Digital Audio Compression Algorithms

Markus Erne

Signal- and Information Processing Laboratory
Swiss Federal Institute of Technology ETHZ
erne@isi.ee.ethz.ch, http://www.isi.ee.ethz.ch

Abstract

Digital audio devices such as CD-players or DAT-recorders have become a standard 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 simultaneously the conflicting demands of high compression ratios and transparent reproduction quality. This tutorial will present the principles of lossless and lossy audio compression schemes whereas we will strongly focus on perceptual compression algorithms. Starting from the definition of entropy, lossless audio coding will be evaluated before perceptual coding will be addressed. Psychoacoustic principles, different quantizing and bit-allocation schemes, subband coding and transform coding will be presented in detail in order to introduce different perceptual coding algorithms. An audio demonstration of psychoacoustic effects such as frequency-domain masking and temporal masking will help the audience to understand the basic psychoacoustic principles. A short overview of the development of perceptual coding algorithms, including the different MPEG-standards and the related concept of different layers 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 (T/F-coders), CELP-coders and parametric coders each of which will be briefly presented before an audio demonstration of the different coding schemes at various bitrates will conclude the presentation.

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 signal reproduction. Conventional digital audio devices (CD, DAT) are typically sampled at 44.1 or 48 KHz, using linear PCM and a quantization of 16 Bits per sample. Therefore data-rates in the range of 1,5 Mbit/s will 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 prefered in order to allow reconstruction at the decoder bit by bit. 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 taking 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
- Polyphase filterbank with uniform bandpass filters
- Time-varying, critically sampled bank of non-uniform bandpass filters
- Hybrid transform/filterbank scheme
- Harmonic/sinusoidal signal analizer
- 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 exhibiting various coding requirements due to their completely different

temporal and spectral fine-structure, suggests 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:

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

Any information can be subdivided into 4 categories:

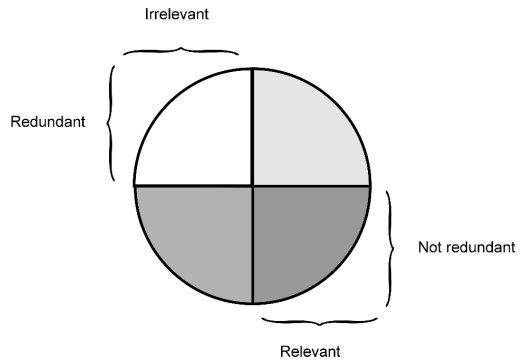


Fig 1: Partition into relevant and irrelevant parts and into redundant and non-redundant parts.

According to these categories, a digital compression scheme will remove either the amount of irrelevant or redundant information in a signal or both. A lossless coding scheme will remove redundant parts in a signal and therefore will come as close as possible to the Shannon-bound of entropy. In practice however, compression ratios between 2 and 3,5 can be achieved and it should be clear that 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 will minimize the variance of the difference signal between successive samples and the entropy coder (Huffmann, LZW) will allocate short codewords to samples with highest probability and longer codewords to samples with lower probability.

Lossless coding schemes allow to reconstruct the coded 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, in multichannel storage or transmission of audio signals and in 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 lossless compressed audio bitstreams some additional error protection is required because error concealment on decorrelated samples will become 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 additionally the perceptual irrelevancies of the signal.

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. The quiet threshold is well approximated by the non-linear function:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$

The absolute threshold of hearing is an important parameter in audio coding algorithms in order to scale the tolerable injected quantization noise accordingly.

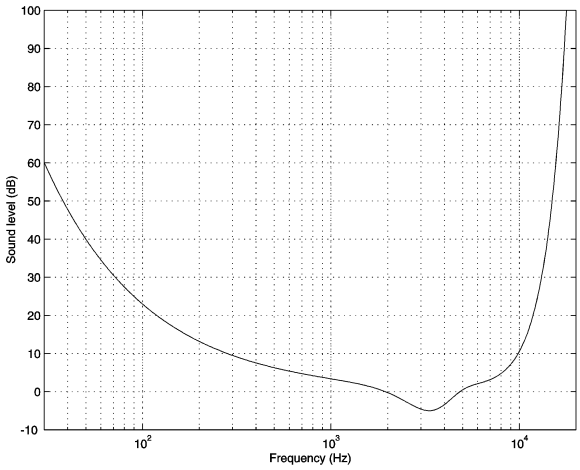


Figure 2: Absolute Threshold of hearing

In the cochlea, a frequency-to place transformation takes place which leads us to the notation of critical bands [4]. Distinct regions in the cochlea are tuned to different frequency bands. Critical bandwidth can be considered as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise at constant sound pressure level remains constant within the

critical band and then, outside the critical band begins to increase [5]. There are approximately 25 critical bands and each critical band is attributed to a Bark-number.
It turns out that the critical band at low frequencies show a constant bandwidth of around 100 Hz up to 500 Hz and at higher frequencies they 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 Bands

Most 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 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.

