State-of-the-Art of Audio Perceptual Compression Systems

Estado del Arte de los sistemas de Compresión de Audio Perceptual

Marcelo Herrera Martínez, Ana María Guzmán Palacios

Abstract

In this paper, audio perceptual compression systems are described, giving special attention to the former one: The MPEG 1, Layer III, in short, the MP3. Other compression technologies are described, especially the technologies evaluated in the present work: OGG Vorbis, WMA (Windows Media Audio) from Microsoft Corporation and AAC (Audio Coding Technologies).

Keywords: Audio Signal, Compression, Frecuency mapping, MP3, MPEG3, Quantization, SBR, WMA.

Resumen

En este trabajo se describen los sistemas de compresión de audio perceptual, con especial atención al último modelo disponible: El MPEG 1 Layer III, en definitiva, el MP3. Así mismo, se describen y evalúan otras tecnologías de compresión, especialmente las siguientes tecnologías: OGG Vorbis, WMA (Windows Media Audio) de Microsoft Corporation y AAC (Audio Coding Technologies).

Palabras clave: señal de audio, compresión, mapeo Frecuencia, MP3, MPEG3, cuantificación, SBR, WMA

Recibido: Abril 30 de 2013 Aprobado: Noviembre 19 de 2013
Tipo de artículo: Artículo de revisión de tema, investigación no terminada.
Afiliación Institucional de los autores: Grupo de Acústica Aplicada – Semillero de investigación Compresión de Audio.
Facultad de Ingeniería. Universidad de San Buenaventura sede Bogotá.
Los autores declaran que no tienen conflicto de interés.

Introduction

The storage and reproduction of sound has evolved since the times of Thomas Alva Edison's phonograph cylinders. The development of the vinyl, as a mechanical storage device, and the cassette, based on magnetization principles were widely distributed and used by the audiophile community. Nevertheless, it was in the 80's, with the incursion of the CD-Audio by Philips, when reducing sizes for storage devices began. In a parallel way, radio-broadcasting (AM-FM) began to visualize the incoming market of internet solutions in order to save bandwidth costs. The development of compression technologies, as MPEG (Fraunhofer Institute), Ogg Vorbis and WMA (Windows Media Audio) among others, initiated. These technologies are based on discarding irrelevancies and redundancies of the audio signal. The irrelevancies of the audio signal can be achieved using psychoacoustic features of the HHS as the masking phenomenon, which was extensively described in the previous chapter. The general configuration of an audio compression system is shown in the appendix. 1.

The PCM signal, a digital representation of the audio signal enters the compression system. A time-to-frequency mapping is performed initially. These values are compared to masking thresholds, obtained by a psychoacoustic model which has implementations based on the previous chapters. In these models, the codec compares at each instance the value of the incoming signal, computes the masking threshold associated with the given signal component, and the components which remain below the computed masking threshold are discarded. This is understood as irrelevance. The non-discarded components of the signal are fed into a re-quantization block, where each of the Fourier or DCT coefficients is re-quantized. The re-quantized coefficients are fed into a Huffman coding block where entropic coding is performed, and more information is discarded based on redundancies, as it is explained in the section I-G.

Compression schemes have reached compression ratios till 1:12, and further development is near with the inclusion of other frequency-time transform algorithms, and quantization schemes. These compression schemes introduces to the musical audio signal undesirable artifacts. Their evaluation is the main part of this research.

These high compression ratios are achieved due to the possibility of discarding components from the signal due to irrelevancies (masking) and redundancies (entropic coding).

MPEG-1

MPEG-1 was the first audio-compression technology to appear at the Fraunhofer Institute, in Erlangen (Germany) in 1987 within the frame of the EUREKA's project (DAB broadcasting system). The audio standard which became available is part of the standard MPEG 1. MPEG-2, and further releases were supposed to improve this first algorithm. The basic scheme of MPEG-1 Layer III is depicted in the appendix. 2.

Table I. *Bit-rates used in MPEG-1 layer III*

Quality	Bandwidth [kHz]	Mode	Bit rate [kbit/s]	Compression Rate
Telephone	2.5	mono	8	1:96
Better than AM	7.5	mono	32	1:24
FM radio	11	stereo	56..64	1:24..26
Near CD	15	stereo	96	1:16
CD	>15	stereo	112..128	1:12..14

MPEG-1 became a standard in 1992, and it is also known as ISO/IEC 11172. The audio signal is then represented by its spectral components on a frame-by-frame basis and encoded exploiting perceptual models. MPEG-1 Audio standard specifications were derived from two main proposals: MUSICAM presented by CCETT, IRT Phillips, and ASPEC presented by AT&T, FhG, and Telefunken [9].

