

INTRODUCTION TO DIGITAL AUDIO CODING AND STANDARDS

**THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE**

INTRODUCTION TO DIGITAL AUDIO CODING AND STANDARDS

Marina Bosi
Stanford University

Richard E. Goldberg
The Brattle Group



SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Bosi, Marina

Introduction to Digital Audio Coding and Standards /by Marina Bosi, Richard E. Goldberg.
p. cm. —(The Kluwer international series in engineering and computer science; SECS 721
Includes bibliographical references and index.

ISBN 978-1-4613-5022-4 ISBN 978-1-4615-0327-9 (eBook)

DOI 10.1007/978-1-4615-0327-9

1. Sound—Recording and reproducing—Digital techniques—Standards. 2. Data
compression (Computer science) I. Goldberg, Richard E. II. Title. III. Series.

TK7881.4 .B59 2002

621.389'32—dc21

2002040686

Copyright © 2003 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 2003

Softcover reprint of the hardcover 1st edition 2003

All rights reserved. No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without the written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper.

Chapter 11

MPEG-1 Audio

1. INTRODUCTION

After the introduction of digital video technologies and the CD format in the mid eighties, a flurry of applications that involved digital audio/video and multimedia technologies started to emerge. The need for interoperability, high-quality picture accompanied by CD-quality audio at lower data rates, and for a common file format led to the institution of a new standardization group within the joint technical committee on information technology (JTC 1) sponsored by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). This group, the Moving Picture Experts Group (MPEG), was established at the end of the eighties with the mandate to develop standards for coded representation of moving pictures, associated audio, and their combination [Chiarioglione 95].

MPEG-1 was the initial milestone achieved by this committee after over three years of concurrent work. MPEG-1 Audio represents the first international standard that specifies the digital format for high quality audio, where the aim is to reduce the data rate while maintaining CD-like quality. Other compression algorithms standardized prior to MPEG-1 addressed either speech-only applications or provided only medium-quality audio performance. The success of the MPEG standard enabled the adoption of compressed high-quality audio in a large range of applications from digital broadcasting to internet applications. Everyone is now familiar with the MP3 format (MPEG Layer III). The introduction of MPEG Audio technology radically changed the perspective of digital distribution of music,

touching diverse aspects of it, including copyright protection, business models, and ultimately our every-day life.

In this chapter and Chapters 12, 13, and 15, we discuss different audio coding algorithms standardized by MPEG. In this chapter, after presenting a brief history of the MPEG standards with emphasis on the MPEG Audio goals and objectives, we discuss in depth the layered approach and attributes of MPEG-1 Audio.

2. BRIEF HISTORY OF MPEG STANDARDS

The Moving Pictures Experts Group (MPEG) was established with the mandate to develop standards for coded representation of moving pictures, associated audio, and their combination. The original group of about 25 people met for the first time in 1988. Later MPEG become working group 11 of ISO/IEC JTC 1 sub-committee 29. Any official document of the MPEG group can be recognized by the ISO/IEC JTC 1/SC 29/WG 11 header. There were originally three work items approved for MPEG:

- The MPEG-1 standard [ISO/IEC 11172] coding of synchronized video and audio at a total data rate of about 1.5 Mb/s was finalized in 1992.
- The MPEG-2 standard [ISO/IEC 13818] coding synchronized video and audio at a total data rate of about 10 Mb/s was finalized in 1994.
- The third work item, MPEG-3, addressing coding of synchronized video and audio at a total data rate of about 40 Mb/s was dropped in July 1993, after being deemed redundant since its attributes were incorporated in the MPEG-2 specifications.

After the initial work started, a proposal for audiovisual coding at very low data rates with additional functionalities, such as scalability, 3-D, synthetic/natural hybrid coding, was first discussed in 1991 and then proposed in 1992 [ISO/IEC MPEG N271]. This phase of MPEG standardization was called MPEG-4 giving origin to the somewhat disconnected numbering of subsequent phases of MPEG. MPEG-4 was finalized in 1998 as [ISO/IEC 14496]. The MPEG-1, 2, 4 standards address video, audio compression as well as synchronization, compliance, and reference software issues. Although MPEG Audio is often utilized as a stand-alone standard, it is one component of a multi-part standard, where typically “part one” describes the system elements (i.e. synchronization of video and audio stream, etc.) of the standard, “part two” the video coding elements, and “part three” the audio coding elements. After MPEG-4 the

work of MPEG started focusing more and more towards coding-related technology rather than coding technology per se. MPEG-7, whose completion was reached in July 2001 [ISO/IEC 15938], addresses the description of multimedia content for multimedia database search. Currently in the developmental stage (only three parts of the standard have been approved), MPEG-21 is addressing the many elements needed to build an infrastructure for the usage of multimedia content, see for example [ISO/IEC MPEG N4318].

The goal of MPEG-1 Audio was originally to define the coded representation of high quality audio for storage media and a method for decoding high quality audio signals. Later the algorithms specified by MPEG were tested within the work of ITU-R for broadcasting applications and recommended for use in contribution, distribution, commentary, and emission channels [ITU-R BS.1115]. Common to all phases of MPEG was the standardization of the bitstream and decoder specifications only, but not of the encoder. A sample encoder algorithm is described in an “informative” part of the standard, but following the sample algorithm is not required to be compliant with the standard. This approach, while allowing for interoperability between implementation from different manufacturers, also allowed encoder manufactures to retain control on the core intellectual property and know-how that contributed to the success of the coding system. The input of the MPEG-1 audio encoder and the output of the decoder are compatible with existing PCM standards such as the CD and the digital audio tape, DAT, formats. MPEG-1 audio aimed to support one or two main channels, depending on the configuration (see more details on channel configuration in the next sections) and sampling frequencies of 32 kHz, 44.1 kHz, and 48 kHz.

In MPEG-2 Audio the initial goal was to define the multichannel extension to MPEG-1 audio (MPEG-2 BC, backwards compatible) and to define audio coding systems at lower sampling rates than MPEG-1, namely at 16 kHz, 22.5 kHz and 24 kHz. This phase of the work of the MPEG audio sub-group was partially motivated by the debut of multichannel de-facto standards in the cinema industry such as Dolby AC-3 (currently known also as Dolby Digital, see also Chapter 14) and the need for lower data rates for the emerging internet applications. After a call for proposals in late 1993, the work on a new aspect of multichannel audio, the so-called MPEG-2 non-backwards compatible, NBC (later renamed MPEG Advanced Audio Coding, AAC), was started in 1994. The objective was to define a higher-quality multichannel standard than achievable with MPEG-1 extensions. A number of studies highlighted the burden in terms of quality, or equivalently in terms of increased data rates demands, suffered by the design of a multichannel audio system when the backwards compatibility requirement

was enforced (see for example [Bosi, Todd and Holman 93 and ISO/IEC MPEG N1229] and see also next chapter for a detailed discussion on this issue). As a result of this phase of work, MPEG-2 AAC was standardized in 1997 [ISO/IEC 13818-7]. In a number of subjective tests MPEG-2 AAC shows comparable or better audio quality than MPEG-2 Layer II BC operating at twice the data rate, see for example [ISO/IEC MPEG N1420].

The MPEG-4 Audio goals were to provide a high coding efficiency, where the data rates introduced ranging from 200 b/s to 64 kb/s reach lower values than the data rates defined in MPEG-1 or 2. In addition to general audio coding technology MPEG-4 also accommodates:

- speech coding technology
- error protection
- content-based interactivity such as flexible access and manipulation, for example pitch/speed modifications;
- universal access, for example access to a subset of data or scalability
- support for synthetic audio and speech, such as in structured audio, SA, and text to speech, TTS, interfaces;
- additional effects such as post-processing (reverberation, 3D, etc.) and scene composition.

From its onset, the MPEG standardization process played a very relevant role in promoting technology across the boundaries of a single organization or country. As a result, teams around the world joined forces and expertise to design algorithms that incorporated the most advanced technology available given a certain range of applications. In the first phases of the MPEG work, the focus was centered on coding technologies. In this chapter and in Chapters 12, 13 and 15 a detailed description of the audio coding algorithms developed during the MPEG-1 through 4 phases are presented. In particular, the next sections of this chapter present the details of the MPEG-1 Audio algorithms.

3. MPEG-1 AUDIO

MPEG-1 is a compression standard that addresses the compression of synchronized video and audio at a total data rate of 1.5 Mb/s. It includes systems, video, and audio specifications. MPEG-1 became a standard in 1992, and is also known as [ISO/IEC 11172]. [ISO/IEC 11172-3] specifies the audio portion of the MPEG-1 standard. It includes the syntax of the audio coded bitstream and a description of the decoding process. In addition, reference software modules and a set of test vectors for assessing the compliance of the decoder are also provided by the standard specifications. The MPEG-1 audio encoder structure is not a mandatory part

of the standard specifications and its description is an informative annex to the standard. While the mandatory nature of the syntax and decoding process ensures interoperability, the encoder implementation is left to the designers of the system, leaving a large degree of differentiation within the boundaries of the standard specifications. The MPEG-1 standard describes a perceptual audio coding algorithm that is designed for general audio signals. There is no specific source model applied as, for example, in speech codecs. It is simply assumed that the statistics of the input signal are quasi-stationary. The audio signal is then represented by its spectral components on a frame-by-frame basis and encoded exploiting perceptual models. The aim of the algorithm is to provide a perceptually lossless coding scheme. The MPEG-1 Audio standard specifications were derived from two main proposals: MUSICAM [Dehéry, Stoll and Kerkhof 91] presented by CCETT, IRT and Philips, which is the basis for the low-complexity first two layers (see also next sections), and ASPEC (see [Brandenburg and Johnston 90] and [Brandenburg et al. 91]) presented by AT&T, FhG, and Telefunken which is the basis for layer III. The quality of the audio standard was tested by extensive subjective listening tests during its development. The resulting data, see for example [ISO/IEC MPEG 91/010], showed that, under strictly controlled listening conditions, experts listeners were not able to distinguish between coded and original sequences with statistical significance at typical codec data rates. Typical data rates for the coded sequences were 192 kb/s per channel for MPEG Layer I, 128 kb/s per channel for Layer II and Layer III (see detailed description of the different MPEG-1 Audio Layers later in this chapter and also the MPEG public documents at [MPEG]).

3.1 Main Features of MPEG-1 Audio

The sampling rates supported by MPEG-1 are 32, 44.1, and 48 kHz. The channel configurations encompass one or two channels. In addition to a monophonic mode for a single audio channel configuration, a dual monophonic mode for two independent channels is included. A stereo mode for stereophonic channels, which shares the available bit pool amongst the two channels but does not exploit any other spatial perceptual model, is also covered. Moreover, joint stereo modes that take advantage of correlation and irrelevancies between the stereo channels are described in the standard. The data rates vary between 32 and 224 kb/s per channel allowing for compression ratios ranging from 2.7 to 24:1 depending on the sampling rate. In addition to the pre-defined data rates, a free format mode can support supplementary, fixed data rates.

MPEG-1 Audio specifies three layers. The different layers offer increasingly higher audio quality at slightly increased complexity. While

Layers I and II share the basic structure of the encoding process having their roots in an earlier algorithm also known as MUSICAM [Dehéry, Stoll and Kerkhof 91], Layer III is substantially different. The Layer III algorithm was derived from the merge of ASPEC [Brandenburg et al. 91] with the Layer I and II filter bank, the idea being that a Layer III decoder should be able to decode Layer I and II bitstreams. Layer I is the simplest layer and it operates at data rates between 32 and 224 kb/s per channel. The preferred range of operation is above 128 kb/s. Layer I finds an application, for example, in the digital compact cassette, DCC, at 192 kb/s per channel. Layer II is of medium complexity and it employs data rates between 32 and 192 kb/s per channel. At 128 kb/s per channel it provides very good audio quality. A number of applications take advantage of Layer II including digital audio broadcasting, DAB, [ETS 300 401 v2] and digital video broadcasting, DVB [ETS 300 421, ETS 300 429, ETS 300 744]. Layer III exhibits the highest quality of the three layers at an increased complexity. The data rates for Layer III are lower than the rates for Layers I and II and they vary between 32 and 160 kb/s per channel. Layer III displays very good quality at rates below 128 kb/s per channel. Applications of Layer III include transmission over ISDN lines and internet applications. A modification of the MPEG Layer III format at lower sampling frequencies gave origin to the ubiquitous MP3 file format.

