Analysis of the MPEG-1 Layer III (MP3) Algorithm Using MATLAB

Synthesis Lectures on Algorithms and Software in Engineering

Editor
Andreas Spanias, ***Arizona State University***

Analysis of the MPEG-1Layer III (MP3) Algorithm Using MATLAB
Jayaraman J. Thiagarajan and Andreas Spanias
2011

Theory and Applications of Gaussian Quadrature Methods
Narayan Kovvali
2011

Algorithms and Software for Predictive and Perceptual Modeling of Speech
Venkatraman Atti
2011

Adaptive High-Resolution Sensor Waveform Design for Tracking
Ioannis Kyriakides, Darryl Morrell, and Antonia Papandreou-Suppappola
2010

MATLAB® Software for the Code Excited Linear Prediction Algorithm: The Federal Standard-1016
Karthikeyan N. Ramamurthy and Andreas S. Spanias
2010

OFDM Systems for Wireless Communications
Adarsh B. Narasimhamurthy, Mahesh K. Banavar, and Cihan Tepedelenlioglu
2010

Advances in Modern Blind Signal Separation Algorithms: Theory and Applications
Kostas Kokkinakis and Philipos C. Loizou
2010

Advances in Waveform-Agile Sensing for Tracking
Sandeep Prasad Sira, Antonia Papandreou-Suppappola, and Darryl Morrell
2008

Despeckle Filtering Algorithms and Software for Ultrasound Imaging
Christos P. Loizou and Constantinos S. Pattichis
2008

Analysis of the MPEG-1Layer III (MP3) Algorithm Using MATLAB

Jayaraman J. Thiagarajan and Andreas Spanias

www.morganclaypool.com

ISBN: 9781608458011 paperback
ISBN: 9781608458028 ebook

DOI 10.2200/S00382ED1V01Y201110ASE009

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON ALGORITHMS AND SOFTWARE IN ENGINEERING

Lecture #9
Series Editor: Andreas Spanias, *Arizona State University*
Series ISSN
Synthesis Lectures on Algorithms and Software in Engineering
Print 1938-1727 Electronic 1938-1735

Analysis of the MPEG-1 Layer III (MP3) Algorithm Using MATLAB

Jayaraman J. Thiagarajan and Andreas Spanias
Arizona State University

SYNTHESIS LECTURES ON ALGORITHMS AND SOFTWARE IN ENGINEERING #9

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT
The MPEG-1 Layer III (MP3) algorithm is one of the most successful audio The MPEG-1 Layer III (MP3) algorithm is one of the most successful audio formats for consumer audio storage and for transfer and playback of music on digital audio players. The MP3 compression standard along with the AAC (Advanced Audio Coding) algorithm are associated with the most successful music players of the last decade. This book describes the fundamentals and the MATLAB implementation details of the MP3 algorithm. Several of the tedious processes in MP3 are supported by demonstrations using MATLAB® software. The book presents the theoretical concepts and algorithms used in the MP3 standard. The implementation details and simulations with MATLAB® complement the theoretical principles. The extensive list of references enables the reader to perform a more detailed study on specific aspects of the algorithm and gain exposure to advancements in perceptual coding.

KEYWORDS
perceptual audio coders, lossy coding, MPEG standards, psychoacustic models, MP3 algorithm, variable-bit-rate coding

Contents

Preface

The MPEG-1 audio Layer-3, popularly referred to as the MP3, is a digital encoding format based on perceptual audio compression. This audio-specific format, designed by the Moving Pictures Expert Group, was standardized as a part of the generic MPEG-1 standard in the early 1990's. MP3 has found widespread application in consumer audio storage and in recording and playback of music on digital audio players. The MP3 compression standard along with the AAC (Advanced Audio Coding) standard is associated with perhaps the most successful music players of the last decade.