The Stereo Audio Signal is encoded using similarities of right and left channels.

MP3 has implemented three modes for this purpose:

1. Stereo
2. Joint Stereo (Middle/Side): In this mode, one signal contains the sum of channel components and a second signal the difference.
3. Joint Stereo (Intensity Stereo): In this mode, high frequency components are coded in mono and the information of panning (direction) is also coded.

MP3 has 3 layers (back compatible), which are physical and mathematical models. The complexity is proportional to the layer number.

The bit-rates used in MPEG-1 Layer III are summarized in the Tab. 1.

The structure of the MP3 file is shown in the appendix. 3.

Layer I and II

The time to frequency mapping is performed by applying a 32-band PQMF to the audio data. Quantization is performed by a mid-tread quantizer. Psychoacoustic Model 1 is applied, with a 512-point FFT (Layer I) or 1024-point FFT (Layer II).

The MPEG-1 Layer I-II structure is depicted in the Appendix 4. Basically, it corresponds to the general configuration of an audio codec, shown in Appendix 5. After splitting the signal into 32 sub-band components, the quantizer recalculates the values of the signal taking into account the psychoacoustic thresholds stored inside the psychoacoustic model.

Layer III

For the Layer III, the output of the PQMF is fed to a MDCT stage. In Layer III, the filter bank is signal adaptive in difference to previous layers.

The block diagram of Layer III filter bank analysis stage is shown in Appendix 5. After the 32-band PQMF filter, blocks of 36 sub-band samples are overlapped by 50 percent, multiplied by a sine window and then processed by the MDCT transform.

The filter bank analysis in MPEG-1 Layer III is shown in Appendix 6.

Compression systems encounter problems while tracking signals containing transients, such as the castanet excerpts. Quantization errors spread over a block of 1152 time- samples. This leads to unmasked temporal noise, and the audibility of the artifact known as pre-echo.

The layer III implemented a shorter block size for avoiding pre-echo. When a transient is detected, the 384 sample window is used; in the opposite case, a 1152 sample window is chosen. This reduces the temporal spreading of quantization noise for sharp attacks. The definition of those windows is treated in [9].

Psychoacoustic Models in MPEG-1

The scheme for the psychoacoustic model in MPEG-1 is shown in the Appendix 7.

Summarizing, the psychoacoustic model 1 performs the SPL computation in each band (the signal level for each spectral line k), the separation of tonal and non-tonal components and the other processes outlined in the Appendix. 14. More information can be found in [9].

Psychoacoustic Model 2

Model 2 process and calculations differ from those of Model 1. The block diagram is depicted in Appendix. 8.

This model uses another mathematical expression for the basic spreading function B(dz) measured in dB. The tonality index in each partition is calculated based on the predictability of the signal from the frequency lines in prior frames. When model 2 is applied to Layer III, it is altered to take into account the different nature of the Layer III (hybrid filter bank, and the switching block). Attacks, for example, are detected based on a psychoacoustic entropy calculation. [9]. The specification of the audio-syntax, the scale factors, bit allocation, quantization, Huffman coding and bit-reservoir are extensively explained in [9].

The Software applications which support MP3 format are Cool Edit Pro (Syntrillium), Sound Forge, Musicmatch Jukebox and Easy MP3, among others.

MP3 Format, Lame Codec

The Lame Codec was created in 1998. It has its own psychoacoustical model (GPSYCHO) which enables the graphic imaging of MDCT coefficients. Some of the softwares which support this codec are Audiograbber, WaveLab, WinLame and RazorLame, among others.

MP3 Pro format

This format uses the MPEG-1 Layer III specification. It was released to the market by Thomson. The available softwares which support the format are Thomson MP3Pro Audio Player, MusicMatch Jukebox and Nero 5.5, among others. It uses the SBR Technology (Spectral Band Replication).

Spectral Band Replication is a technique by which the non-transmitted higher frequency range of a compressed audio file is deduced by some helper bits and the transmitted base band. The bandwidth of the audio signal is limited in or prior to the coding process.

It is based on the dependencies between lower and higher frequency components of an audio musical signal, as Appendix. 10 shows. The localization of the SBR module inside the MP3 Pro format is shown in Appendix 10.