In spite of the differences in complexity, single-chip, real-time decoder implementations exist for all three layers. It should be noted that, in addition to the main audio data, all three layers provide a means of including auxiliary data within the bitstream syntax. Finally it should be mentioned that, MPEG-1 Layers II and III were also selected by ITU-R task group, TG, 10/2 for broadcasting applications in recommendation BS.1115. In ITU-R BS.1115, Layer II is recommended for emission at the data rate of 128 kb/s per channel, and for distribution and contribution at data rates above 180 kb/s per channel. Layer III is also recommended in BS.1115 for commentary broadcasting at data rates of about 60 kb/s per channel.

The main building blocks of the MPEG-1 audio coding scheme are shown in *Figure 1* and *Figure 2*. The basic building blocks include a time to frequency mapping stage followed by a bit or noise allocation stage. The input signal also feeds a psychoacoustic model block whose output determines the precision of the allocation stage. The bitstream formatting stage interleaves the representation of the quantized data with side information and optional ancillary data. The decoder interprets the bitstream, restores the quantized spectral components of the signal and finally reconstructs the time domain representation of the audio signal from its frequency representation.

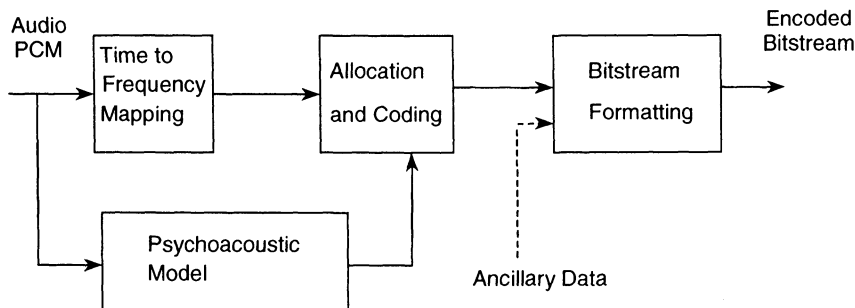


Figure 1. MPEG-1 Audio encoder basic building blocks

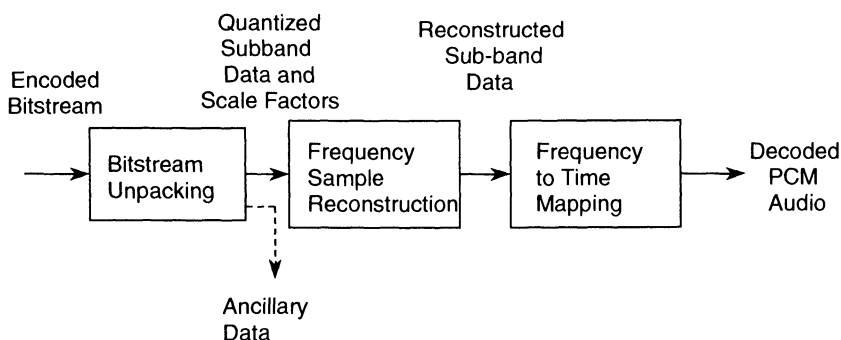


Figure 2. MPEG-1 Audio decoder basic building blocks

3.2 Different Layers Coding Options

The general approach to the coded representation of audio signals is the same for all layers. Based on the time to frequency mapping of the signals with a source model design based on statistics of generic audio signals, they share the basic building blocks and group the input PCM samples into frames of samples for analysis/synthesis. There are, however, a number of differences in the different layers' algorithms going from the simple approach of Layer I to the more sophisticated approach of Layer III at increased complexity. In *Figure 3* and *Figure 4* the block diagrams of Layers I, II, and III in single channel mode are shown.

3.2.1 Layers I and II

For Layers I and II, the time to frequency mapping is performed by applying a 32-band PQMF (see also Chapter 4) to the main audio path data. The frequency representation of the signal is scaled and then quantized with a uniform midtread quantizer (see also Chapter 2) whose precision is determined by the output of the psychoacoustic model. Typically, Psychoacoustic Model 1 (see also next sections) is applied, where the psychoacoustic analysis stage is performed with a 512-point FFT (Layer I) or 1024-point FFT (Layer II). In order to further reduce the data rate, Layer II applies group coding of consecutive quantized samples certain levels (see also next sections).

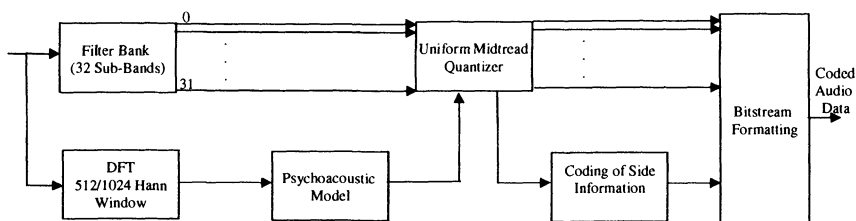


Figure 3. Block diagram of Layers I and II (single channel mode)

3.2.2 Layer III

For Layer III the output of the PQMF is fed to an MDCT stage (see also Chapter 5). In addition, the Layer III filter bank is not static as in Layers I and II, but it is signal adaptive (see next sections). The output of this hybrid filter bank is scaled and then non-uniformly quantized with a midtread quantizer. Noiseless coding is also applied in Layer III. In an iterative loop that performs the synthesis of the Huffman-encoded, quantized signal and compares its relative error levels with the masked thresholds levels, the quantizer step size is calculated for each spectral region. The quantizer step is once again determined by the output of the psychoacoustic model, however, the nature of the psychoacoustic model (Model 2, see next sections) applied to Layer III is substantially different from the model applied for Layers I and II. *Figure 4* highlights one of the differences, the analysis stage, which is performed by applying two, 1024-point FFTs. In all layers the audio data together with the side information such as bit allocation

and control parameters are multiplexed with the optional ancillary data and then stored or transmitted.

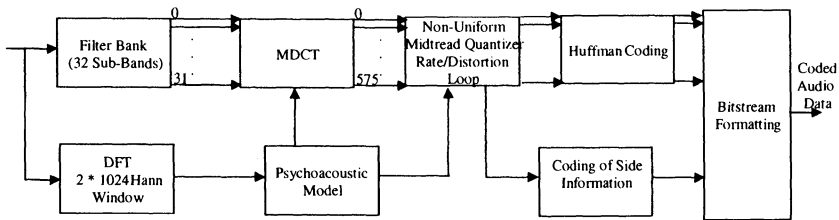


Figure 4. Block diagram of Layer III (single channel mode)

In the next sections, we describe the common characteristics of the audio coding algorithms in the three layers.

4. TIME TO FREQUENCY MAPPING

A PQMF filter bank (see also Chapter 4) is part of the time to frequency mapping stage for all three MPEG layers. This filter divides the frequency spectrum into 32 equally spaced frequency sub-bands. For Layers I and II the output of the PQMF represents the signal spectral data to be quantized. The frequency resolution of the Layer I and II filterbank is 750 Hz at a 48 kHz sampling rate. For Layer III, the PQMF is cascaded with an 18 frequency-line MDCT for a total of 576 frequency channels in order to increase the filter bank resolution.

4.1 Layer III Hybrid Filter Bank

The block diagram of Layer III filter bank analysis stage is shown in Figure 5. After the 32-band PQMF filter, blocks of 36 sub-band samples (for steady state conditions) are overlapped by 50 percent, multiplied by a sine window and then processed by the MDCT transform (see also Chapter 5). It should be noted that, in addition to the potential frequency aliasing introduced by the PQMF, the OTDAC transform introduces also time aliasing that cancels out between adjacent time-blocks in absence of quantization in the overlap-add stage of the decoder process. In order to lessen some of the artifacts potentially introduced by the overlapping bands of the PQMF, for long blocks (steady state conditions) the Layer III filter bank multiplies the MDCT output by coefficients that reduce the signal aliasing [Edler 92].

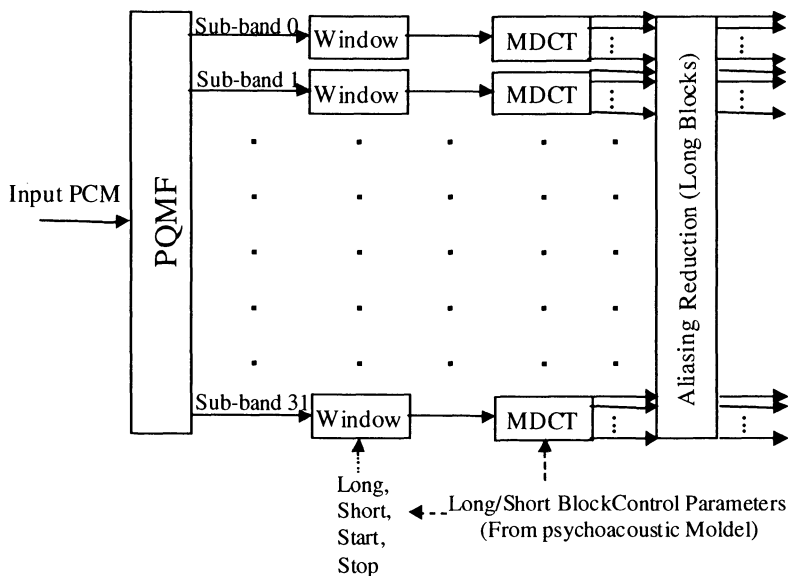


Figure 5. MPEG Audio Layer III analysis filter bank structure

In the decoder, the inverse aliasing reduction process is applied prior to the IMDCT in order to provide the correct sub-band samples for the PQMF synthesis stage for aliasing cancellation (see Figure 6). A pure sine wave signal processed by the hybrid PQMF/MDCT filter bank without aliasing reduction can present a spurious component as high as -12 dB with respect to the original signal. After the aliasing reduction process, the spurious component magnitude is reduced significantly. It should be noted, however, that, although the aliasing reduction process greatly improves the frequency representation of the signal, residual aliasing components might still be present. In the synthesis stage, the IMDCT is applied prior to the reconstruction PQMF. After the de-quantization of the spectral components and, when applicable, the joint stereo processing of the signal, the inverse aliasing reduction is applied. Next, the IMDCT is employed followed by the windowing process, where the windows applied are defined in the same manner as the analysis windows. The first half of the current windowed block is overlapped and added to the second half of the windowed samples of the previous block. For the long block the output of the overlap and add stage consists of 18 samples for each of the 32 synthesis PQMF sub-bands.

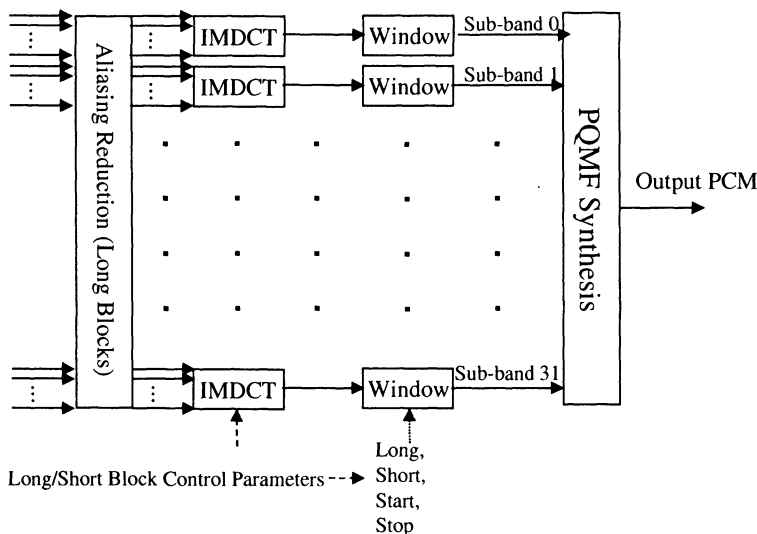


Figure 6. MPEG Audio Layer III synthesis filter bank structure

4.1.1 Block Switching

The total block size processed by the Layer III filter bank is given by $32 * 36 = 1152$ time-samples. This block length ensures a frequency resolution of about 41.66 Hz at 48 kHz sampling rate. The increased frequency resolution for Layer III is much better suited to accommodate allocation of the bit pool based on psychoacoustic models. One drawback of this approach is that quantization errors can now be spread over a block of 1152 time-samples. For signals containing transients, such as castanet excerpts, this translates into unmasked temporal noise, specifically pre-echo. In the case of transient signals, the Layer III filter bank can switch to a higher time resolution in order to avoid pre-echo effects. Namely, during transients, Layer III utilizes a shorter block size of $32 * 12 = 384$ time samples, reducing the temporal spreading of quantization noise for sharp attacks. The short block size represents a third of the long block length. During transients, a sequence of three short windows replaces the long window, maintaining the same total number of samples per frame. In order to ensure a smooth permutation between long and short blocks and vice versa, two transition blocks, long-to-short and short-to-long, which have the same size as the long block are employed. This approach was first presented by Edler [Edler 89] and, based on the shape of the windows and overlap regions, it

maintains the time domain aliasing cancellation property of the MDCT (see also Chapter 5). In addition, the frame size is kept constant during the allowed window size sequences. This characteristic ensures the subsistence of a simple overall structure of the algorithm and bitstream formatting routines.