This book presents MATLAB® software that implements several important functions of the MPEG-1 Layer 3 encoding and decoding algorithms. We describe the fundamentals and implementation details of the algorithm along with several MATLAB® demonstrations. In the First chapter, we present a brief discussion on the history of audio coders and describe the architectural overview of perceptual coders. Furthermore, we describe the principles of psychoacoustics and provide details of the *psychoacoustic model I* used in the earlier MPEG audio coders. From Chapter 2 and on, we discuss the various modules of the MP3 algorithm. The theory sections in most chapters provide description of the necessary concepts required to understand the algorithm. The implementation details and simulations with MATLAB® functions complement these descriptions. The extensive list of references will enable the reader to perform a more detailed study on specific aspects of the algorithm and gain exposure to some of the recent advancements in perceptual coding. Finally, a detailed analysis of the computational complexity of the MP3 algorithm, for both the C and MATLAB implementations, is presented

Audio coders, in addition to being an interesting application of signal processing principles for students, present a valuable resource to both practitioners and researchers in several other sound related implementations. In particular, the functions and software provided in this book will enable practitioners and algorithm developers to understand and optimize/modify sections of the algorithm in order to achieve improved performance and design computationally efficient implementations. Furthermore, understanding the principles of the MP3 algorithm will enable the reader to understand and analyze several of the more recent perceptual coders. Students will be able to visualize and understand the various modules of MP3 by using the software and the associated simulations.

Jayaraman J. Thiagarajan and Andreas Spanias
October 2011

Acknowledgments

The authors acknowledge Ramapriya Rangachar, a Masters graduate student that worked in our lab, who was involved in developing some of the MATLAB® code for the MP3 algorithm, and ISO for developing the C code for the algorithm. We note that the MATLAB® code presented in this book is organized in modules by the authors in order to facilitate the learning of the MP3 functions. In fact, several functions have been developed directly from the original standard documentation. The authors also like to thank Ted Painter and Venkatraman Atti for providing descriptions of some of the foundations of psychoacoustics and audio coding. In addition, the first author acknowledges Karthikeyan N. Ramamurthy, Harini Sridhar and Vivek Iyer for their support. Portions of this work and associated software have been supported by the SenSIP Center and NSF grants 0817596, 1035086.

Jayaraman J. Thiagarajan and Andreas Spanias
October 2011

CHAPTER 1

Introduction

Audio coding or *audio compression* algorithms obtain reduced bit rate representations of high-fidelity audio signals and have applications in transmission, storage, streaming and broadcasting. The objective of audio coding algorithms is to represent the signal with a small number of bits while maintaining its perceptual quality such that it is indistinguishable from the original. The compact disk (CD) introduced true high-fidelity at high data rates. Conventional digital audio signals are associated with sampling frequencies of either 44.1 (CD) or 48 kHz (DAT) and the samples are encoded with pulse code modulation (PCM) at 16-bits per sample. This results in very high data rates amounting to 1.41 Mbits/s for a stereo-pair sampled at 44.1 kHz. Motivated by the need for compression algorithms for network and portable applications several codecs have been established. The focus of this book is to present a detailed analysis of several aspects of the MPEG-1 Layer-III audio coding standard. To facilitate the understanding of MP3, the theoretical concepts discussed in the following chapters are accompanied by various simulation examples. Furthermore, this book emphasizes the implementation specifics of the MP3 codec by including MATLAB code snippets and a detailed complexity analysis of the encoder and decoder functions.

1.1 A BRIEF HISTORY OF AUDIO CODERS

Digital audio compression relies on sophisticated time-frequency analysis techniques that use transform coders, filterbanks or hybrid signal-adaptive coding techniques. Audio coding relies heavily on exploiting properties of psychoacoustics [1, 2, 3, 4]. Foundations of audio coding using filter banks have been established both in the time-domain [5, 6, 7, 8, 9, 10, 11, 12, 13] and the frequency domain [14, 15]. Furthermore, Discrete Wavelet Transform (DWT) based subband coders [16, 17, 18] have also been considered for audio coding due to the flexibility in the choice of filter coefficients. The characterization of the auditory filterbank in terms of critical bands [2] has been used in audio compression and over the years several filterbanks [19, 20, 21, 22, 23] have been proposed to mimic the critical band structure. Furthermore, several quantization strategies have also been applied to transform-coders that use the discrete cosine transform (DCT) [24] and the modified DCT (MDCT) [25].