MPEG-2

This standard was developed to extend the MPEG-1 Audio functionality to lower data rate and for extending to the multi-channel applications.

MPEG-2 LSF, "MPEG-2.5" and MP3

MPEG-2 LSF has lower sampling data rates, 24, 22.05 and 16 kHz. Therefore, the other parts of the chain are also adapted to this fact, as the psychoacoustic model. Even a lower sampling rate modification in "MPEG-2.5" was created. Sample rates of 12, 11.025 and

8 kHz were allowed. The MPEG-2 Multi-channel BC Configuration, Multilingual Channels is extensively described in [9].

MPEG-2 AAC

Started in 1994, the MPEG-2 Audio committee wanted to define a higher quality multi-channel standard than MPEG-1 backwards compatibility. The MPEG-2 non-backwards compatible audio standard ISO/IEC 13818-7, renamed MPEG-2 Advanced Audio Coding (MPEG-2 AAC) was finalized in 1997 [9]. The structure is depicte in the Appendix 18.

The AAC system offers three profiles: Main Profile, Low Complexity (LC) Profile, and Scalable Sampling Rate (ScSR) Profile. AAC uses three advanced coding techniques, as SBR (Spectral Band Replication), TNS (Temporal Noise Shaping) and PNS (Perceptual Noise Substitution).

Perceptual Noise Substitution can be summarized as the junction between the Perceptual coder and the parametric representation of noise-like signals.

The Noise-like signal components are detected on a scalefactor band basis. The corresponding groups of spectral coefficients are excluded from the quantization/coding process. Instead, only a noise substitution flag plus total power of the substituted band is transmitted in the bitstream. The Decoder inserts pseudorandom vectors with desired target power as spectral coefficients.

LTP: Long Term Prediction is applied for the prediction of tonal-like/noise-like signals. Since tonal-like signals require much higher coding precision than noise-like, higher bit rate is necessary. Nevertheless these components are predictable. The Prediction of each spectral coefficient is done by a backward adaptive second order lattice predictor.

Transform-Domain Weighted Interleave VQ (Vector Quantization) After the spectral coefficients are normalized, the interleaving of spectral coefficients are fed into new sub-vectors. These sub-vectors are quantized with vector quantization [9].

OGG VORBIS

Ogg Vorbis is a license-free audio compression format which was originated as an alternative to the commercial audio codecs as MP3 and WMA. It uses the DCT transform for the time-to-frequency mapping, and vector quantization in the quantization stage. The complete OGG Vorbis structure is depicted in Appendix 12. As it can be seen in the figure, OGG Vorbis uses MDCT (Modified Discrete Cosine Transform) instead of FFT, and Vector Quantization.

Ogg Vorbis allows maximum encoder flexibility, allowing excellent performance over a wide range of bit-rates. At the high quality/bit-rate end of the scale (CD or DAT rate stereo, 16/24 bits) it performs similarly to MPEG-2 and to MPC. The 1.0 encoder can encode high-quality CD and DAT rate stereo below 48 kbps without re-sampling to a lower rate. Vorbis also supports lower and higher sample rates (from 8 kHz telephony to 192 kHz digital masters) and a range of channel representations (monaural, polyphonic, stereo, quadraphonic, 5.1 and ambisonic, up to 255 discrete channels). The codec is structured to allow the addition of a hybrid wavelet filterbank in Vorbis II to offer better transient response and reproduction using a transform better suited to localize time events. [29]

Vorbis decoding is computationally simpler than mp3, but it requires more working memory as Vorbis has no static probability model; the vector codebooks used in the first stage of decoding from the bit-stream are packed into the Vorbis bit-stream headers. Vorbis is a method of accepting input audio, dividing it into individual frames and compressing these frames into raw, unformatted packets. A brief description of some terms belonging to the Vorbis specification will be held next, since this codec was found to have the best performance during subjective evaluations.

The Decoder Configuration

The decoder setup consists of the configuration of multiple, self-contained component abstractions that perform specific functions in the decode pipeline, as it is depicted in the Appendix 13.

The Global Configuration Mode consists of some audio related fields (sample rate, channels).

The mode mechanism is used to encode a frame according to one of multiple possible methods with the intention of choosing a method best suited to that frame.