3.1.1 Frequency domain masking

Frequency domain masking can be observed within critical bands (simultaneous masking) or across critical bands (inter band masking).
In the presence of a masker, the absolute threshold of hearing will be modified to the masking threshold. All signals which are below the masking threshold will not be perceived by the human auditory system and therefore quantization noise in every subband can be as high as the masking threshold permits due to the individual quantizer in each subband.
Some algorithms use a subband decomposition equal to the critical bands in order take advantage of frequency domain masking.

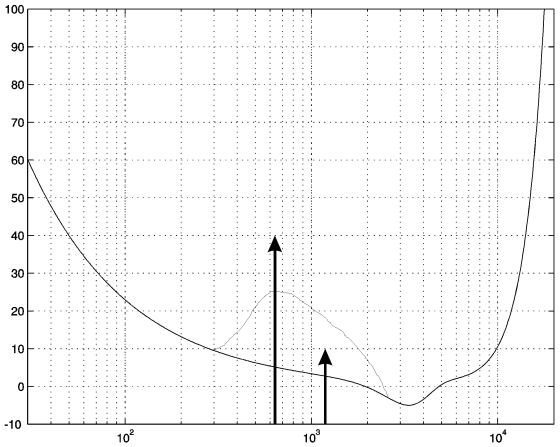


Figure 4: Masking effect, highlighted by the modified masking threshold in the presence of a masker and a maskee

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 from the masker to the minimum masking threshold in this particular critical band and the noise to mask ratio (NMR) denotes the log distance from the minimum masking threshold to the quantization noise level.

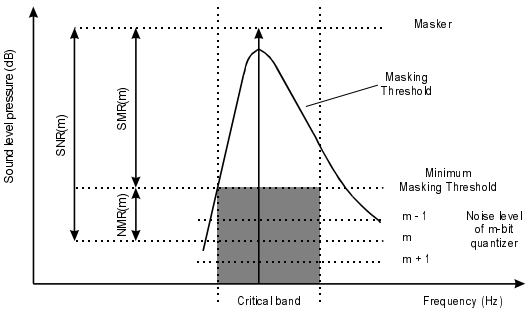


Figure 5: Frequency domain masking showing the signal to mask ratio and the noise to mask ratio

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 beneath the audibility threshold in the pre- and post-masking regions. 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.

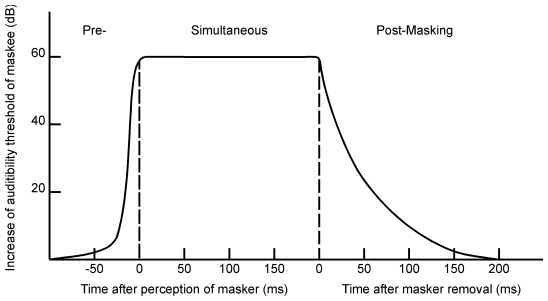


Figure 6: Temporal masking

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 transform spreads the quantization noise throughout the reconstructed block resulting in an unmasked distortion preceding the signal attack.

4 Subband Decomposition

The filterbank is certainly one of the most important part of a perceptual coding scheme. Despite the fact that there exist audio coding schemes using signal transforms (MDCT, Lapped Transform, Wavelet Transform) and audio coding devices using filterbanks (Polyphase, Time varying, QMF), they are mathematically almost equal. Transform coders process the signals in blocks of samples whereas filterbanks based on convolutions may operate on a sample by sample basis. There are several options for the implementation of a filterbank which directly will affect the computational complexity:

Uniform or Non-uniform filterbanks: uniform filterbanks are rather easy to implement and they are widely used in the ISO-MPEG-standard for Layer1 and Layer2. Non-uniform filterbank oftenly mimic the response of the human auditory system and they therefore better approximate the critical bands. Nevertheless, they may not be superior in terms of coding gain compared to a uniform filterbank.