4.1.1.1 Window Sequence

In *Figure 7* a typical window sequence from long to short windows and from short to long windows is shown together with the corresponding amount of overlapping between adjacent windows in order to maintain the time-domain aliasing cancellation for the transform. The basic window, $w[n]$, utilized by Layer III is a sine widow. The different windows are defined as follows:

$$w[n] = \sin\left(\frac{\pi}{36}\left(n + \frac{1}{2}\right)\right) \quad n = 0, \dots, 35 \quad (\text{long window})$$

$$w[n] = \sin\left(\frac{\pi}{12}\left(n + \frac{1}{2}\right)\right) \quad n = 0, \dots, 11 \quad (\text{short window})$$

$$w[n] = \begin{cases} \sin\left(\frac{\pi}{36}\left(n + \frac{1}{2}\right)\right) & n = 0, \dots, 17 \\ 1 & n = 18, \dots, 23 \\ \sin\left(\frac{\pi}{12}\left(n - 18 + \frac{1}{2}\right)\right) & n = 24, \dots, 29 \\ 0 & n = 30, \dots, 35 \end{cases} \quad (\text{start window})$$

$$w[n] = \begin{cases} 0 & n = 0, \dots, 5 \\ \sin\left(\frac{\pi}{12}\left(n - 6 + \frac{1}{2}\right)\right) & n = 6, \dots, 11 \\ 1 & n = 12, \dots, 17 \\ \sin\left(\frac{\pi}{36}\left(n + \frac{1}{2}\right)\right) & n = 18, \dots, 35 \end{cases} \quad (\text{stop window})$$

It should be noted that the Layer III filter bank allows for a mixed block mode. In this mode, the two lower frequency PQMF sub-bands are always processed with long blocks, while the remaining sub-bands are processed with short blocks during transients. This mode ensures high frequency

resolution at low frequencies where it is most needed and high time resolution at high frequencies.

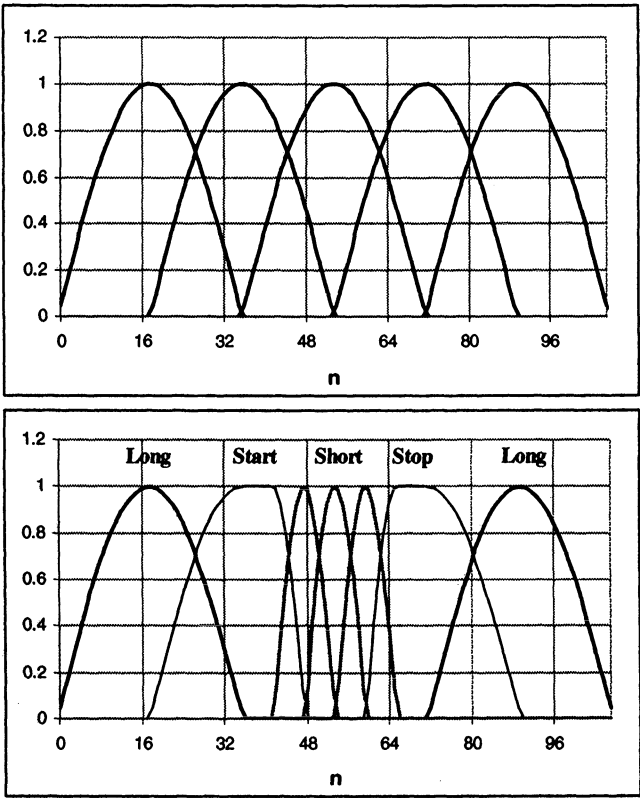


Figure 7. Typical Layer III window sequence: top for steady state signal, bottom for transients occurring in the time region between $n = 45$ and $n = 60$

4.1.2 Hybrid Filter Bank Versus PQMF Characteristics

The Layer III hybrid filter bank provides much higher frequency resolution than the Layers I and II PQMF. The time resolution, however, is decreased. At 48 kHz sampling rate, the time resolution for Layers I and II is 0.66 ms, for Layer III is 4 ms. The decreased time resolution renders Layer III more prone to pre-echo. A number of measures to reduce pre-echo are incorporated in Layer III including a detection mechanism in the

psychoacoustic model and the ability to “borrow” bits from the bit reservoir (see also next sections) in addition to block switching. The inherent filter bank structure with long impulse response of $384 + 512 = 896$ samples even in the short block mode, however, makes the encoding of transients a challenge for Layer III.

In summary, the Layer III hybrid filter bank approach offers advantages such as high frequency resolution, a dynamic, adaptive trade-off between time and frequency resolution, and full compatibility with Layers I and II. The shortcomings include potential aliasing effects exposed by the MDCT stage and long impulse response filters. Both shortcomings are reduced in the standard specifications by adopting procedures to mitigate them. The complexity of the Layer III filter bank is increased with respect to the complexity of Layers I and II. In addition to the PQMF, the MDCT stage contributes to its complexity. In general, fast implementations of the MDCT exploit the use of FFTs. It should be noted that the size of the MDCT is non-power of two, therefore the implementation via FFT requires a decomposition to a power-of-two length sequence if a radix-2 FFT is utilized. Considering the different stages of the filter bank implementation and assuming that the MDCT is implemented via an FFT, the complexity for a long window is given by $(18 + 9 + 18)$ additional complex multiplications and additions per sub-band block with respect to the PQMF alone, or equivalently a little over 1 additional multiplication and addition per sub-band sample.

5. MPEG AUDIO PSYCHOACOUSTIC MODELS

The goal of MPEG Audio is to provide perceptually lossless quality. In other words, the output of the MPEG coder should be a signal statistically indistinguishable from its input. In order to achieve this objective at relatively low data rates, MPEG Audio exploits the psychoacoustic principles and models we discussed in Chapter 7. During the encoding process, the input signal is analyzed on a frame-by-frame basis and the masking ability of the signal components is determined. For each frame, based on the computed masked thresholds, the bits available are distributed through the signal spectrum in order to best represent the signal. Although the encoder process is not a mandatory part of the MPEG standard, two psychoacoustic models are described in the informative part of its specifications. Either model works for all layers, but typically Model 1 is applied to Layers I and II and Model 2 to Layer III. There is a large degree of freedom in the psychoacoustic model implementation. At high data rates, the psychoacoustic model can be completely bypassed, leaving the task of

assigning the available resource to the iterative process in the allocation routines simply based on the strength of the signal spectral components.

5.1 Psychoacoustic Models Analysis Stage

The input to the psychoacoustic model is the time representation of the audio signal over a certain time interval and the corresponding outputs are the signal to mask ratios (SMRs) for the coder's frequency partitions. Based on this information, the bit (Layers I and II) and noise (Layer III) allocation is determined for each block of input data (see Chapter 8 for a discussion on perceptual bit allocation). In order to provide accurate frequency representation of the input signal, a discrete Fourier transform is computed in parallel to the main audio path time to frequency mapping stage. One might argue that the output of the PQMF or the hybrid filter bank could be utilized for this purpose in order to simplify the structure of the algorithm. In the case of the evaluation of the masking thresholds, the aim is to have maximum accuracy in the signal representation. While issues like critical sampling etc. play a fundamental role in the design of the time to frequency mapping in the main audio path, they are irrelevant in the frequency representation of the audio signal for analysis only purposes. On the other hand, inadequate frequency resolution and potential aliasing can irremediably confound the evaluations of the psychoacoustic model. It should be noted that different approaches are found in the literature. For example, in Dolby AC-3 [Fielder et al. 96] and PAC [Sinha, Johnston, Dorward and Quackenbush 98] the output of the MDCT is employed for the psychoacoustic analysis; in the advanced PEAQ version the DFT analysis is employed along with a filter bank that mirrors the auditory peripheral filter bank [Thiede et al. 00].

The first step in both MPEG psychoacoustic models is to time-align the audio data used by the psychoacoustic model stage with the main path audio data. This process must take into account the delay through the filter bank and the time offset needed so that the psychoacoustic analysis window is centered on the current block of data to be coded. For example, in Layer I the delay through the filter bank is 256 samples and the block of data to be coded is 384 samples long (see also next section). The analysis window applied to Layer I data in Psychoacoustic Model 1 is 512 samples long. The offset to be applied for time alignment is therefore $256 + (512 - 384)/2 = 320$ samples.

5.2 Psychoacoustic Model 1

The block diagram for Psychoacoustic Model 1 is shown in *Figure 8*. The first stage, the analysis stage, windows the input data and performs an FFT. The analysis window is a Hanning window of length N equal to 512 samples for Layer I and 1024 for Layers II and III. The overlapping between adjacent windows is $N/16$. Since Layers II and III utilize a 1152 sample frame, the 1024 sample analysis window does not cover the entirety of the audio data in a frame. If, for example, a transient occurs at the tail end of the main path audio frame, the relative sudden energy change would be undetected in the Psychoacoustic Model 1 analysis window. In general, however, the 1024 sample analysis window proved to be a reasonable compromise.

5.2.1 SPL Computation

After applying the FFT, the signal level is computed for each spectral line k , L_k as follows

$$L_k = 96 \text{ dB} + 10 \log_{10} \left(4 / N^2 |X[k]|^2 8/3 \right) \quad \text{for } k=0, \dots, N/2-1$$

where $X[k]$ represents the FFT output of the time-aligned, windowed input signal and N equals 512 for Layer I and 1024 for Layers II and III. The signal level is normalized so that the level of an input sine wave that just overloads the quantizers, here defined as being at $x[n] = \pm 1.0$, has a level of 96 dB when integrated over the peak. In this equation, the factor of $1/N^2$ comes from Parseval's theorem, one factor of 2 comes from only working with positive frequency components, another factor of 2 comes from the power of a unit amplitude sinusoid being equal to $1/2$, and the factor of $8/3$ comes from the reduction in gain from the Hanning window (see also Chapter 9). Since the description of Model 1 in the standard absorbs the factor of $8/3$ into the Hanning window definition, a natural way to take the other factors into account is to include a factor of $2/N$ in the forward transform of the FFT.

The sound pressure level in each sub-band m , $L_{sb}[m]$ is then computed as the greater of the SPL of the maximum amplitude FFT spectral line in sub-band m and the lowest level that can be described with the maximum scale factor for that frame in sub-band m as follows:

$$L_{sb}[m] = \max \{ L_k, 20 \log_{10} (\text{scf}_{\max}[m] 32,768) - 10 \} \text{dB}$$

where L_k represents the level of the k th line of the FFT in sub-band m with the maximum amplitude and $scf_{\max}$ is the maximum of the scale factors for sub-band m (see the next section for a discussion on MPEG Layers I and II “scale factors” which differ somewhat from the scale factors discussed in Chapter 2 in that these “scale factors” represent the actual factor that the signal is scaled by as opposed to the number of factors of 2 in the scaling). In Layers I and II coders, the scale factors range from a very small number up to 2.0 so the multiplication by 32,768 just normalizes the power of the scale factor so that the largest possible scale factor corresponds to a level of 96 dB. The -10 dB term is an adjustment to take into consideration the difference between peak and average levels. The reason for taking the scale factor into account in the above expression can be explained by closely examining the block floating quantization process, since block floating point quantization cannot scale the signal to lower amplitudes than can be represented by the scale factors themselves. This implies that for low amplitude frequency lines the quantization noise is of a size determined by the maximum scale factor.

5.2.2 Separation of Tonal and Non-Tonal Components

Having found the sound pressure level in the sub-band, we next compute the masking threshold in order to calculate the signal to mask ratio (SMR) for the sub-band. Since noise is a better masker than tones, a search for tonal maskers in the signal is performed in Model 1. This evaluation is based upon the assumption that local maxima within a critical band represent the tonal components of the signal. A local maximum L_k , is included in the list of tonal components if $L_k - L_{k+j} \geq 7$ dB where the index j varies with the center frequency of the critical band examined. If L_k represents a tonal component, then the index k , the sound pressure level derived from the sum of three adjacent spectral components centered at k , L_T , and a tonal flag define the tonal masker.