The mapping contains a channel coupling description and a list of "sub-maps" that bundle sets of channel vectors together for grouped encoding and decoding. A sub-map is a configuration/grouping that applies to a subset of floor and residue vectors within a mapping.

Vorbis encodes a spectral floor vector for each PCM Channel. This vector is a low-resolution representation of the audio spectrum for the given channel in the current frame. There are two types of floors. Floor 0 uses a packed LSP representation on a dB amplitude scale and Bark frequency scale. Floor 1 represents the curve as a piecewise linear interpolated representation on a dB amplitude scale and linear frequency scale.

The spectral residue is the fine structure of the audio spectrum once the floor curve has been subtracted out.

Codebooks perform entropic decoding. It is provided by the Huffman coding.

More specifications about the coding/decoding procedure of Vorbis, as well as the Bit-packing, the Probability Model, the Codebooks, the Codec Setup, the Packet Decode, the Comment Field, the Header Specification, the Floor specifications, the Residue setup and the Helper Equations can be found in the ogg vorbis specification [29].

Some of the softwares which support this format are Cdex, Siren Jukebox, GoldWave, WinAmp, and SoundForge.

Windows Media Audio

Windows Media Audio is the commercial codec of Microsoft Company. Its contemporary version, version 9, is available in Windows 2000 and XP. WMA is not just audio, but a complex multimedia format. The programs which also support WMA are Nero 5.5 or WMA Encoder/ Decoder from Media Twins.

Windows Media also includes tools for video compression (Windows Media Video) and tools for DRM (Digital Rights Management). The actual version of WMA 9 is conformed by four coding algorithms:

WMA 9 – It is a basic lossy codec. It is used for the compression of two channel sound signals, with CD parameters (16 bit/44,1 kHz).

WMA 9 Professional – It is used for the coding of more than two channels. It is able to compress even signals of high quality (24 bit/96 kHz).

WMA 9 Lossless.

WMA 9 Voice.

Other formats and implementations

There exists a wide variety of audio-coding formats available at the internet, or inside some audio packages. Among them, Dolby AC-3, MPEG-4 and ATRAC can be mentioned.

The state-of-the-art audio compression systems include implementations of suitable transforms for the pre-echo detection with the use of energy calculation in neighboring blocks, birdies cancellation, the use of sinusoidal lapped transforms, etc. Hybrid implementations with more than one transform are also alternatives. Vector quantization as in Ogg Vorbis has been also lately preferred.

Conclusions

A complete state-of-the-art of the available perceptual audio coders has been made. The work comprises the former audio coders, as MPEG 1- Layer3, describing all the technical aspects and the mathematical apparatus that lie behind them. At the end some free-source implementations are reviewed, and information probabilistic aspects are described in full detail in order to completely characterize them.

References

[1] GUILFORD, J.P., Psychometric methods, McGraw-Hill, Second Edition, 1954.

[2] ZWICKER, E., FASTL, H. Psychoacoustics: Facts and Models, Springer, 1990.

[3] STEPANEK, J., "Interpretation and comparison of perceptual spaces" , in Inter-Noise 2005, Rio de Janeiro, 2005, CD-ROM, file:/papers/doc_1627.pdf

[4] BECH, S., ZACHAROV, N. Perceptual Audio Evaluation. Theory, Method and Application. Wiley, 2006

[5] CRONBACH, L.J. Prospects for a psychometric theory based on utility measure.1954.

[6] HINTON R. P., Statistics Explained, Routledge, New York 2004.

[7] BRANDENBURG, K: Scalable Audio Coding MPEG-4, Fraunhofer, Erlangen, Germany.

[8] KIRBY, D.G., WATANABE, K.: "Subjective testing of MPEG-2 NBC multichannel audio coding", Broadcasting Convention, 1997.

[9] DOLEJSI, P., Psychoacoustic assessment of audio compressors, Diploma Thesis, CVUT, January 2004.

[10] SVITEK, J., Methods for the subjective assessment of audio compressors, Diploma Thesis, CVUT, May 2005.

[11] KUDLACEK, F., Applications of psychometric methods to the evaluation of quality of compressed signals. Diploma Thesis, CVUT, January 2006

[12] HAVELKA, J.: Artefakty kompresoru zvukovych signalu. Diplomova prace, CVUT-FEL, Praha, 2007.

[13] SUK, P. Objektivni metody vyhodnocovani zvukove kvality kompresor zvuku. Diplomova prace CVUT, May 2004.

