Many thanks to Dr. Juergen Herre for reviewing this document.

REFERENCES

[1] Brandenburg K., Stoll G. et al., “The ISO/MPEG-Audio Codec: A Standard for Coding of High Quality Digital Audio”, *AES Convention Preprint 3336,* March 1992

[2] Jayant N., Noll P., “Digital Coding of Waveforms”, *Englewood Cliffs: Prentice Hall,* 1984

[3] Fletcher H., “Auditory Patterns”, *Rev. Mod. Phys.,* January 1940, pp. 47-65.

[4] Scharf B., “Critical Bands”,. *Foundations of Modern Auditory Theory, Academic Press,* 1970

[5] Zwicker E., Fastl H, “Psychoacoustics, Facts and Models” *Springer Verlag,* 1990

[6] Gilchrist N., Grewin C., “Collected Papers on Digital Audio Bit-rate Reduction”, *Audio Engineering Society*, September 1996

[7] Noll P., “MPEG Digital Audio Coding”, *IEEE SP Magazine,* September 1997, pp. 59-81

[8] Kahrs M., Brandenburg K.H. “Applications of Digital Signal Processing to Audio and Acoustics” . *Kluwer Academic Press,* 1999

[9] Erne M., Moschytz G. “Best Wavelet-Packet Bases for Audio Coding using Perceptual and Rate-distortion Criteria, *Proceedings of ICASSP 99*, May 1999

[10] Malvar H.S., “Signal Processing with Lapped Transforms”, *Artech House,* Norwood, 1992.

[11] Sinha D., Tewfik A., “Low Bit Rate Transparent Audio Compression using Adapted Wavelets”,. *IEEE Trans. on ASSP*, Vol. 41, No.12, December 1993, pp. 3463-3479.

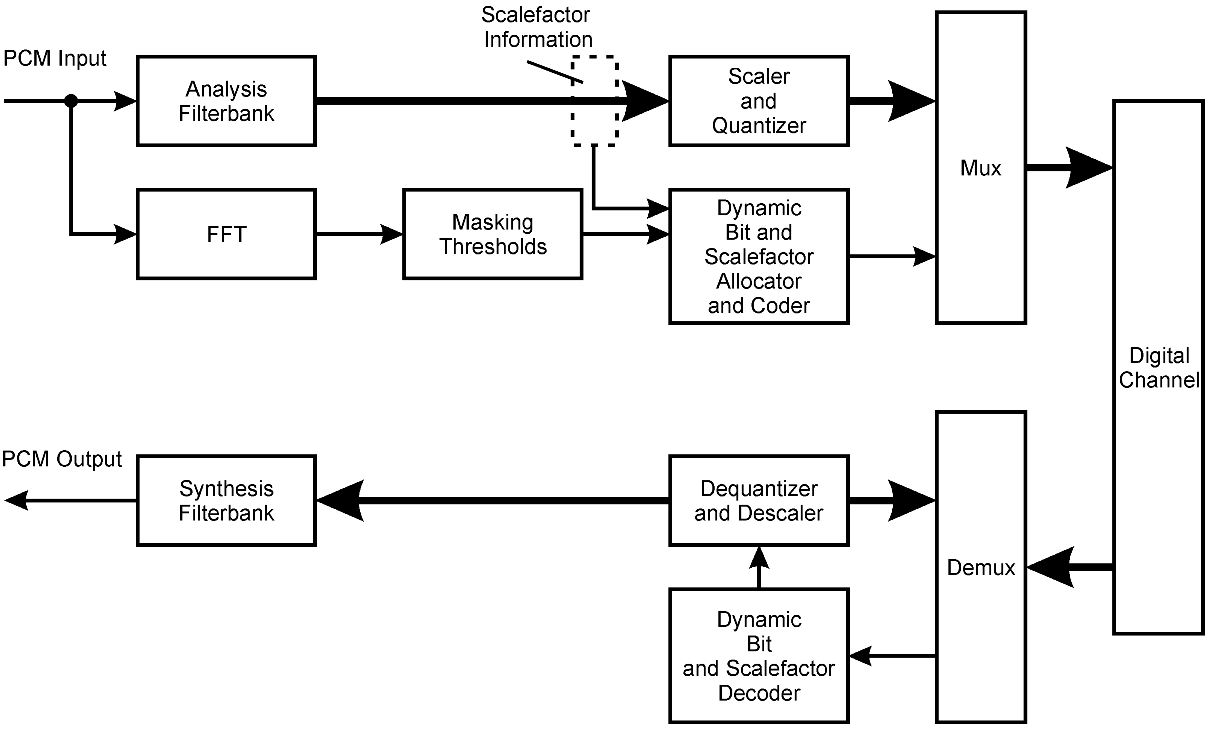
[12] Erne M., Moschytz G., “Perceptual and Near-Lossless Audio Coding based on Signal-adaptive Wavelet filterbank, *106-th AES Conference, May 1999*, Munich, Preprint No. 4934