During the early nineties, several workgroups and organizations such as the ISO/IEC and the International Telecommunications Union (ITU) became actively involved in developing perceptual audio coding standards. The Masking pattern adapted Universal Subband Integrated Coding and Multiplexing (MUSICAM) algorithm [20, 26] had an initial influence on the ISO/IEC (International Organization for Standardization/International Electro-technical Commission) MPEG

(Moving Pictures Experts Group) audio standards, i.e., MPEG-1 [27] and the MPEG-2 [28]. Furthermore, several successful commercial audio standards have been published including Sony's Adaptive TRansform Acoustic Coding (ATRAC), DTS Coherent Acoustics (DTS-CA) and Dolby's Audio Coder-3 (AC-3). Elements or entire algorithms for perceptual coding have also appeared in [21, 23], [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53]. With the emergence of surround sound systems, multi-channel encoding formats also gained interest [54, 55, 56]. The advent of ISO/IEC MPEG-4 standardization [45, 47] established new research goals for high-quality coding of general audio signals even at low bit rates. MPEG-4 audio encompasses an integrated family of algorithms with wide ranging provisions for scalable, object-based speech and audio coding at bit rates from 200 bps up to 64 kbps per channel [57, 58].

1.1.1 RECENT AUDIO CODECS

The older MPEG-1 hybrid audio coding technique (ISO/IEC 11172-3) incorporates subband filter bank decomposition, signal transforms such as the FFT and psychoacoustic analysis. MPEG-1 audio operates on 16-bit PCM input audio data and accommodates sample rates of 32, 44.1, and 48 kHz. Operating modes of this algorithm include mono, stereo, dual independent mono, and joint stereo. The target bit rates are programmable in the range of 32-192 kbits/s for mono and 64-384 kbits/s for stereo. Despite the fact that MPEG-1 Layer-III (MP3) is still an active and popular standard, several new algorithms have been shown to perform better. Advanced Audio Coding (AAC) is a standardized, lossy compression scheme that generally achieves better sound quality than MP3 at similar bit rates. It has been standardized by the ISO and IEC as part of the MPEG-2 and MPEG-4 standards. Designed as a successor to the MP3 algorithm, AAC allows more sampling frequencies (8 kHz to 96 kHz) and supports up to 48 channels.

Though perceptual audio coders such as the MP3 and AAC offer reasonably good quality at bit rates down to 80 kbps, they are associated with an algorithmic delay that exceeds 120 ms. Applications such as two-way communications or broadcasting require low end-to-end delays of the order of 20 ms. As a result, Low Delay (LD) audio coding schemes have been developed and they provide comparable perceptual quality to MP3 or AAC with a very low algorithmic delay. The MPEG-4 AAC audio coder is used as a basis to build the low delay functionality preferable in end-to-end applications such as teleconferencing and telephony. Typical bit rates of AAC-LD start at 32 kbps for a mono signal with 22 kHz sampling rate and reach 128 kbps providing excellent audio quality [59]. AAC-ELD (Enhanced Low Delay) was standardized as part of MPEG in January 2008. AAC-ELD has an algorithmic delay of 32 ms at 24 kbps down to 15 ms at 64 kbps. AAC-ELD combines the advantages of AAC-LD for low encoding/decoding purposes and Spectral Band Replication (SBR) for preserving high quality at low bit rates. Delay critical applications such as wideband audio/video conferencing, broadcasting which require high quality audio at low bit rates can benefit from this scheme [60]. The Ultra Low Delay (ULD) AAC [61] was developed at Fraunhofer and attains delays of the order of 8 ms.