[14] KLIMA, M., BERNAS, M., HUSNIK, L. PATA, P. ROUBIK, K.: Qualitative Aspects of Image Compression Methods in Multimedia Systems. In Proceedings of Workshop 2005 - Part A,B. Prague: CTU, 2005, s. 188-189. ISBN 80-01-03201-9.

[15] FLIEGEL, K.: Image and Video Processing for Image Quality Evaluation Using Artificial Neural Network. In Workshop 2006 [CD-ROM]. Prague: CTU, 2006, vol. A, s. 150-151. ISBN 80-01-03439-9.

[16] HERRERA, M. Evaluation of compression artifacts. In Proceedings of the 10th International Student Conference on Electrical Engineering POSTER 2006. Prague (Czech Republic), 2006.

Los Autores

Marcelo Herrera Martínez

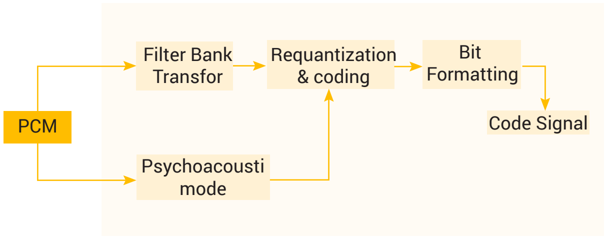
Ingeniero Electrónico, Magíster en Radioelectrónica y Doctor en Acústica de la Universidad Técnica de Praga. Profesor Titular de la Facultad de Ingeniería de la Universidad de San Buenaventura, Sede Bogotá. Líder del Semillero de Investigación de "Sistemas de Compresión Perceptual de Audio". Member (M) of IEEE in 2013.

Ana María Guzmán Palacios

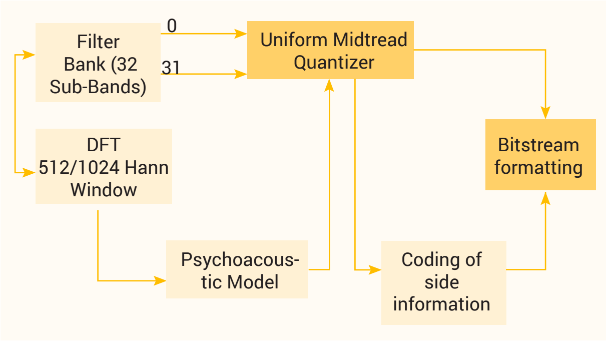
Nació en Popayán, Colombia. Obtuvo su título de Ingeniería Física en la Universidad del Cauca, Popayán. Posteriormente, consiguió su título de Maestría en Mecatrónica en la Universidad de Brasilia, Brasil. Actualmente, es profesora de tiempo completo en la Facultad de Ciencias Naturales, Exactas y de la Educación en la Universidad del Cauca, y trabaja con el Grupo de Investigación Sistemas de Comprensión perceptual de audio de la Universidad de San Buenaventura. Correo Electrónico: amguzmanp@gmail.com