[13] Erne M., Moschytz G., “Audio Coding Based on Rate-Distortion and Perceptual Optimization Techniques”, *Proc. of the AES 17-th International Conference on High Quality Audio Coding”,* Florence, September 1999, pp. 220- 225

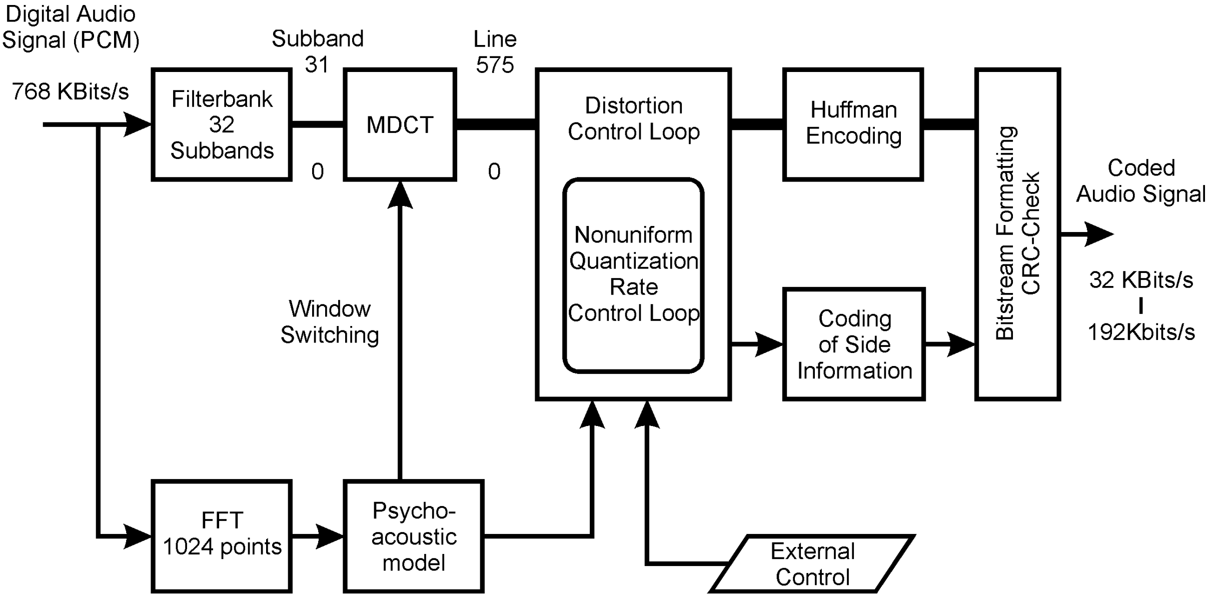
[14] Herre, J.; Jonston J.; “Enhancing the performance of perceptual coders by using Noise substitution”, *AES 104-th Convention Preprint,* #4720, 1998