Static or Time-varying filterbank: Ideally the filterbank should adapt to the signal-statistics in order to optimize the time-frequency tiling. Only a few 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.

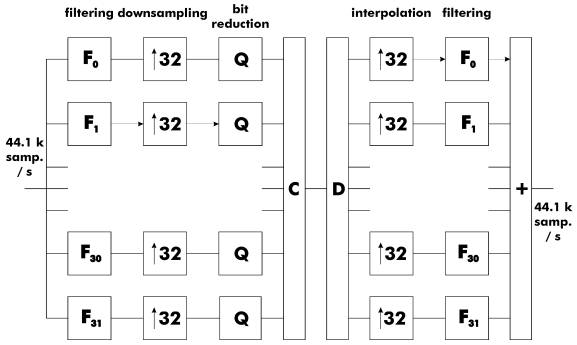


Figure 7: Audio Coder using a uniform polyphase filterbank

5 Block diagram

Combining the individual building blocks, we can now derive a block diagram of a basic audio coder.

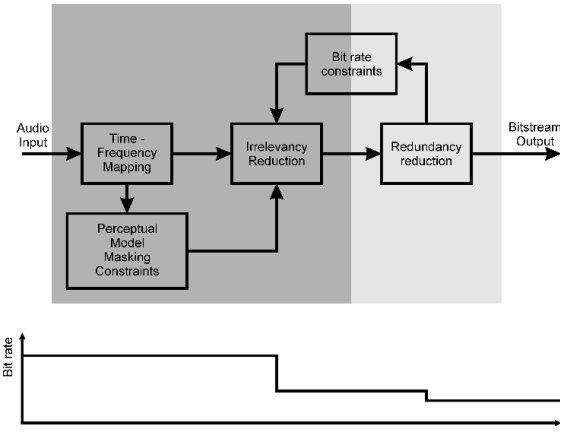


Figure 8: Block diagram of a basic perceptual audio coder

The time-frequency mapping or filterbank decomposes the audio signal into subbands and therefore decorrelates successive samples. In the perceptual model, the masking threshold is computed based on the presence of frequency domain- and temporal masking present in the current block of audio data. The reduction in irrelevance is perforemd in the quantizer. The quantizer can be a uniform quantizer, a non-uniform quantizer or an adaptive quantizer. The quantizer is controlled by the perceptual model in order to guarantee that the quantization noise in each subband is below the masking threshold. After the quantizer, the quantized subband samples can be grouped in a most 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) based on a combination of QMF-filters and MDCT is used in Minidisc-recorders. AC-2 and AC-3, developed by Dolby 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 MPEG-1 can be implemented using 3 different layers each having different target-bitrates, different coding delays and different complexity.
Layer1, 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.
Layer2, a derivative of the MUSICAM-algorithm has a target bitrate of 128 Kbit/s per channel and is widely used in broadcast applications and DAB (Digital Audio Broadcasting).
Layer3, the most complex algorithm will allow a transparent quality at bitrates down to 64 Kbit/s or even less per channel. Layer3 therefore can be used for low bitrate applications (satellite transmission, ISDN-applications, solid state memory recorders etc.).

MPEG-2 Audio is an extension of MPEG-1 for the MPEG-2 BC (Backwards Compatible) standard. The first extension allows to use lower sampling frequencies (16 KHz, 22 KHz and 24 KHz) for low bitrate applications. Additionally, multichannel coding for surround-sound applications (5.1. configuration) has been added in the MPEG-2 BC standard.
MPEG-2 NBC (Non Backward Compatible) makes use of AAC (Advanced Audio Coding), a new coding scheme, offering some advanced features which can not 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 transport up to 48 channels of compressed audio.

6.3 MPEG-4

The upcoming MPEG-4 standard is still under development and will 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 it offers a complete framework for synthesis, rendering, transport, compression and the integration of audio (speech, music etc.) and visual data. 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 is used. There exist 3 different categories of audio coders in MPEG-4: A parametric core for very low bitrate applications, a CELP-core for speech applications and a T/F (Time/Frequency) core for high quality audio applications. For the T/F-part, AAC with some modifications is proposed in the draft-standard. MPEG-4 AAC [8] includes new tools such as PNS (Perceptual Noise Substitution) which allows to save transmission bandwidth for noise-like signals. Instead of coding the noise-like signals, 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 scalability. Part of the bitstream can be 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 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 advancement-layers.