Anexos

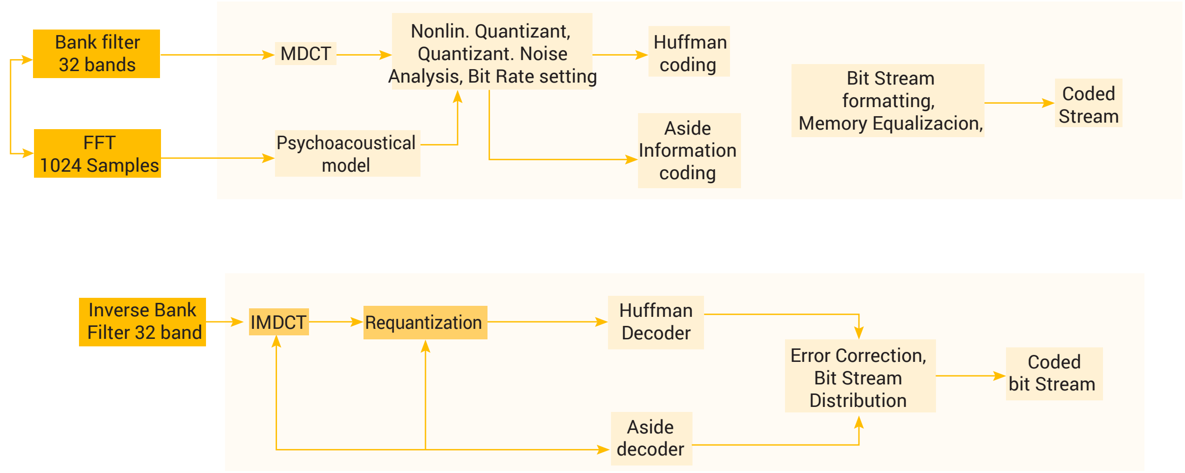
Appendix 1. *Audio Compression System*



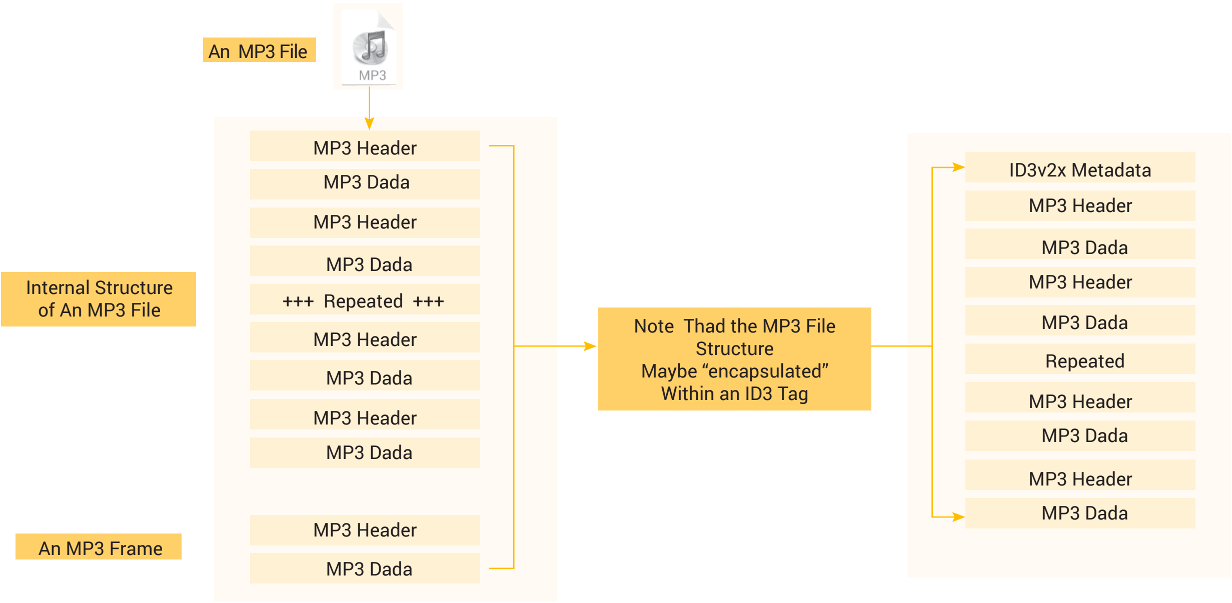
Appendix 4. *MPEG-1 Layer I-II*


Filter Bank (32 Sub-Bands)
0
31
Uniform Midtread Quantizer
Bitstream formatting
DFT 512/1024 Hann Window
Psychoacous-tic Model
Coding of side information

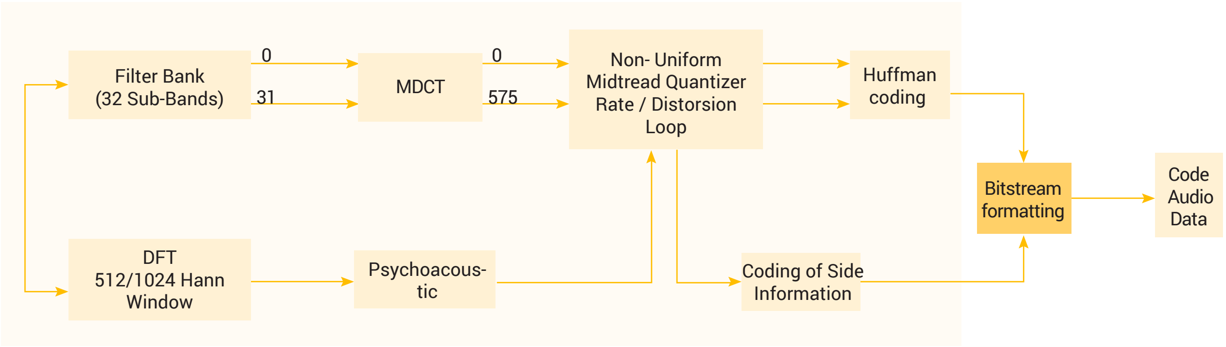
Appendix 2. *Scheme of MPEG-1 Layer III*