APPENDIX

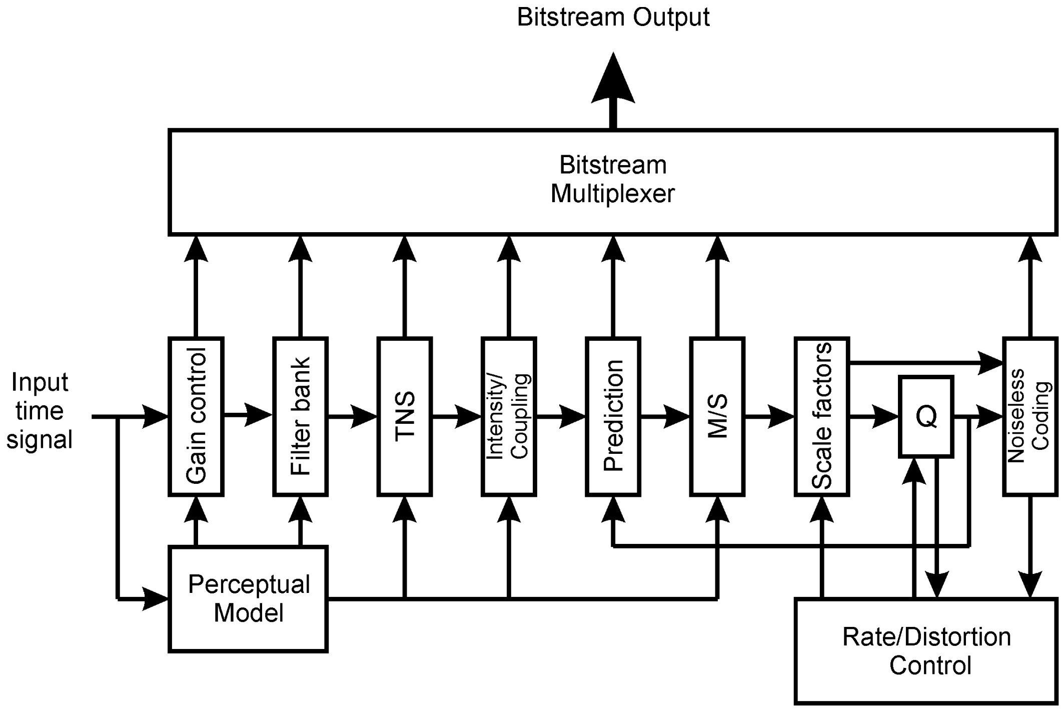
MPEG-1, Layer2 Coder



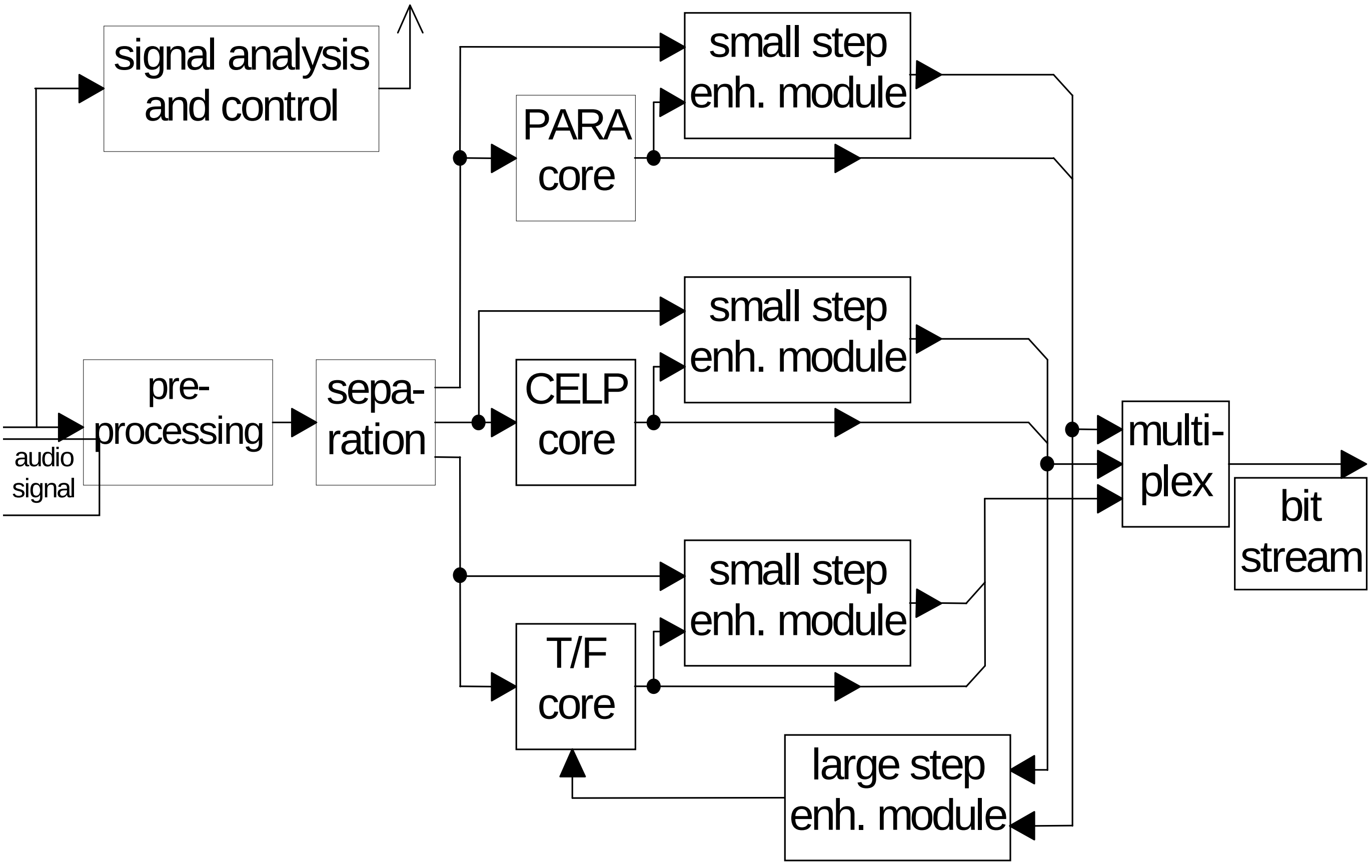
MPEG-1, Layer3 Encoder



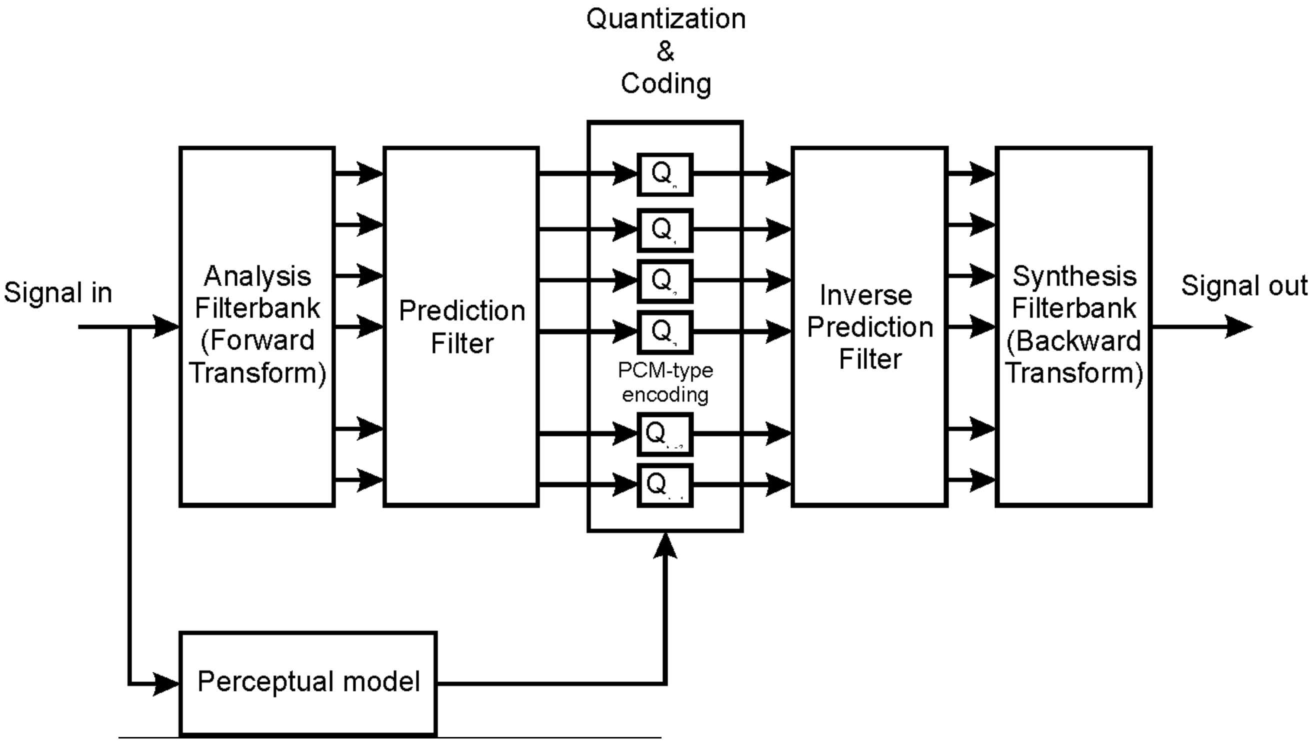