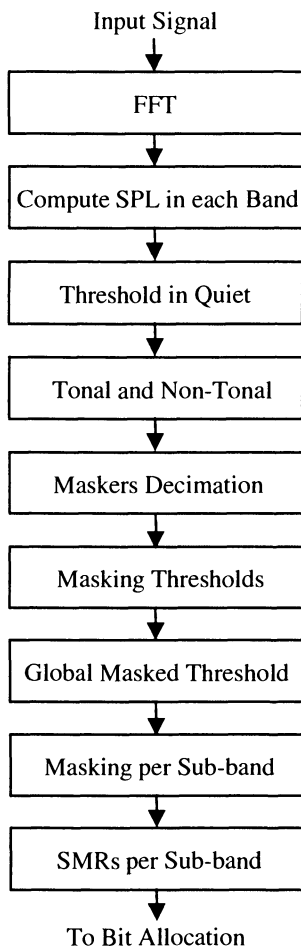


Figure 8. Block diagram of MPEG Psychoacoustic Model 1

The noise maskers in the signal are derived from the remaining spectral lines. Within a critical band, the power of the spectral components remaining after the tonal components are removed is summed to form the sound pressure level of the noise masker, L_N , for that critical band. The noise masker components for each critical band are centered at the geometric mean of the FFT spectral line indices for each critical band.

5.2.3 Maskers Decimation

Having defined the tonal and noise maskers in the signal, the number of maskers is then reduced prior to computing the global masked threshold. First, tonal and non-tonal maskers are eliminated if their levels do not exceed the threshold in quiet. Second, maskers extremely close to stronger maskers are eliminated. If two or more components are separated in frequency by less than 0.5 bark, only the component with the highest power is retained. The masking thresholds are then computed for the remaining maskers by applying a spreading function and shifting the curves down by a certain amount of dB which depends on whether the masker is tonal or noise-like and the frequency position of the masker. To keep the calculation time manageable, the masking thresholds are evaluated at only a sub-sampling of the frequency lines. The number of sub-sample lines depends on which layer is being implemented and on the sampling rate, ranging from 102 to 108 sub-sample lines in Layer I and ranging from 126 to 132 sub-sample lines in Layer II.

5.2.4 Model 1 Spreading Function and Excitation Patterns

The spreading function used in Model 1 is defined as follows:

$$B(dz, L) = -17 \, dz + 0.15L \, (dz - 1) \, \theta(dz - 1) \quad \text{for } dz \geq 0$$

$$B(dz, L) = -(6 + 0.4L) \, |dz| - (11 - 0.4L) \, (|dz| - 1) \, \theta(|dz| - 1) \quad \text{for } dz < 0$$

where $\theta(x) = 1$ for $x \geq 0$ and 0 otherwise; L represents the sound pressure level of the masker; dz represents the distance in bark between the maskee and the masker. Notice that dz assumes positive values when the masker is located at a lower frequency than the maskee and negative values when the masker is located at a higher frequency than the maskee. The two-piece linear spreading function for upper and lower frequencies in Model 1 seems to be consistent with the masking data for tones masking tones (see also Chapter 6). The excitation patterns for a signal positioned at 8 bark with different sound pressure levels, $L = 20, 40, 60, 80, 100$ dB, is shown in *Figure 9*. As shown in *Figure 9* the excitation patterns are level dependent, being nearly symmetrical for levels below 40 dB and increasingly spreading towards high frequencies at higher levels.

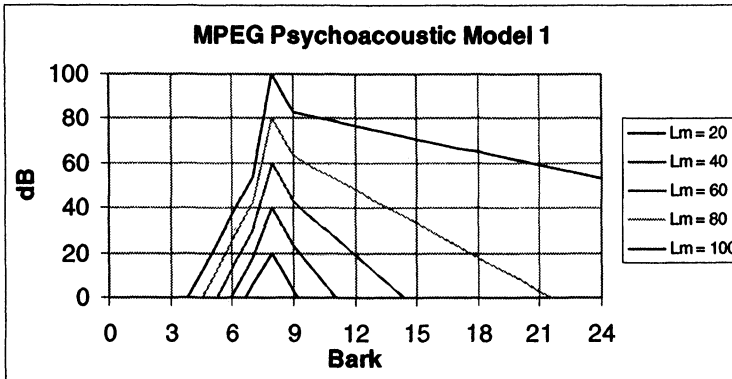


Figure 9. MPEG Psychoacoustic Model 1 excitation patterns for different SPLs

5.2.5 Masking Thresholds

In order to derive the masking threshold relative to the masker of level L , one needs to shift the excitation pattern relative to the masker by an appropriate amount. This shift depends on the tonal versus noise-like characteristics of the masker, since we know from experimental data that noise is a better masker than tones. In Model 1 the shift Δ is defined as follows:

$$\Delta_T(z) = -6.025 - 0.275 z \text{ dB}$$

$$\Delta_N(z) = -2.025 - 0.175 z \text{ dB}$$

where Δ_T represents the shift for tonal maskers and Δ_N represents the shift for noise-like maskers, and z is the frequency index of the masker in the bark scale. The masking threshold M relative to a masker of level L at a frequency equal to bark z can be expressed as:

$$M_{T,N}(L, z_i, z_j) = B(z_i - z_j, L) + \Delta_{T,N}(z_j)$$

Where z_j is the bark index of the masker and z_i is the bark index of the maskee, $\Delta = \Delta_T$ for tonal maskers and $\Delta = \Delta_N$ for noise-like maskers. In Figure 10 the masking curve relative to a 70 dB noise-like masker centered at bark 8 is shown.

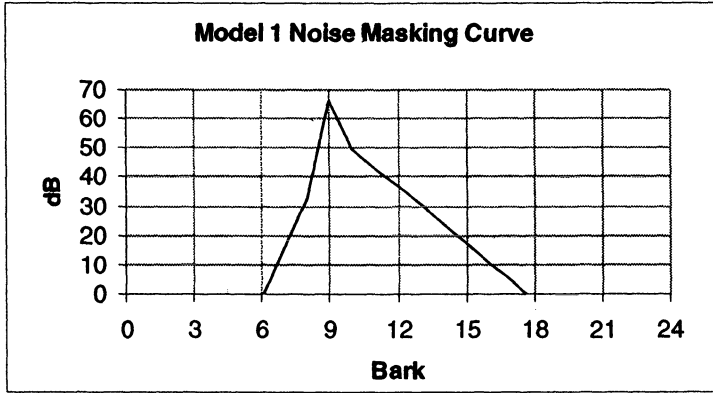


Figure 10. MPEG Psychoacoustic Model 1 masking curve for a noise-like masker centered at 8 Bark with a sound pressure level of 70 dB

5.2.6 Masked Threshold

The global masked threshold $M_G(z_i)$ at the bark location z_i of each sub-sampled frequency line is then computed in Model 1 by summing the power of the individual masking thresholds and the threshold in quiet as follows:

$$M_G(z_i) = 10 \log_{10} \left[10^{\frac{M_q(z_i)}{10}} + \sum_{j=1}^m 10^{\frac{M_{T_j}(L_j, z_i, z_j)}{10}} + \sum_{k=1}^n 10^{\frac{M_{N_k}(L_k, z_i, z_k)}{10}} \right]$$

where M_q represents the threshold in quiet at the bark location z_i , M_{T_j} the masking threshold from the j^{th} tonal masker, M_{N_k} the masking threshold from the k^{th} noise-like masker, m is the total number of tonal maskers, and n is the total number of noise-like maskers.

5.2.7 SMR Computation

Since at low frequency some of the main path PQMF sub-bands span more than one critical band, the masking level $M_{Gmin}(sb)$ in each sub-band, sb , is determined based on the minimum masking level at the sub-sampled lines in that sub-band and used to determine the signal to mask ratio $SMR(sb)$ to be transmitted to the bit allocation routine:

$$SMR(sb) = L(sb) - M_{Gmin}(sb)$$

where $L(sb)$ is the signal level for sub-band sb .

5.2.8 An Example

To help clarify the Model 1 calculations, *Figure 11* through *Figure 13* give an example taken from [Pan 95]. *Figure 11* shows the frequency spectrum of an input signal consisting of a single tonal component overlaying a shaped noise spectrum. The top part of *Figure 12* shows the signal level of each masker identified in the signal and its associated frequency location (measured in terms of the k index in the $N = 1024$ FFT). The tonal masker is clearly visible as are the noise maskers from each critical band. The middle part of *Figure 12* shows the maskers remaining after eliminating maskers below the threshold in quiet or too close to stronger maskers. The bottom part of *Figure 12* shows the resulting global masked threshold at each of the sub-sampled frequency locations. The top of *Figure 13* shows the resulting SMRs calculated for each of the 32 critical bands from the input signal spectrum and the masked curve. The bottom of *Figure 13* shows the resulting decoded data after being quantized using bit allocations determined by the SMRs in the top of the figure at a data rate of 64 kb/s.

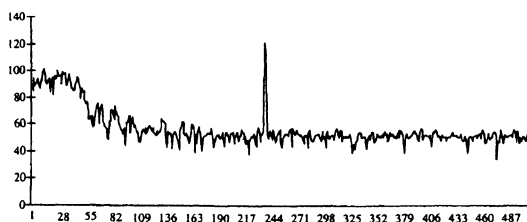


Figure 11. Low-pass filtered noise plus 11.250 kHz sine wave at a sampling rate of 48kHz; the abscissa represents the FFT frequency line indices, the ordinate the strength of the signal in dB. (From [Pan 95] © 1995 IEEE.)

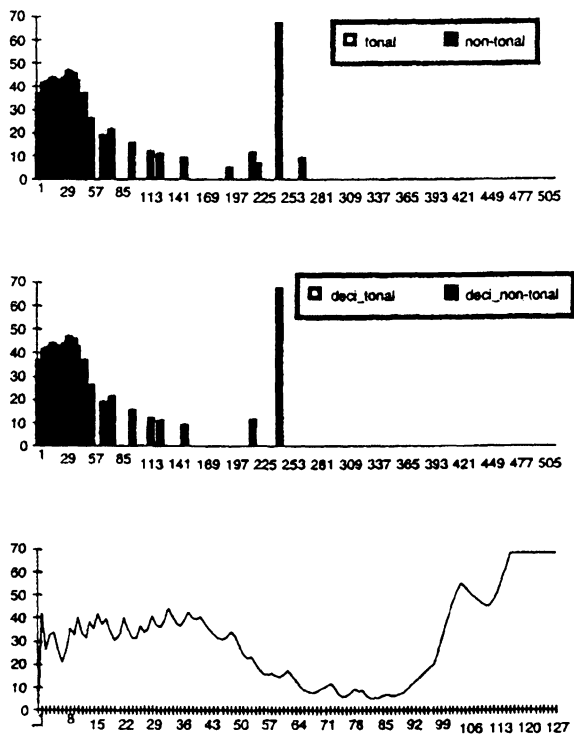


Figure 12. MPEG Psychoacoustic Model 1 processing from [Pan 95] © 1995 IEEE.

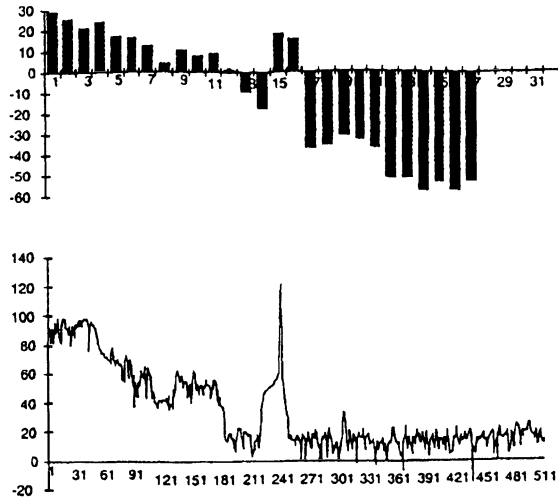


Figure 13. MPEG Psychoacoustic Model 1 SMRs (top) and coded audio data (bottom) at 64 kb/s. (From [Pan 95] © 1995 IEEE.)

5.3 Psychoacoustic Model 2

Model 2's process and calculations differ substantially from those of Model 1. A block diagram of Model 2 is shown in *Figure 14*. Notice that there are now two parallel calculation paths: that of the masking energy and that of the tonality index. In addition, Model 2 does not have a masker decimation stage.

5.3.1 Model 2 Analysis Stage

As discussed in Model 1, Model 2 also applies a Hanning-windowed FFT to the time-aligned signal samples. The size of the FFT window is now 1024 for all Layers, however for Layers II and III this model computes two FFTs for each frame. The centers of the FFT input blocks are aligned with the center of the first and second half of the main data path input block. Model 2 uses the output of the FFT analysis to calculate the masking curves and their associated signal to mask ratios for the coder sub-bands. The higher of the two resulting signal to mask ratios, or the lower of the two masked thresholds, are then selected per each sub-band.