Bank filter 32 bands
MDCT
Nonlin. Quantizant, Quantizant. Noise Analysis, Bit Rate setting
Huffman coding
Bit Stream formatting, Memory Equalizacion,
Coded Stream
FFT 1024 Samples
Psychoacoustical model
Aside Information coding
Inverse Bank Filter 32 band
IMDCT
Requantization
Huffman Decoder
Error Correction, Bit Stream Distribution
Coded bit Stream
Aside decoder

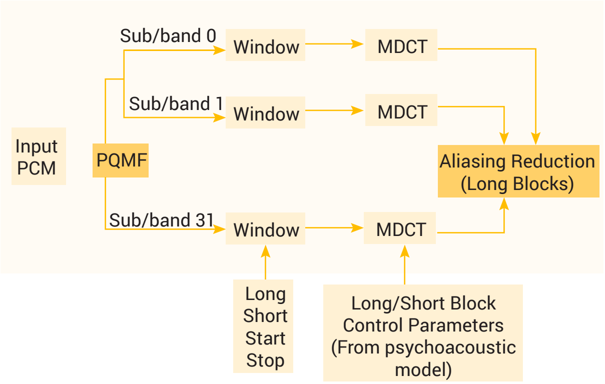
Appendix 3. *Structure of a MP3 file*



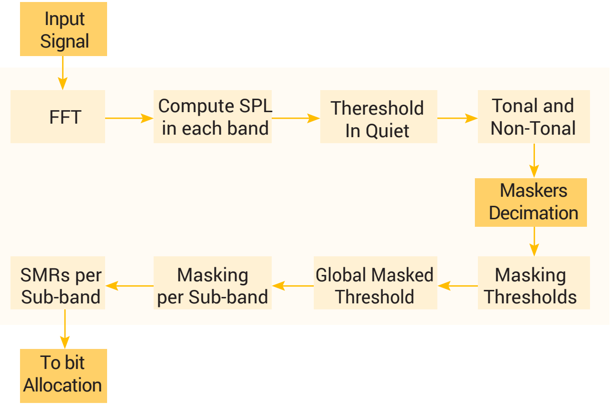
Appendix 5. *MPEG-1 Layer III*



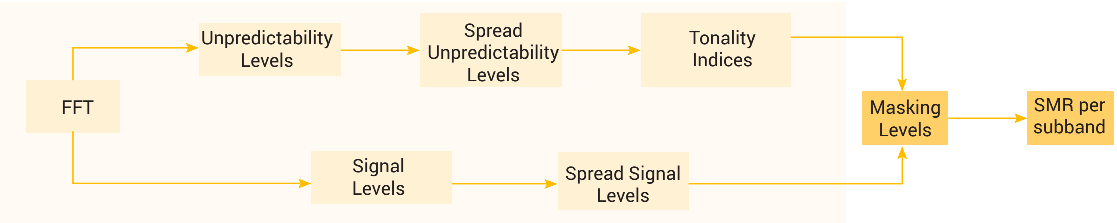
Appendix 6. *Filter Bank Analysis in MPEG-1 Layer III*