The need for an interface to exchange multimedia content through the internet resulted in the development of the MPEG-7 audio standard [48]. MPEG-7 supports a broad range of applications [62] that include the following: multimedia indexing/searching, multimedia editing, broadcast media selection, and multimedia digital library sorting. Issues such as the "interoperability" and multimedia resource delivery over a wide range of networks and terminals motivated the MPEG-21 Framework [53].

As mentioned earlier, Adaptive Transform Acoustic Coding (ATRAC) is a family of audio compression algorithms developed by Sony. Though the initial versions of ATRAC were used with the MiniDisc in the early 1990s, today the recent advanced ATRAC algorithms are used in several Sony-branded audio players, the Real Audio 8 and the native audio compression format for audio rendering in PS3 [63]. The MPEG-4 parametric audio codec, called Harmonic and Individual Lines plus Noise (HILN), enables coding of general audio signals at bitrates as low as 4 kbit/s using a parametric representation [64]. The encoder assumes that the audio signals can be synthesized using only sinusoids and noise. The input signal is decomposed into components based on appropriate source models and represented by model parameters. This approach utilizes more advanced source modeling than just assuming a stationary signal for the duration of a frame.

The launch of storage formats (in 1999) such as the DVD-Audio and the Super Audio CD (SACD) provided the audio codec designers with enormous storage capacity. This motivated an effort for *lossless* coding of digital audio [46, 51, 65]. A lossless audio coding system is able to reconstruct perfectly a "bit-for-bit representation" of the original input audio from the coded bitstream. In contrast, a coding scheme incapable of perfect reconstruction from the coded representation is called *lossy*. Several commercially successful lossless codecs have been developed in the last decade. Some of the earliest lossless audio coders include the Apple Lossless Audio Codec (ALAC) [66] and the Windows Media Audio 9 (WMA 9) lossless codec [13]. ALAC is an audio codec developed by Apple Inc. for lossless data compression of digital music. Typically, it is stored within an MP4 container with the filename extension .m4a. Though this extension is also used by AAC, ALAC employs linear prediction similar to other lossless codecs. All current iPod and iPhone devices can play Apple Lossless-encoded files. The WMA 9 lossless codec was released by Microsoft in early 2003 and it supports up to 96 kHz, 24-bit, 5.1 discrete channels with full dynamic range compression control. It can compress this multichannel signal audio CD at bit rates of 470 to 940 kbit/s.

Dolby TrueHD is an advanced lossless multi-channel audio codec developed by Dolby Laboratories [67]. It is primarily intended for high-definition home-entertainment equipment such as the Blu-ray Disc and the HD DVD. Though Dolby TrueHD is based on Meridian Lossless Packing (MLP) [46], it is significantly different from DVD-Audio. This variable bit-rate codec can support up to 14 discrete sound channels in its bitstream. Another important audio codec is DTS-HD Master Audio, developed by Digital Theater System [11]. It is an optional audio format for the Blu-ray Disc format exclusively. This format aims to allow a bit-to-bit representation of the original movie's studio master soundtrack. To accomplish this, DTS-HD MA supports variable bit rates up to 24.5 Mbit/s on a Blu-ray Disc and up to 18.0 Mbit/s for HD DVD. The DTS-HD Master Audio contains 2 data streams: the original DTS core stream and the residual stream which contains the difference between the original signal and the lossy compression DTS core stream [68]. The residual data is then encoded by a lossless encoder and packed together with the core. The most recent version of the Real player also supports lossless coding. The RealAudio lossless codec is designed primarily for high-quality music downloads in mono or two-channel stereo format (multichannel output is not supported). It replicates CD-quality sound in a format that takes less time for the user to download. Although the lossless audio codec is designed for high-fidelity music downloads, it can also be used for broadcasts in high-bandwidth environments.