5.3.2 SPL Computation

In Model 2 the frequency-lines are grouped into “threshold calculation partitions” whose widths are roughly 1/3 of a critical band or one FFT line, whichever is wider at that frequency location. For each partition, one single masker SPL is derived from the sum of energy densities in the partition.

The total masking energy for the signal frame is computed by first convolving a spreading function with each of the maskers in the signal. This process is equivalent to spreading in frequency the energy of each masker and adding up the relative energies. The spreading function used in Model 2 differs from that used in Model 1 and is a variant of the Schroeder spreading function we discussed in Chapter 7.

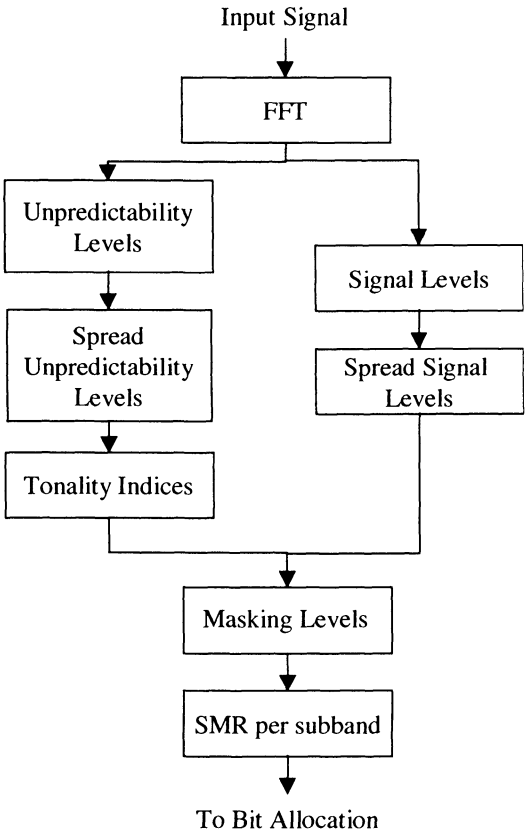


Figure 14. Block diagram of MPEG Psychoacoustic Model 2

5.3.3 Model 2 Spreading Function and Excitation Patterns

The basic spreading function $B(dz)$ measured in dB used by Model 2 is (see also *Figure 15*):

$$B(dz) = 15.8111389 + 7.5 * (1.05 * dz + 0.474) - 17.5 * \sqrt{1.0 + (1.05 * dz + 0.474)^2} \\ + 8 * \text{MIN}(0, (1.05 * dz - 0.5)^2 - 2 * (1.05 * dz - 0.5))$$

where dz is the bark distance between the maskee and the masker. Notice that $dz < 0$ has the masker at a higher frequency than the maskee. Notice also that, other than a minor factor of 1.05, the first line is exactly equal to Schroeder's spreading function. The second line lowers the curve for values of dz between 0.5 and 2.5, causing somewhat of a two-sloped shape for positive dz (see *Figure 15*). Furthermore, notice that this spreading function is not level dependent in contrast with the clear level dependence seen in experimental data (see also Chapter 6).

In Model 2 the spreading function is modified by a normalization that preserves power near zero frequency and the upper frequency limit. Namely, a flat spectrum over the positive frequency range of the FFT is convolved with $B(dz)$. The power of the convolved flat spectrum drops at the edges of the positive frequency range due to a lack of spectrum energy outside this range. The spreading function $B(dz)$ is adjusted to remove this drop by dividing the convolved signal power by the power of the convolved flat spectrum. Given the behavior of the threshold in quiet at the boundaries of the human audible range (see Chapter 6), this normalization may not be essential. In order to derive the global masked threshold, this convolved, normalized signal energy is then lowered in each partition by an amount that depends on the tonality of the spectrum in that partition.

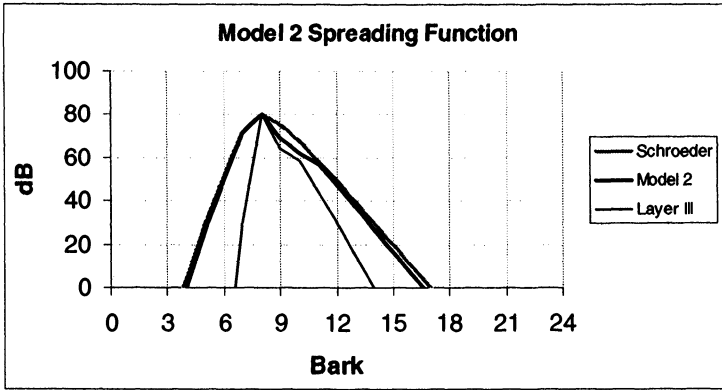


Figure 15. Basic spreading function used by MPEG Psychoacoustic Model 2 compared with the Schroeder spreading function and with the modified version utilized in Layer III.

5.3.4 Tonality Index

The tonality index in each partition is calculated based on how predictable the signal is from the frequency lines in the two prior frames, see also [Brandenburg and Johnston 90]. For each frame m and for each frequency line k , the signal amplitude, $A_m[k]$, and phase, $\phi_m[k]$, are predicted by linear extrapolation from the two prior values as follows:

$$A'_m[k] = A_{m-1}[k] + \{A_{m-1}[k] - A_{m-2}[k]\}$$

$$\phi'_m[k] = \phi_{m-1}[k] + \{\phi_{m-1}[k] - \phi_{m-2}[k]\}$$

where $A'_m[k]$, and $\phi'_m[k]$ represent the predicted values. The predicted values are then mapped into an “unpredictability measure” defined as:

$$c_m[k] = \frac{\sqrt{\{A'_m[k] \cos \phi'_m[k] - A_m[k] \cos \phi_m[k]\}^2 + \{A'_m[k] \sin \phi'_m[k] - A_m[k] \sin \phi_m[k]\}^2}}{A_m[k] + |A'_m[k]|}$$

where $c_m[k]$ is equal to zero when the current value is exactly predicted and equal to 1 when the power of either the predicted or actual signal is dramatically higher than the other.

The unpredictability measure is first weighted with the energy in each partition, deriving a partition unpredictability measure. This partition

unpredictability measure is then convolved with the spreading function. The result of this convolution is normalized using the signal energy convolved with the spreading function and then mapped onto a tonality index which is a function of the partition number and whose values vary between zero and one. The tonality index has the property that high unpredictability goes towards zero and low unpredictability goes towards one. Notice that, since the unpredictability measure is convolved using the spreading function that determines the masking energy at a certain frequency location, the resulting tonality index at a frequency location reflects the tonality of the dominant maskers at that frequency location.

5.3.5 Masking Thresholds

The tonality indices for each partition created by this calculation process are utilized to determine the shift $\Delta(z)$ in dB of the re-normalized convolved signal energy that converts it into the global masking level. The values for $\Delta(z)$ are linearly interpolated based on the tonality index from the value of 5.5 dB for zero tonality (a noise masker) to a frequency-dependent value defined in the standard for tonality index equal to one (a tonal masker). The interpolated $\Delta(z)$ is then compared with a frequency-dependent minimum value, also defined in the standard, that controls stereo unmasking effects (see also Section 8 later in this chapter) and the larger value is used for the shift.

5.3.6 SMR Computation

The masking levels in each threshold calculation partition are compared with the threshold in quiet for that partition and the larger is used. Note that this differs from the addition of masking power from the threshold in quiet to that of the maskers in Model 1. This curve represents the masked threshold. For each partition, the masking energy is evenly mapped to each frequency line by dividing the partition masking threshold values by the number of frequency lines in that partition. Finally, the masked threshold and the corresponding power spectral densities are mapped onto the coder scale factors sub-bands (see next section) to calculate the signal to mask ratios (SMRs) for each sub-band. These SMRs are then sent to the allocation routine to determine the number of bits allocated to each sub-band.

5.3.7 An Example

In *Figure 16* and *Figure 17*, an example of Model 2 processing for the signal shown in *Figure 11* is illustrated from [Pan 95]. In *Figure 16*, the

signal level and its spread level (top) and the tonality index (bottom) are shown versus the 62, approximately one-third critical-band frequency partitions. The masked threshold is plotted in *Figure 17* (top) as derived from the convolved energy and the tonality index versus the FFT frequency lines. *Figure 17* (center) shows the SMR values for each coder scale factor band. Finally, in *Figure 17* (bottom) the coded audio data at 64 kb/s is shown. Note that the dB values shown differ by roughly 50 dB from those in the Model 1 example due to a different choice of normalization in the SPL calculations. This can be most easily seen by comparing the level of the tonal peak in both cases.

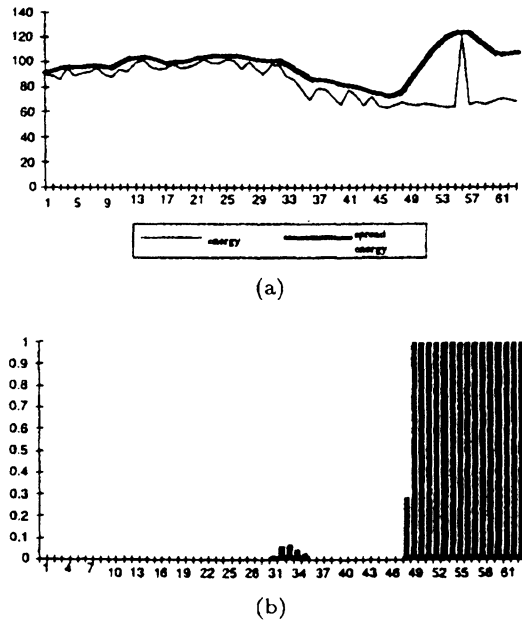


Figure 16. MPEG Psychoacoustic Model 2 processing: the abscissa represents the partition index and the ordinate the signal energy and spread energy in dB (top) and the tonality index (bottom). (From [Pan 95] © 1995 IEEE.)

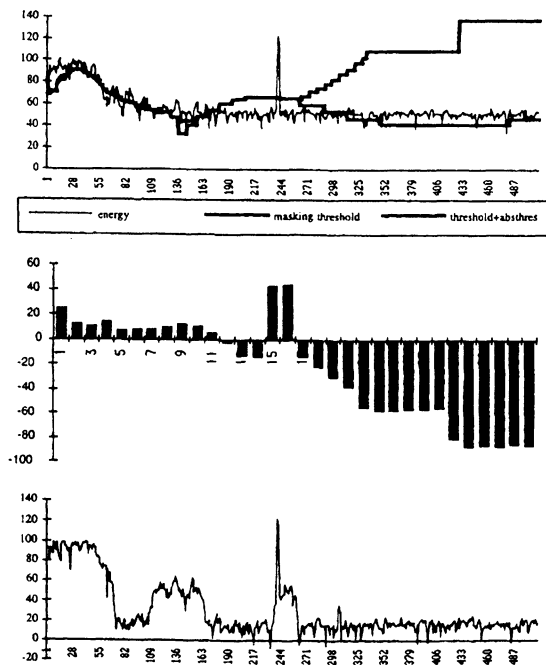


Figure 17. MPEG Psychoacoustic Model 2 processing: original signal and computed masking thresholds (top); SMRs (center); coded signal at 64 kb/s. (From [Pan 95] © 1995 IEEE.)

5.3.8 Model 2 Applied to Layer III

Model 2 is slightly altered when applied to Layer III to take into account the different nature of the Layer III algorithm including the hybrid filter bank and the block-switching capability of Layer III. The main changes are:

- The model is calculated twice in parallel – once for using a long block FFT and once using a short block FFT, where the short block consists of 256 samples.
- The results of both FFTs are combined in calculating the unpredictability measures.
- The spreading function (see Figure 15) is changed to drop off faster. Terms of $1.05 \cdot dz$ in the spreading function are replaced with terms of $1.5 \cdot dz$ for negative dz (masker frequency location lower than maskee frequency location) and $3.0 \cdot dz$ for positive dz (masker frequency location higher than maskee frequency location).

- The noise-masking-tone drop has been changed from 5.5 dB to 6.0 dB and the tone-masking-noise drop has been changed from a frequency-dependent table lookup to a constant 29.0 dB.
- Pre-echoes are reduced by comparing the masking threshold with scaled versions of the last 2 frames' thresholds and setting the current masking threshold to the smallest of these before comparing it with the threshold in quiet. (Twice the previous frame's value is used and 16 times the value of the frame before that.) This serves to reduce the masking thresholds following very quiet signals.