7 Conclusions

A lot of effort has been put in the development of high quality and/or low-bitrate audio compression algorithms during the past few years. Audio coding still is a young field of research but 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 on the psychoacoustic principles for advanced psychoacoustic models and on the combination of coding and synthesis methods.

REFERENCES

[1] Brandenburg K., Stoll G., "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

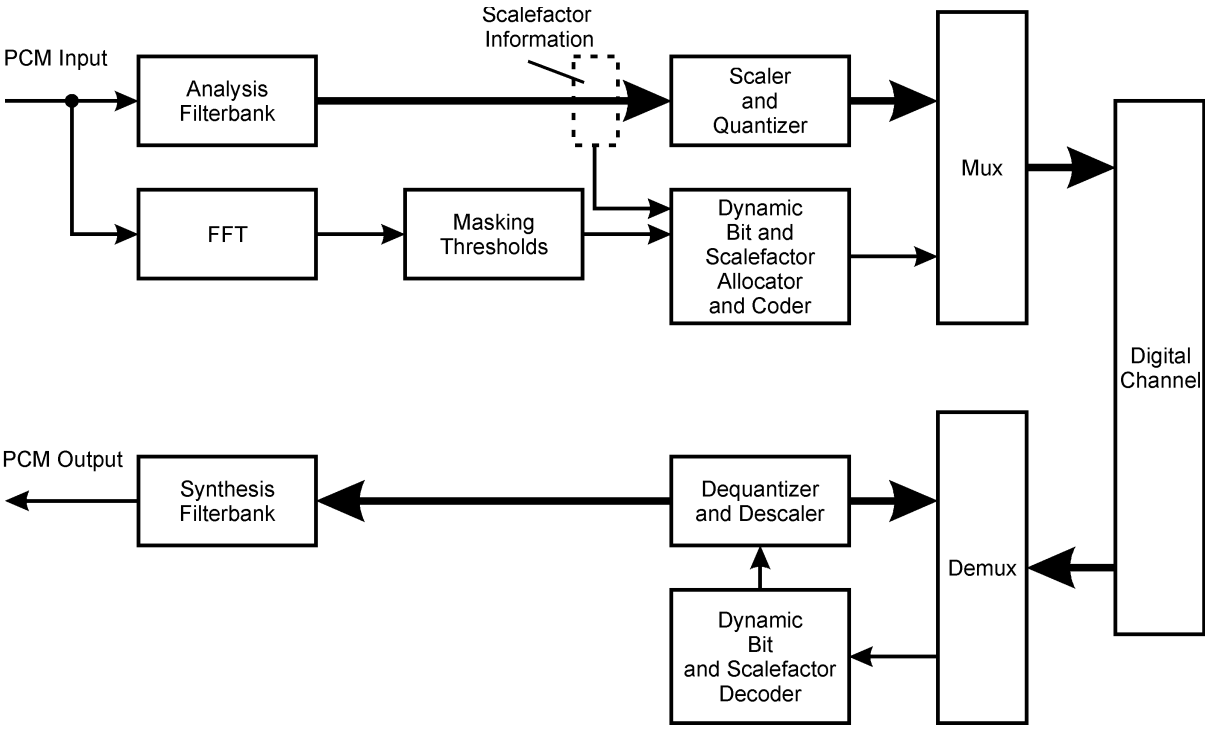
[6] Gilchrist N., Grewin C., "Collected Papers on Digital Audio Bit-rate Reduction",. *Audio Engineering Society*, September 1996