The MPEG-4 Audio Lossless Coding, also referred as MPEG-4 ALS [12], extends the MPEG-4 Part 3 audio standard to perform lossless audio compression. It comprises of a short-term predictor, which is a quantized LPC predictor with a lossless residual, and a long term predictor modeled by 5 long-term weighted residues, each with its own delay. The long term predictor improves the compression for sounds with rich harmonics found in several musical instruments and human voice.

1.2 A GENERAL PERCEPTUAL AUDIO CODING ARCHITECTURE

It is important to note the architectural similarities that characterize most perceptual audio coders before we describe the MP3 audio codec in the following chapters. Over the last few years, researchers have proposed several efficient signal models and compression standards/methodologies for high-quality digital audio reproduction. Most of these algorithms are based on the generic architecture shown in Figure 1.1. Most coders typically segment input signals into quasi-stationary frames ranging from 2 to 50 ms in duration. This is followed by a time-frequency analysis to estimate the temporal and spectral components of each frame. Often, the time-frequency mapping is matched to the analysis properties of the human auditory system, although this is not always the case. The objective is to extract a set of time-frequency parameters that can be efficiently coded based on perceptual criteria. The time-frequency analysis module can typically comprise of time-invariant or time-varying filterbanks, harmonic analyzers and hybrid transforms.

The choice of time-frequency analysis methodology always involves a fundamental tradeoff between time and frequency resolution requirements. The time-frequency analysis module employed in the MPEG-1 codec is described in Chapter 2 and the strategies to handle the different resolution requirements are presented in Chapter 4. Perceptual distortion control is achieved by a psychoacoustic signal analysis module that estimates signal masking power based on psychoacoustic principles. The psychoacoustic model quantifies the maximum amount of distortion at each point in the time-frequency plane such that quantization of the time-frequency parameters does not introduce audible artifacts. The steps involved in the estimation of the masking thresholds using the psychoacoustic model – II are explained in Section 3.2 of this book. The quantization and encoding module can also exploit statistical redundancies through classical techniques such as DPCM or ADPCM. The redundancies in the quantized parameters can be removed using run-length and entropy coding

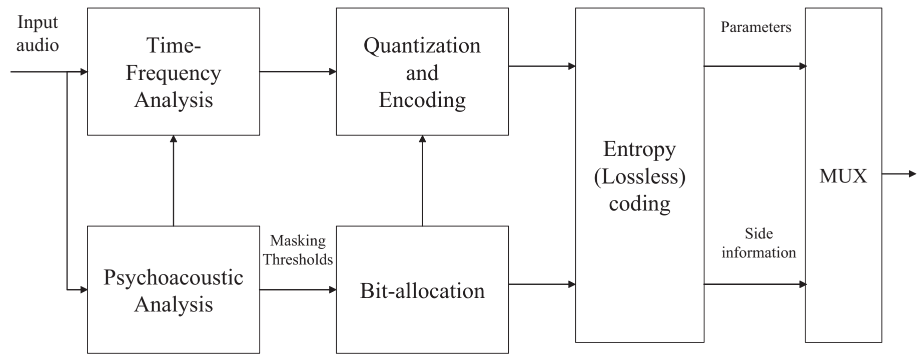

Figure 1.1: A generic block diagram of a perceptual audio encoder.

strategies [69, 70, 71]. Since the psychoacoustic module is signal dependent, most audio coding algorithms are variable rate. However, fixed channel rates can be achieved by efficient management of bit allocation using buffer feedback schemes. The coding methodology and the bit management techniques employed in the MP3 algorithm to achieve a fixed average bit rate are discussed in Sections 5.4 and 5.5, respectively.