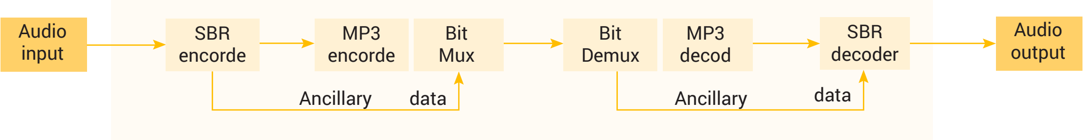
Appendix 7. *Psychoacoustic Model 1*

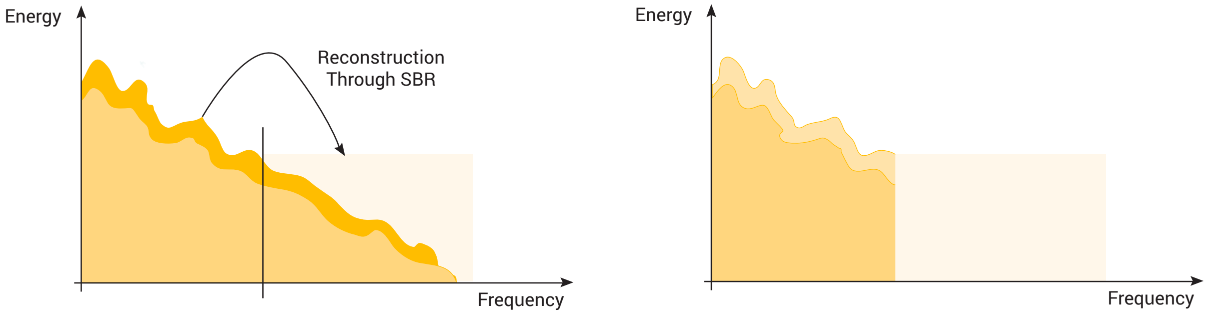


Appendix 8. *Psychoacoustic Model 2*

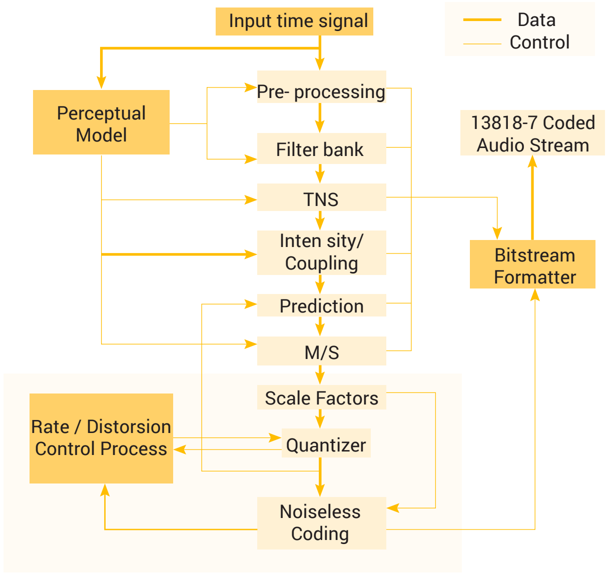


Appendix 9. *Localization of the SBR Module*

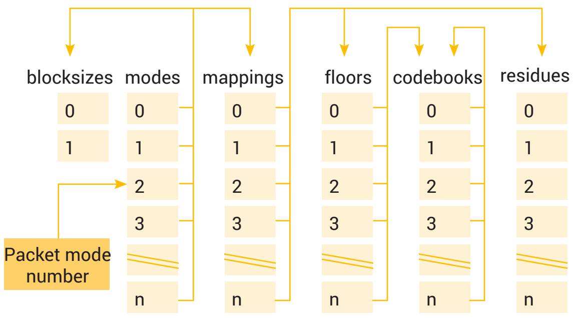




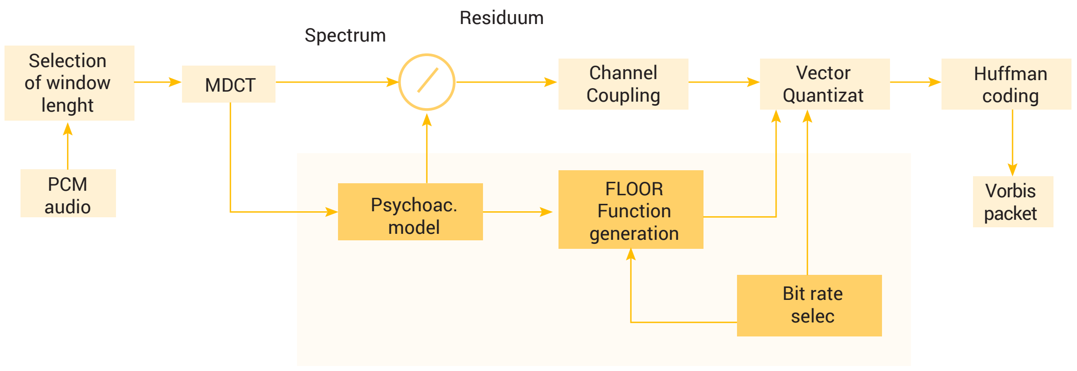
Appendix 10. *Reconstruction of the signal in a coder using SBR as MP3 Pro*



Appendix 11. *MPEG-2 AAC*



Appendix 13 OGG Vorbis Decoder



Appendix 12 OGG Vorbis Scheme