Attacks are detected based on a “psychoacoustic entropy” or perceptual entropy, PE (see also Chapter 7), calculation [Johnston 88] equal to the logarithm of a geometric mean of the threshold-weighted energy across the block. The exact formula as described in the informative section of the standard is

$$PE = \sum_{\text{partition } b} n_b \log_2 (1 + \sqrt{\text{energy}_b / \text{threshold}_b})$$

where n_b is the number of frequency lines in partition b , energy_b is the aggregate signal energy in partition b and threshold_b is the energy relative to the masked threshold in partition b . The PE as described above measures the minimum number of bits per block required to achieve transparency in a signal given a masking curve. Psychoacoustic entropy above a certain trigger (PE > 1800 bits from [ISO/IEC 11172]) signals an attack. The PE is effectively a measure of the perceptual flatness of the frequency spectrum since it effectively measures the geometric mean of the spectral energy (normalized by the masking threshold). Recall from Chapter 8 that a high geometric mean signifies a flat spectrum. A flat spectrum is associated with a sharp attack in this model since sudden changes lead to broad frequency content. Detection of an attack determines which window-type is used according to *Figure 18*. Signal-to-mask ratios are computed for each scale factor partition, rather than for each PQMF sub-band, corresponding to the chosen window-type.

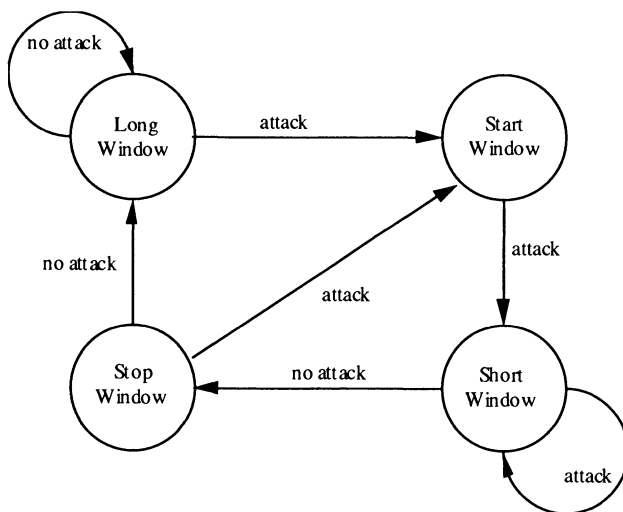


Figure 18. Block switching state diagram for Layer III from [ISO/IEC 11172-3]

6. MPEG-1 AUDIO SYNTAX

The MPEG Audio bitstream provides periodically spaced frame headers to identify the coded data. In *Figure 19* the audio data frame structure is shown for all Layers. For Layer I the frame size consists of $12 * 32 = 384$ samples. For Layers II and III, three granules of 12 samples per sub-band are grouped into a frame for a total of 1152 samples. The 32-bit MPEG Audio header is shown in *Figure 20*. After a 12-bit string of 1 which represents the synch word, there is an MPEG ID bit (=1 implies an MPEG-1 Audio bitstream) and the layer identification (two bits). The error protection bit, the bitrate index (four bits), the sampling frequency bits (two), padding and private bit and mode (stereo, joint stereo, dual channel, single channel), mode extension, etc., follow. The frame format of the three layers is shown in *Figure 21*. In order to understand the underlying structure of the MPEG Audio layers syntax, we now examine the details of different coding options adopted by the layers.

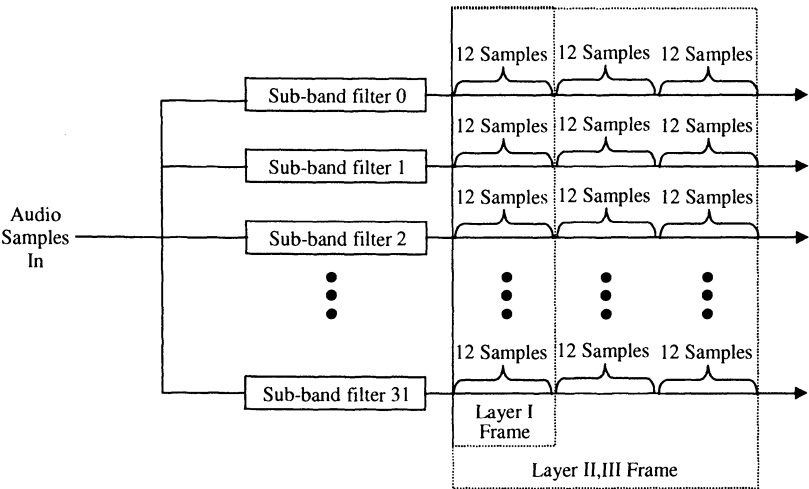


Figure 19. Data frame structure for MPEG-1 Audio Layers I, II, and III

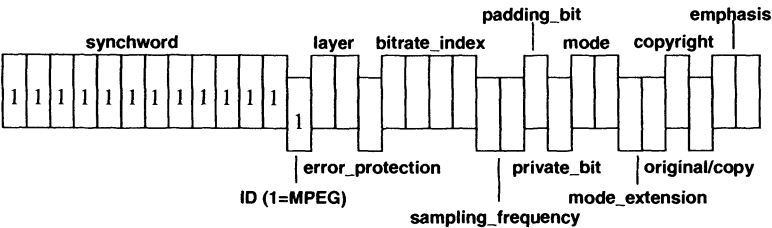


Figure 20. MPEG-1 Audio Header

LAYER I	Header (32)	CRC (0,16)	Bit Allocation (128-256)	Scale Factors (0-384)	Samples	Ancillary Data
LAYER II	Header (32)	CRC (0,16)	Bit Allocation (26-188)	SCFSI (0-120)	Scale Factors (0-1080)	Samples
LAYER III	Header (32)	CRC (0,16)	Side Information (130-246)	Main Data (May start at a previous frame)		

Figure 21. MPEG-1 Audio frame format

6.1 Layer I

Layer I encodes the audio samples in groups of 384 samples. The audio data in each frame represent 8 ms of audio at a 48 kHz sampling rate. The encoded audio data is described in terms of bit allocation information, scale factors, and quantized samples. In addition to the header and the audio data, each Layer I frame contains (see top part of *Figure 21*) a 16-bit optional cyclic redundancy code (CRC) error check word, and optional ancillary data.

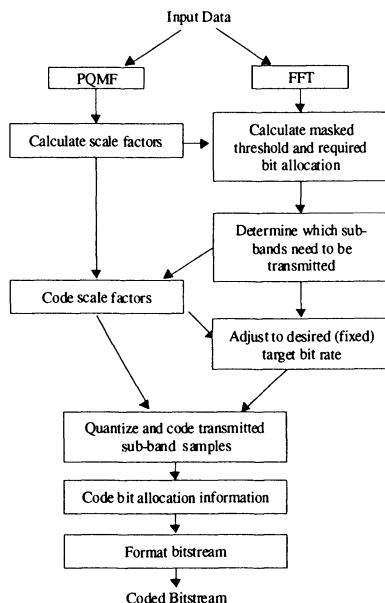


Figure 22. Basic structure of the encoding process for MPEG-1 Audio Layers I and II from [ISO/IEC 11172-3]

6.1.1 Scale Factors Computation

The Layer I (and Layer II) encoding process is shown in *Figure 22*. The analysis PQMF stage is followed by the computation of scale factors for each sub-band. One scale factor is computed for each 12 sub-band frequency samples (called a “granule”) and it is represented using 6 bits. The maximum absolute value of the 12-sample granule is determined and mapped to a scale factor value via a look up table defined in the standard and shown in *Table 1*. The samples in the granule are divided by the scale factor

prior to the quantization stage. The dynamic range covered by the scale factors is 120 dB.

The scale factor is transmitted only if the bit allocation for that band is non-zero. The maximum number of bits used in a Layer I frame to transmit scale factors (in the case where the bit allocation for each sub-band is non-zero) is $6 \times 32 \times 2 = 384$ bits for stereo mode and $6 \times 32 = 192$ bits for mono mode.

Table 1. MPEG Audio Layers I and II Scale Factors [ISO/IEC 11172-3]

Scale Factor Index	Scale Factor Value	Scale Factor Index	Scale Factor Value
0	2.00000000000000	32	0.00123039165029
1	1.58740105196820	33	0.00097656250000
2	1.25992104989487	34	0.00077509816991
3	1.00000000000000	35	0.00061519582514
4	0.79370052598410	36	0.00048828125000
5	0.62996052494744	37	0.00038754908495
6	0.50000000000000	38	0.00030759791257
7	0.39685026299205	39	0.00024414062500
8	0.31498026247372	40	0.00019377454248
9	0.25000000000000	41	0.00015379895629
10	0.19842513149602	42	0.00012207031250
11	0.15749013123686	43	0.00009688727124
12	0.12500000000000	44	0.00007689947814
13	0.09921256574801	45	0.00006103515625
14	0.07874506561843	46	0.00004844363562
15	0.06250000000000	47	0.00003844973907
16	0.04960628287401	48	0.00003051757813
17	0.03937253280921	49	0.00002422181781
18	0.03125000000000	50	0.00001922486954
19	0.02480314143700	51	0.00001525878906
20	0.01968626640461	52	0.00001211090890
21	0.01562500000000	53	0.00000961243477
22	0.01240157071850	54	0.00000762939453
23	0.00984313320230	55	0.00000605545445
24	0.00781250000000	56	0.00000480621738
25	0.00620078535925	57	0.00000381469727
26	0.00492156660115	58	0.00000302772723
27	0.00390625000000	59	0.00000240310869
28	0.00310039267963	60	0.00000190734863
29	0.00246078330058	61	0.00000151386361
30	0.00195312500000	62	0.00000120155435
31	0.00155019633981		

6.1.2 Bit Allocation and Quantization

The number of bits available to quantize audio samples in each frame is determined by the data rate less the bits needed to transmit the header, CRC, bit allocations, scale factors, and any ancillary data.

The number of bits that can be used to quantize sub-band samples in Layer I ranges between zero and 15, excluding the allocation of 1 bit because of the nature of the midtread quantizer (see also Chapter 2). The bit allocation for each sub-band of each channel is communicated through a four-bit code where, other than a code of zero representing zero bits, the value of the four-bit code is equal to one less than the number of allocated bits. For example, a code of 5 indicates six bits for that sub-band. A code value of 15 is forbidden. The number of bits used to transmit bit allocation information is equal to $4 \times 32 = 128$ bits for single channel mode and $4 \times 32 \times 2 = 256$ bits for stereo mode.

The bit allocation routine is an iterative process where, after initializing the process by setting all bit allocation codes to zero and assuming no bits are needed to transmit scale factors, in each iteration additional bits are allocated to the sub-band with the highest noise-to-mask ratio (NMR) until there are not enough additional bits left for the next iteration pass.

The NMR for each sub-band is calculated as the difference between the signal-to-mask ratio (calculated in the psychoacoustic model) and the signal-to-noise ratio (estimated from a table lookup based on the number of bits allocated to the sub-band). In each iteration, an additional quantization bit is allocated to the sub-band with the highest NMR that can still accept an additional bit (i.e. has less than 15 bits already). If a sub-band that has not already been given any bits is the one with the highest NMR, that sub-band is given two bits and bits are also set aside to transmit the scale factor for that sub-band.

Once the bit allocations are determined, each sub-band sample is divided by its scale factor and quantized using a midtread uniform quantizer having a number of steps determined by the bit allocation for that sub-band.

6.2 Layer II

The Layer II algorithm builds on the basic structure of Layer I (see *Figure 22*). The frame size is increased for Layer II to 3 granules of 12 sub-band samples corresponding to a total of $12 \times 3 \times 32 = 1152$ input samples per frame so that one can take advantage of commonality between consecutive granules. The audio encoded data in each frame represent 24 ms of data at a 48 kHz sampling rate.

6.2.1 Scale Factors and Scale Factors Select Information

The scale factors in Layer II can be shared among the three consecutive granules. The scale factors are computed in the same manner as in Layer I (see *Table 1*). When the values of the scale factors for the consecutive granules are sufficiently close or when temporal post-masking can hide distortion, only one or two scale factors need to be transmitted.

The scale factor select information, SCFSI (see mid portion of *Figure 22*), is coded with two bits and determines whether one, two, or three scale factors will be transmitted for the three consecutive granules in the frame. A SCFSI equal to zero corresponds to three scale factors, a SCFSI equal to one corresponds to transmitting two scale factors (the first for the first two granules and the second for the third granule), a SCFSI equal to two corresponds to transmitting a single scale factor to be used for all granules, and a SCFSI equal to three corresponds to transmitting two scale factors (where the first is used for the first granule and the second is used for the last two granules).

Both SCFSI and scale factors are transmitted only if the bit allocation is different from zero for that sub-band. The bits used for SCFSI vary between zero and $30 * 2 * 2 = 120$ bits for each frame and the bits used for scale factors vary between zero and $6 * 3 * 30 * 2 = 1080$ bits per frame. (Note: Layer II does not allow any bits to be allocated to the two highest sub-bands so there are at most only 30 sub-bands that are allocated bits. See [ISO/IEC 11172-3].)