MPEG-2 AAC Encoder



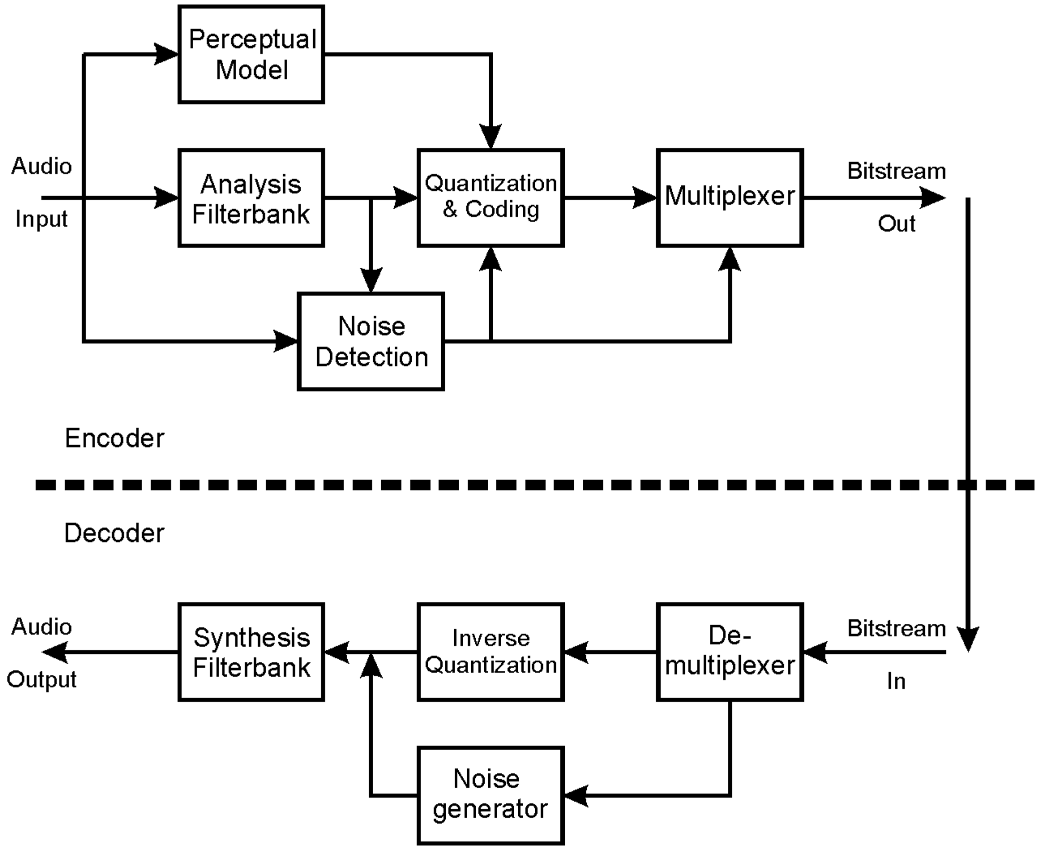
MPEG-4 Scalable Encoder



MPEG-2/4 AAC,Temporal Noise Shaping



MPEG-4, Perceptual Noise Substitution


Perceptual
Model
Audio
Input
Analysis
Filterbank
Quantization
& Coding
Multiplexer
Bitstream
Out
Noise
Detection
Encoder
Decoder
Audio
Output
Synthesis
Filterbank
Inverse
Quantization
De-
multiplexer
Bitstream
In
Noise
generator