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

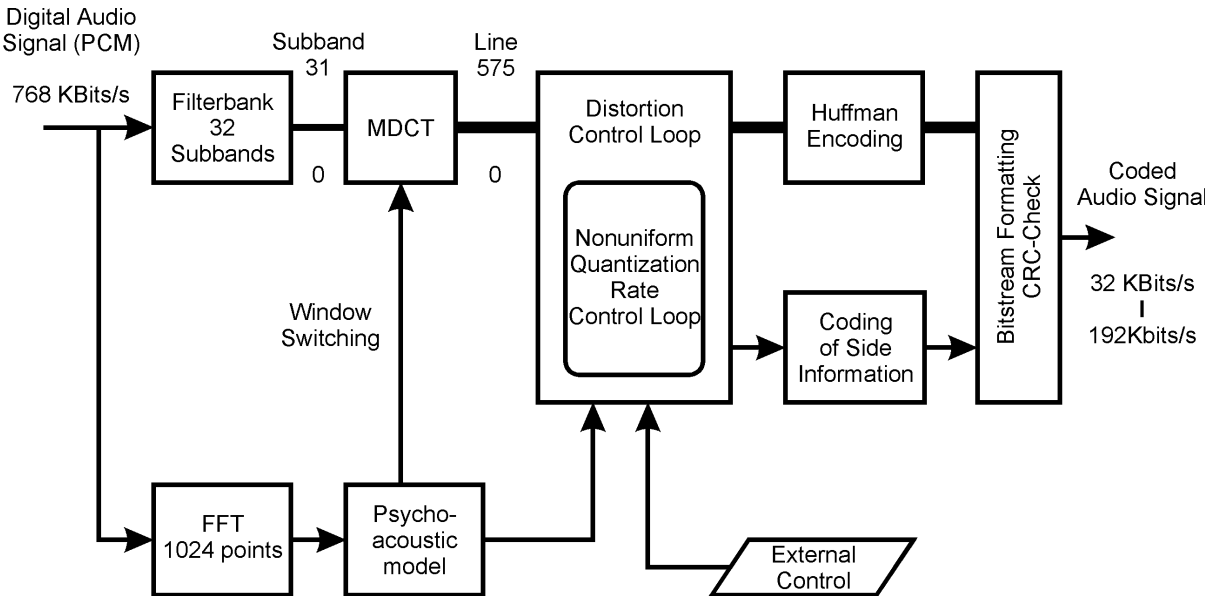
[8] Kahrs M., Brandenburg K.H. "Applications of Digital Signal Processing to Audio and Acoustics" . *Kluwer Academic Press,* 1998

APPENDIX

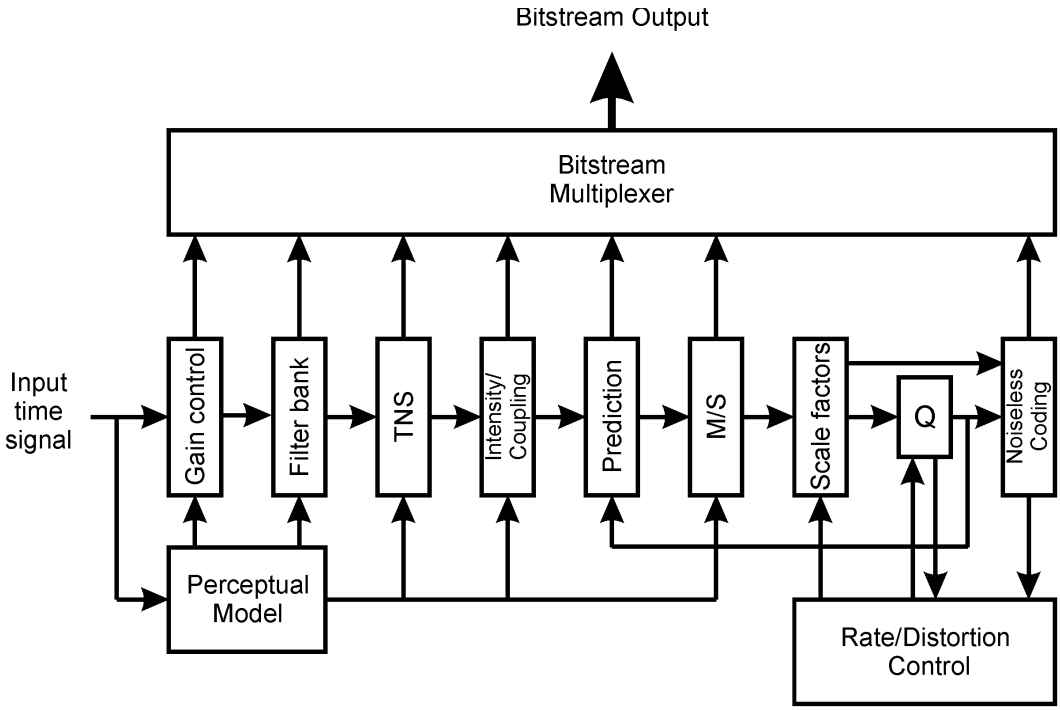
MPEG-1, Layer2 Coder



MPEG-1, Layer3 Encoder



MPEG-2 NBC, AAC Encoder



MPEG-4 scalable Encoder