6.2.2 Bit Allocation and Quantization

The psychoacoustic model and bit allocation process are computed in a similar manner for Layer II as for Layer I, but they are now computed relative to a frame of 36 sub-band samples. Also, bit allocation tables are now specified in the standard that determine the possible quantization levels that can be used to quantize the samples in any sub-band. As in Layer I, quantization is carried out by dividing each sub-band sample by its scale factor and then using a uniform mid-tread quantizer with the prescribed number of steps.

Bit allocations are coded based on tables specified in the standard that depend on the sample rate and the target data rate. Bit allocation codes range from zero to four bits per sub-band, where a greater number of bits are used to code the bit allocations of lower frequency sub-bands. In all cases, no bits are allocated to the two highest frequency sub-bands in Layer II. In other sub-bands, the number of quantization levels can reach that of 16-bit quantization. The total number of bits employed to describe the bit

allocation for one frame of the stereo signal varies between 26 and 188 depending on which table is applicable.

An example of a Layer II bit allocation table is shown in *Table 2*. This particular table is valid for sample rates of 44.1 and 48 kHz when they are running at data rates of 32 or 48 kb/s per channel. The first column represents the sub-band number and the second column represents the number of bits used to transmit that sub-band's bit allocation. The successive columns in each row represent the possible number of quantization levels that can be used to quantize the 36 samples in that frame's sub-band. The numbers in the top row indicate the bit allocation code corresponding to that number of quantization steps. Notice that four bits are used to encode the bit allocation for the first two sub-bands, three bits are used to encode the bit allocation for the next six sub-bands, and no bits are allocated to any higher frequency sub-bands in this particular bit allocation table. For this bit allocation table, the total number of bits employed to describe the bit allocation is equal to 26.

Table 2. Example of MPEG Audio Layer II bit allocation table [ISO/IEC 11172-3]

B	N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	4	-	3	5	9	15	31	63	127	255	511	1023	2047	4095	8191	16383	32767
1	4	-	3	5	9	15	31	63	127	255	511	1023	2047	4095	8191	16383	32767
2	3	-	3	5	9	15	31	63	127								
3	3	-	3	5	9	15	31	63	127								
4	3	-	3	5	9	15	31	63	127								
5	3	-	3	5	9	15	31	63	127								
6	3	-	3	5	9	15	31	63	127								
7	3	-	3	5	9	15	31	63	127								
8	0	-															
9	0	-															
10	0	-															
11	0	-															
12	0	-															
13	0	-															
14	0	-															
15	0	-															
16	0	-															
17	0	-															
18	0	-															
19	0	-															
20	0	-															
21	0	-															
22	0	-															
23	0	-															
24	0	-															
25	0	-															
26	0	-															

B	N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
27	0	-															
28	0	-															
29	0	-															
30	0	-															
31	0	-															

Notice also that the bit allocation table shown in *Table 2* has entries with numbers of levels not usually seen in midtread quantizers. Recall that an N -bit mid-tread quantizer usually has 2^N-1 levels. For example, a 4-bit quantizer would have 15 levels. In contrast, this table has cases with 5 and 9 levels. A 5-step midtread quantizer needs 3 bits to encode its values (2 bits can only count up to 4 levels) while a 9-step quantizer needs 4 bits to encode its values (3 bits can only count up to 8 levels). At first blush, it seems quite wasteful of bits to allow such quantizers, however, grouping of consecutive samples allows for quite efficient packing of the quantized data.

In the cases of 3, 5, and 9-step quantization, three consecutive samples are (vector) coded with one single code word, v , consisting of 5, 7, 10 bits respectively as follows:

$$\begin{aligned} v_3 &= 9z + 3y + x \\ v_5 &= 25z + 5y + x \\ v_9 &= 81z + 9y + x \end{aligned}$$

where x, y, z , are the quantization levels corresponding to three consecutive sub-band samples (counting from zero up to $N_{\text{levels}}-1$). The result of this vector coding is that quantization can be carried out with few quantizer levels without wasting too many bits.

This vector coding approach saves bits by allocating bits to enumerate possible triplets of coded values rather than allocating bits to each quantized value individually. For example, 3 consecutive samples using a 3-step quantizer would require $3*2 = 6$ bits but are here encoded using only 5 bits. This packing can be achieved since there are only $3*3*3 = 27$ possible code triplets for 3 samples – a number of possibilities that easily can be enumerated with a 5-bit code (which can enumerate up to 32 items). Similarly, 3 consecutive 5-step quantized values would require $3*3 = 9$ bits but the $5*5*5 = 125$ possible code triplets can be enumerated with a 7-bit code (which can enumerate up to 128 items). Finally, 3 consecutive 9-step quantized values could require $3*4 = 12$ bits while the $9*9*9 = 729$ possible triplets can be enumerated with a 10-bit code (which can enumerate up to 1024 items).

In general Layer II represents bit allocation, scale factors, and quantized samples in a more compact way than Layer I, so that more bits are available

to represent the coded audio data. In general, Layer II provides higher quality than Layer I at any given data rate.

6.3 Layer III

In addition to the differences in the filter bank and psychoacoustic model computation (see previous sections), Layer III core allocation/quantization routines are more complex. Moreover, a locally variable data rate is employed in Layer III to respond to the demand of difficult signals. The mechanism employed to dynamically provide additional bits is defined in the standard as a bit reservoir mechanism. The frame size for Layer III is 1152 samples as in Layer II, but the coded audio data may expand beyond the current frame, their starting point being identified by a backwards pointer in the bitstream.

6.3.1 Scale Factors

The basic Layer III algorithm (in the single channel mode) feeds the output of the hybrid filter bank to the quantization and coding stage (see also *Figure 4*). The spectrum is subdivided in 21 or 12 scale factor bands (rather than the 32 PQMF sub-bands as in Layers I and II) for long blocks and short block respectively, where the width of these bands loosely follows the critical bandwidth rate and is specified in look-up tables in the standard. In *Table 3* and *Table 4* an example of Layer III scale factors partition at 48 kHz for long and short blocks is shown.

Table 3. MPEG Audio Layer III scale factors partition for long blocks, $F_s = 48$ kHz [ISO/IEC 11172-3]

Scale factor band	Width	Start	End
0	4	0	3
1	4	4	7
2	4	8	11
3	4	12	15
4	4	16	19
5	4	20	23
6	6	24	29
7	6	30	35
8	6	36	41
9	8	42	49
10	10	50	59
11	12	60	71
12	16	72	87
13	18	88	105
14	22	106	127

Scale factor band	Width	Start	End
15	28	128	155
16	34	156	189
17	40	190	229
18	46	230	275
19	54	276	329
20	54	330	383

Table 4. MPEG Audio Layer III scale factors partition for short blocks, $F_s = 48$ kHz [ISO/IEC 11172-3]

Scale factor band	Width	Start	End
0	4	0	3
1	4	4	7
2	4	8	11
3	4	12	15
4	6	16	21
5	6	22	27
6	10	28	37
7	12	38	49
8	14	50	63
9	16	64	79
10	20	80	99
11	26	100	125

The scale factors in Layer III are employed to modify the quantization step size to ensure that the resulting quantization noise level falls below the computed masked threshold [Brandenburg 87 and Johnston 89], see also next section. Layer III scale factors differ from Layers I and II scale factors in that they are not the result of a normalization process. A set of scale factor select information is coded every two granules, where for Layer III a granule is defined as 576 frequency lines. The SCFSI in Layer III is represented by one bit for each sub-band, which is set to 0 if different scale factors are transmitted for each granule, or to 1 if the scale factor transmitted for the first granule is valid also for the second. If the coder operates in short block mode, then different scale factors are always transmitted for each granule. In addition to the sub-band scale factors, a global gain scale factor for the entire granule determines the overall quantization step size.

6.3.2 Non-Uniform Quantization and Huffman Coding

A non-uniform midtread quantizer is used in Layer III. The quantizer raises its input to the $3/4$ power before applying a midtread quantizer (see also Chapter 2). In the decoder before inverse quantizing, the quantized values are re-linearized by raising them to the $4/3$ power. In this fashion,

bigger values are quantized with less accuracy than smaller values, increasing the signal to noise ratio at low level input.

In addition, in Layer III static Huffman coding is also employed. The encoder subdivides the quantized spectral values into three regions and each region is coded with a different set of Huffman code tables tuned for the statistics of that region. At high frequencies, the encoder identifies a region of “all zeros”. The size of this region can be deduced from the size of the other two regions and does not need to be coded, the only restriction being that it must contain an even number of zeros since the other two regions group their values in even numbered sequences. The second region, called “count 1” region, contains a series of contiguous values consisting only of -1 , 0 , $+1$. In this region, four consecutive values are Huffman encoded at the same time and the size of this region is a multiple of four. Finally the third and last region, the “big values” region, covers the remaining spectral values which are encoded in pairs. This region is further subdivided in three parts each covered by a separate Huffman table. A set of 16 different Huffman code tables is utilized. For each partition, the Huffman table which best matches the signal statistics is selected during the encoder process. This dynamic search allows for both an increased coding efficiency and a decreased error sensitivity. The largest Huffman table carries 16 by 16 entries. Larger values are accommodated by using an escape mechanism.

The encoder iteratively quantizes the spectral values based on the data rate available, computes the number of Huffman codes needed to represent the quantized spectral values, and derives the distortion noise. If the distortion noise in certain scale factor bands is above the masked threshold values estimated in the psychoacoustic model, it varies the quantizer step size for each scale factor band by amplifying the relative scale factor. The quantization process is repeated until none of the scale factor bands have more than the allowed distortion, or the amplification for any band exceeds the maximum allowed value.

6.3.3 Bit Reservoir

The mechanism put in place in Layer III bitstream formatting routine allows for donating bits to the bit reservoir when the signal analyzed at a certain time requires less than the average number of bits to code a frame or utilizing bits from the reservoir for peak demands. In this fashion, the bitstream formatting routine is designed to support a locally-variable data rate. This mechanism not only allows to respond to local variations in the bit demand, but also is utilized to mitigate possible pre-echo effects. Although the number of bits employed to code a frame of audio data is no longer constant, its long term average is. The deviation from the target data rate is

always negative, i.e. the channel capacity is never exceeded. The maximum deviation from the target bit rate is fixed by the size of the maximum allowed delay in the decoder. This value is fixed in the standard by defining a code buffer limited to 7680 bits. For the data rate of 320 kb/s at 48 kHz sampling rate the average number of bits per frame is

$$\text{Number of bits per frame} = 1152 \cdot 320 / 48 = 7680$$

In this case, no variation is allowed. The first element in the Layer III audio data bitstream is a 9-bit pointer (`main_data_begin`) which indicates the location of the starting byte of the audio data for that frame. In *Figure 23* an example of the Layer III frame structure is shown.

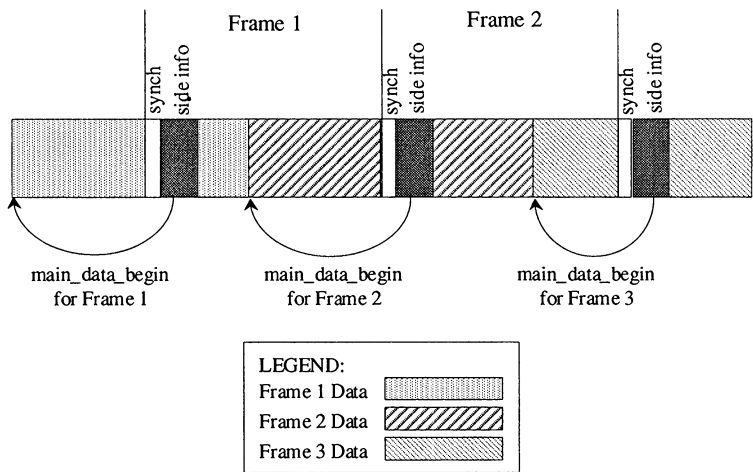


Figure 23. Example of Layer III frame structure from [ISO/IEC 11172-3]

7. STEREO CODING

Spatial redundancies and irrelevancies are exploited in joint stereo coding in order to further reduce the audio coding system data rates. While for most stereo signals there is little correlation between the time representation of the left and the right channels, typically for the signal power spectra strong correlations exist [Brandenburg 98]. Regarding stereo irrelevancies, we know that at high frequencies the ability of the human auditory system to discriminate the exact source location is decreased [Blauert 83]. Above about 2 kHz, for each critical band the human auditory system bases its

perception of stereo imaging from power maxima in space rather than the signal fine temporal structure.