1.3 PRINCIPLES OF PSYCHOACOUSTICS

Audio coding algorithms rely on generalized models of human hearing to optimize coding efficiency. The receiver is ultimately the human ear, and sound perception is affected by its masking properties. The field of psychoacoustics has made significant progress toward characterizing the time-frequency analysis capabilities of the inner ear. This, in turn, enabled audio coders to achieve compression by exploiting "irrelevant" information that is not detectable by even a trained listener. Irrelevant information is identified by incorporating psychoacoustic principles in quantization rules, including critical band frequency analysis and masking. The psychoacoustic model described in Chapter 3 relies on the principles discussed in this section. The *Sound Pressure Level* (SPL) is a standard metric that quantifies the intensity of an acoustic stimulus. The SPL provides the level (intensity) of sound pressure in decibels (dB) relative to an internationally defined reference level, i.e., $L_{SPL} = 20\log_{10}(p/p_0)$, where L_{SPL} is the SPL of a stimulus, p is the sound pressure of the stimulus in Pascals (Pa - equivalent to Newton/m^2), and p_0 is the standard reference level of $20\mu Pa$. Loosely speaking, about 150 dB SPL spans the dynamic range of the auditory system; an SPL reference of a quiet environment is around 0 dB SPL while a stimulus of 140 dB SPL approaches the threshold of pain. The absolute threshold of hearing shown in Figure 1.2 characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment.

The curve for the absolute threshold of hearing alone cannot be used for audio coding. Typically, music records require spectrally complex quantization rules and hence one has to modify the

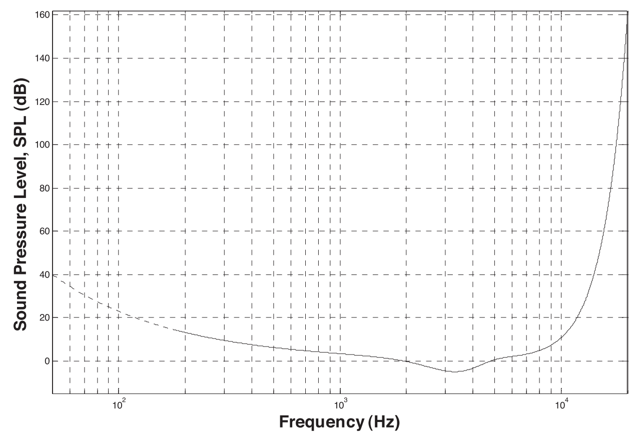

Figure 1.2: The absolute threshold of hearing in a noiseless environment.

absolute threshold in a dynamic manner. In order to estimate a time-varying threshold, one must use models for human hearing that take into account how the human ear performs spectral analysis. A frequency-to-place transformation takes place in the cochlea (inner ear), along the basilar membrane [72]. The loudness (perceived intensity) remains constant for a narrowband noise source presented at a constant SPL even as the noise bandwidth is increased up to the critical bandwidth. For any increase beyond the critical bandwidth, the loudness begins to increase. Critical bandwidth tends to remain constant (about 100 Hz) up to 500 Hz, and increases to approximately 20% of the center frequency above 500 Hz. The width of a critical band is commonly referred to as one *Bark*. The nonlinear function,

$$H_z(f) = 1.3 \arctan(0.00076 f) + 3.5 \arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{ (Bark)} \tag{1.1}$$

is often used to convert frequency from the Hertz to the Bark scale. Table 1.1 shows the idealized critical band filter bank in terms of band edges and center frequencies for a collection of 26 critical bandwidth auditory filters that span the audio spectrum. The frequency resolution of the auditory filter bank largely determines which portions of a signal are perceptually irrelevant. The auditory time-frequency analysis that occurs in the critical band filter bank induces simultaneous and non-simultaneous masking phenomena that are routinely used by modern audio coders to shape the coding distortion spectrum. As we will discuss in Section 5.1. the perceptual models allocate bits for signal components such that the quantization noise is shaped to exploit the *masking* thresholds for a complex sound.

Masking refers to a process where one sound is rendered inaudible because of the presence of another sound. In audio coding typically we distinguish between only three types of simultaneous