In general, jointly coding the left and right channel provides higher coding gains. There are instances, however, in which stereo coding data rates may exceed twice the rate needed to transparently code one mono signal. In other words, the artifacts masked in single channel coding may become audible when presented as a stereo signal encoded as a dual mono. This effect is related to differences between the masked threshold recorded when the signal is presented as a single channel signal and the masked threshold under binaural conditions. This difference is called the binaural masking level difference (BMLD) [Blauert 83] and it is most pronounced at low frequencies. An example of this effect is related to “the cocktail party effect”, where a listener is capable to focus on a conversation in spite of the louder background noise. If the listener plugs one ear, suddenly the same level conversation becomes more difficult to understand. The conversation is less effectively masked when the subject listens to it binaurally than when the subject listens to it monaurally. Applying these findings to stereo coding, one may find that coding artifacts that are masked in single channel mode can become unmasked when presented simultaneously in two channels, i.e. in a dual mono system.

The basic idea in joint stereo coding is to appropriately rotate the stereo plane for each critical band into the main axis direction of the input stereo signal. Applying this idea in audio coding means transmitting side information to convey the direction parameters. This sometimes translates into a large amount of additional side information with no net coding gain. Simplified approaches can be found in literature such as mid/sum, M/S, and intensity stereo coding [Johnston and Ferreira 92 and van der Waal and Veldhuis 91]. While these approaches have some overlapping, their main focus is very different.

M/S stereo coding focuses mainly on redundancies removal and it is a transparent process in absence of quantization and coding. In M/S stereo coding only two directions are considered. Instead of transmitting separately the left and the right signal, the normalized sum and difference signals are transmitted. These signals are also referred to as the middle or mid (sum) and side (difference) signals. The coding gain achieved by utilizing M/S stereo coding is signal dependent. The maximum gain is reached when the left and right signals are equal or phase shifted by π . M/S stereo processing can be applied to the full signal spectrum since it is a lossless process and, in particular, it preserves the spatial attributes of the signal. In some cases, however, stereo irrelevancy effects can also be utilized in this approach.

Intensity stereo coding focuses mainly on irrelevancy removal, where at high frequencies the signal is coded with reduced spatial resolution. This

approach is based on the fact that the human auditory system detects direction at high frequency primarily based upon relative intensity in each ear rather than using phase cues. In intensity stereo coding, only one channel resulting from the combination of the right and left channel is transmitted for each critical band. The directional information is conveyed with independent scale factors for the left and the right channels. While the main spatial cues are retained in this method, some details may be missing. Intensity stereo coding preserves the energy of the stereo signal, but some of the signal components may not be properly transmitted, resulting in a potential loss of spatial information. In general, the loss of spatial information is considered less annoying than coding artifacts from bit starvation. For this reason, intensity stereo coding is employed mainly at low data rates. It is important to emphasize that intensity stereo coding is applied for high frequency only. Extending this approach to low frequencies may cause severe distortion such as a major loss in spatial information. Assuming that intensity coding is applied to half of the signal spectrum, a saving of about 20% in the system data rate can be achieved [Brandenburg 98]. Finally it should be mentioned that extensions to multichannel of the general intensity coding technique were developed in more recent years and different flavors are known as coupling channels [Davis 93], dynamic crosstalk [Stoll et al. 94] or generalized intensity coding (see also next chapter for more details).

7.1 MPEG Audio Stereo Coding

In MPEG-1 audio coding mid/sum, MS, stereo coding and intensity stereo coding techniques are applied. M/S stereo coding is employed in Layer III only, while intensity stereo coding is employed in all MPEG-1 Audio layers. In the joint stereo mode (bits 24 and 25 in the MPEG Audio header) the sum of the left and right sub-band samples is transmitted with scale factors for the left and the right channels for selected sub-bands. First an estimate of the required data rate for the left and right channels is performed. If the data rate exceeds the target data rate, a number of sub-bands are set to intensity stereo mode. Depending on the data rate, the allowed set of sub-bands for intensity coding is specified in the standard (bits 26 and 27 in the MPEG Audio header). For the quantization of the combined sub-bands, the higher of the bit allocation for the left and right channel is used. The combined sub-bands are coded in the same fashion as the independent sub-bands. In addition to the common scale factors, directional scale factors are unique to the sub-bands coded in the joint stereo mode and they are transmitted in the bitstream.

7.1.1 M/S Stereo Coding

M/S stereo coding is applied in Layer III when in joint stereo mode when:

$$\sum_{k=0}^{511} (l_k^2 - r_k^2) < 0.8 \sum_{k=0}^{511} (l_k^2 + r_k^2)$$

Where l_k and r_k correspond to the FFT spectral line amplitudes computed in the psychoacoustic model. The values M_i and S_i are transmitted, instead of the left and right channel values L_i and R_i , where L_i and R_i correspond to the hybrid filter bank spectral line amplitudes as follows:

$$M_i = \frac{R_i + L_i}{\sqrt{2}} \quad \text{and} \quad S_i = \frac{L_i - R_i}{\sqrt{2}}$$

In order to code the M/S signals, specially tuned thresholds are computed in the psychoacoustic model and utilized in the allocation routines.

8. SUMMARY

In this chapter we reviewed the main features of MPEG-1 Audio. MPEG-1 is the first international standard of high quality audio and opened the door to a variety of applications from digital audio broadcasting to internet distribution of music. Organized in three layers of increasing sophistication and complexity design, the basic building blocks consist of a time to frequency mapping stage, psychoacoustic model, quantization and coding and bitstream formatting. MPEG-1 allows for audio coding between 32 and 448 kb/s per channel targeting perceptually lossless quality. In the next chapters, we discuss further developments in MPEG Audio including MPEG-2 and MPEG-4.

9. REFERENCES

- [Blauert 83]: J. Blauert, *Spatial Hearing*, MIT Press, Cambridge, MA 1983.
- [Bosi, Todd and Holman 93]: M. Bosi, C. Todd and T. Holman, "Aspects of Current Standardization Activities for High-Quality, Low Rate Multichannel Audio Coding,"

Proc. of the 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2a.3, New Paltz, New York, October 1993.

[Brandenburg 87]: K. Brandenburg, "OCF-A New Coding Algorithm for High Quality Sound Signals", in Proc. ICASSP, pp. 141-144, May 1987.

[Brandenburg 98]: K. Brandenburg, "Perceptual Coding of High Quality Digital Audio", in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg (ed.), pp. 39-83, 1998.

[Brandenburg and Johnston 90]: K. Brandenburg and J. D. Johnston, "Second Generation Perceptual Audio Coding: The Hybrid Coder", presented at the 88th AES Convention, Preprint 2937, Montreux 1990.

[Brandenburg et al. 91]: K. Brandenburg, J. Herre, J. D. Johnston, Y. Mahieux, and E. F. Schroeder, "ASPEC-Adaptive Spectral Perceptual Entropy Coding of High Quality Music Signals", presented at the 90th AES Convention, Preprint 3011, May 1991.

[Chiariglione 95]: L. Chiariglione, "The development an integrated audio-visual coding standard", Proceedings of the IEEE, Vol. 83 n. 2, pp. 151-157, February 1995.

[Davis 93]: M. Davis, "The AC-3 Multichannel Coder", presented at the 95th AES Convention, Preprint 3774 New York, October 1993.

[Dehéry, Stoll and Kerkhof 91]: Y. F. Dehéry, G. Stoll and L. v.d. Kerkhof, "MUSICAM Source Coding for Digital Sound", In Symp. Rec. Broadcast Sessions of the 17th Int. Television Symp., pp. 612-617, Montreux Switzerland, June 1991.

[Edler 92]: B. Edler, "Aliasing reduction in sub-bands of cascaded filter banks with decimation ", Electronics Letters, Vol. 28, pp. 1104-1105, 1992.

[Edler 89]: B. Edler, "Coding of Audio Signals with Overlapping Transform and Adaptive Window Shape" (in German), Frequenz, Vol. 43, No. 9, pp. 252-256, September 1989.

[ETS 300 401 v2]: The European Telecommunications Standards Institute (ETSI), ETS 300 401 v2, "Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to Mobile, Portable and Fixed Receivers", 1995-1997.

[ETS 300 421]: The European Telecommunications Standards Institute (ETSI), ETS 300 421, "Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for 11/12 GHz Satellite Services", August 1997.

[ETS 300 429]: The European Telecommunications Standards Institute (ETSI), ETS 300 429, "Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Cable Systems", August 1997.

[ETS 300 744]: The European Telecommunications Standards Institute (ETSI), ETS 300 744, "Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television", August 1997.

[Fielder et al. 96]: L. D. Fielder, M. Bosi, G. A. Davidson, M. Davis, C. Todd, and S. Vernon "AC-2 and AC-3: Low Complexity Transform-Based Audio Coding," in N. Gielchrist and C. Grewin (ed.), *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 54-72, AES 1996.

[ISO/IEC 11172]: ISO/IEC 11172, Information Technology, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s", 1993.

[ISO/IEC 11172-3]: ISO/IEC 11172-3, Information Technology, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 3: Audio", 1993.

[ISO/IEC 13818]: ISO/IEC 13818, Information Technology, "Generic coding of moving pictures and associated audio", 1994-1997.

[ISO/IEC 13818-7]: ISO/IEC 13818-7, Information Technology, "Generic coding of moving pictures and associated audio, Part 7: Advanced Audio Coding", 1997.

[ISO/IEC 14496]: ISO/IEC 14496, Information Technology, "Coding of audio-visual objects", 1999-2001.

[ISO/IEC 15938]: ISO/IEC 15938, Information Technology, "Multimedia content description interface", 2001.

[ISO/IEC MPEG 91/010]: ISO/IEC JTC 1/SC 29/WG 11 MPEG 91/010, "The MPEG/AUDIO Subjective Listening Test" Stockholm, April/May 1991.

[ISO/IEC MPEG N1229]: ISO/IEC JTC 1/SC 29/WG 11 N1229, "MPEG-2 Backwards Compatible CODECS Layer II and III: RACE dTTb Listening Test Report" Florence, March 1996.

[ISO/IEC MPEG N1420]: ISO/IEC JTC 1/SC 29/WG 11 N1420, "Overview of the Report on the Formal Subjective Listening Tests of MPEG-2 AAC multichannel audio coding" Maceio, November 1996.

[ISO/IEC MPEG N4318]: ISO/IEC JTC 1/SC 29/WG 11 N4318, "MPEG-21 Overview" Sydney, July 2001.

[ISO/IEC MPEG N271]: ISO/IEC JTC 1/SC 29/WG 11 N271, "New Work Item Proposal for Very-Low Bitrates Audiovisual Coding" London, November 1992.

[ITU-R BS.1115]: International Telecommunication Union BS.1115, "Low Bitrate Audio Coding", Geneva, 1994.

[Johnston 88]: J. D. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria", in Proc. ICASSP, pp. 2524-2527, May 1988.

[Johnston 89]: J. D. Johnston, "Transform Coding of Audio Using Perceptual Noise Criteria", in IEEE J. Select. Areas Common., vol. 6, pp. 314-323, 1989.

[Johnston and Ferreira 92]: J. D. Johnston and A. J. Ferreira, "Sum-Difference Stereo Transform Coding", in Proc. ICASSP pp. 569-571, May 1992.

[MPEG]: MPEG Home Page, <http://mpeg.telecomitalia.com/>.

[Pan 95]: D. Pan, "A Tutorial on MPEG/Audio Compression", in IEEE Multimedia, pp. 60-74, Summer 1995.

[Searing 91]: S. Searing, "Suggested Formulas for Audio Analysis and Synthesis Windows", ISO/IEC JTC 1/SC 29/WG 11 MPEG 91/328, November 1991.

[Sinha, Johnston, Dorward and Quackenbush 98]: D. Sinha, J. D. Johnston, S. Dorward and S. R. Quackenbush, "The perceptual Audio Coder (PAC)", in *The Digital Signal Processing Handbook*, V. Madiseti and D. Williams (ed.), CRC Press, pp. 42.1-42.18, 1998.

[Stoll et al. 94]: G. Stoll, G. Theile, S. Nielsen, A. Silzle, M. Link, R. Sedlmayer and A. Breford, "Extension of ISO/MPEG-Audio Layer II to Multi-Channel Coding: The Future Standard for Broadcasting, Telecommunication, and Multimedia Applications", presented at the 94th AES Convention, Preprint 3550, Berlin 1994.

[Thiede et al. 00]: T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg and B. Feiten., "PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality", J. Audio Eng. Soc., Vol. 48, no. 1/2, pp. 3-29, January/February 2000.