
AUDIO SIGNAL PROCESSING AND CODING

**Andreas Spanias
Ted Painter
Venkatraman Atti**



WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

AUDIO SIGNAL PROCESSING AND CODING



THE WILEY BICENTENNIAL—KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

A handwritten signature in black ink that reads "William J. Pesce".

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

A handwritten signature in black ink that reads "Peter Booth Wiley".

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

AUDIO SIGNAL PROCESSING AND CODING

**Andreas Spanias
Ted Painter
Venkatraman Atti**



WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Spanias, Andreas.

Audio signal processing and coding/by Andreas Spanias, Ted Painter, Venkatraman Atti.

p. cm.

“Wiley-Interscience publication.”

Includes bibliographical references and index.

ISBN: 978-0-471-79147-8

1. Coding theory. 2. Signal processing—Digital techniques. 3. Sound—Recording and reproducing—Digital techniques. I. Painter, Ted, 1967-II. Atti, Venkatraman, 1978-III. Title.

TK5102.92.S73 2006
621.382'8—dc22

2006040507

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

*To
Photini, John and Louis
Lizzy, Katie and Lee
Srinivasan, Sudha, Kavitha, Satis and Ammu*

CONTENTS

PREFACE	xv
1 INTRODUCTION	1
1.1 Historical Perspective	1
1.2 A General Perceptual Audio Coding Architecture	4
1.3 Audio Coder Attributes	5
1.3.1 Audio Quality	6
1.3.2 Bit Rates	6
1.3.3 Complexity	6
1.3.4 Codec Delay	7
1.3.5 Error Robustness	7
1.4 Types of Audio Coders – An Overview	7
1.5 Organization of the Book	8
1.6 Notational Conventions	9
Problems	11
Computer Exercises	11
2 SIGNAL PROCESSING ESSENTIALS	13
2.1 Introduction	13
2.2 Spectra of Analog Signals	13
2.3 Review of Convolution and Filtering	16
2.4 Uniform Sampling	17
2.5 Discrete-Time Signal Processing	20

2.5.1	Transforms for Discrete-Time Signals	20
2.5.2	The Discrete and the Fast Fourier Transform	22
2.5.3	The Discrete Cosine Transform	23
2.5.4	The Short-Time Fourier Transform	23
2.6	Difference Equations and Digital Filters	25
2.7	The Transfer and the Frequency Response Functions	27
2.7.1	Poles, Zeros, and Frequency Response	29
2.7.2	Examples of Digital Filters for Audio Applications	30
2.8	Review of Multirate Signal Processing	33
2.8.1	Down-sampling by an Integer	33
2.8.2	Up-sampling by an Integer	35
2.8.3	Sampling Rate Changes by Noninteger Factors	36
2.8.4	Quadrature Mirror Filter Banks	36
2.9	Discrete-Time Random Signals	39
2.9.1	Random Signals Processed by LTI Digital Filters	42
2.9.2	Autocorrelation Estimation from Finite-Length Data	44
2.10	Summary	44
	Problems	45
	Computer Exercises	47

3 QUANTIZATION AND ENTROPY CODING 51

3.1	Introduction	51
3.1.1	The Quantization–Bit Allocation–Entropy Coding Module	52
3.2	Density Functions and Quantization	53
3.3	Scalar Quantization	54
3.3.1	Uniform Quantization	54
3.3.2	Nonuniform Quantization	57
3.3.3	Differential PCM	59
3.4	Vector Quantization	62
3.4.1	Structured VQ	64
3.4.2	Split-VQ	67
3.4.3	Conjugate-Structure VQ	69
3.5	Bit-Allocation Algorithms	70
3.6	Entropy Coding	74
3.6.1	Huffman Coding	77
3.6.2	Rice Coding	81
3.6.3	Golomb Coding	82

3.6.4	Arithmetic Coding	83
3.7	Summary	85
	Problems	85
	Computer Exercises	86
4	LINEAR PREDICTION IN NARROWBAND AND WIDEBAND CODING	91
4.1	Introduction	91
4.2	LP-Based Source-System Modeling for Speech	92
4.3	Short-Term Linear Prediction	94
4.3.1	Long-Term Prediction	95
4.3.2	ADPCM Using Linear Prediction	96
4.4	Open-Loop Analysis-Synthesis Linear Prediction	96
4.5	Analysis-by-Synthesis Linear Prediction	97
4.5.1	Code-Excited Linear Prediction Algorithms	100
4.6	Linear Prediction in Wideband Coding	102
4.6.1	Wideband Speech Coding	102
4.6.2	Wideband Audio Coding	104
4.7	Summary	106
	Problems	107
	Computer Exercises	108
5	PSYCHOACOUSTIC PRINCIPLES	113
5.1	Introduction	113
5.2	Absolute Threshold of Hearing	114
5.3	Critical Bands	115
5.4	Simultaneous Masking, Masking Asymmetry, and the Spread of Masking	120
5.4.1	Noise-Masking-Tone	123
5.4.2	Tone-Masking-Noise	124
5.4.3	Noise-Masking-Noise	124
5.4.4	Asymmetry of Masking	124
5.4.5	The Spread of Masking	125
5.5	Nonsimultaneous Masking	127
5.6	Perceptual Entropy	128
5.7	Example Codec Perceptual Model: ISO/IEC 11172-3 (MPEG - 1) Psychoacoustic Model 1	130
5.7.1	Step 1: Spectral Analysis and SPL Normalization	131

5.7.2	Step 2: Identification of Tonal and Noise Maskers	131
5.7.3	Step 3: Decimation and Reorganization of Maskers	135
5.7.4	Step 4: Calculation of Individual Masking Thresholds	136
5.7.5	Step 5: Calculation of Global Masking Thresholds	138
5.8	Perceptual Bit Allocation	138
5.9	Summary	140
	Problems	140
	Computer Exercises	141
6	TIME-FREQUENCY ANALYSIS: FILTER BANKS AND TRANSFORMS	145
6.1	Introduction	145
6.2	Analysis-Synthesis Framework for M -band Filter Banks	146
6.3	Filter Banks for Audio Coding: Design Considerations	148
6.3.1	The Role of Time-Frequency Resolution in Masking Power Estimation	149
6.3.2	The Role of Frequency Resolution in Perceptual Bit Allocation	149
6.3.3	The Role of Time Resolution in Perceptual Bit Allocation	150
6.4	Quadrature Mirror and Conjugate Quadrature Filters	155
6.5	Tree-Structured QMF and CQF M -band Banks	156
6.6	Cosine Modulated “Pseudo QMF” M -band Banks	160
6.7	Cosine Modulated Perfect Reconstruction (PR) M -band Banks and the Modified Discrete Cosine Transform (MDCT)	163
6.7.1	Forward and Inverse MDCT	165
6.7.2	MDCT Window Design	165
6.7.3	Example MDCT Windows (Prototype FIR Filters)	167
6.8	Discrete Fourier and Discrete Cosine Transform	178
6.9	Pre-echo Distortion	180
6.10	Pre-echo Control Strategies	182
6.10.1	Bit Reservoir	182
6.10.2	Window Switching	182
6.10.3	Hybrid, Switched Filter Banks	184
6.10.4	Gain Modification	185
6.10.5	Temporal Noise Shaping	185
6.11	Summary	186
	Problems	188
	Computer Exercises	191

7 TRANSFORM CODERS	195
7.1 Introduction	195
7.2 Optimum Coding in the Frequency Domain	196
7.3 Perceptual Transform Coder	197
7.3.1 PXFM	198
7.3.2 SEPXF M	199
7.4 Brandenburg-Johnston Hybrid Coder	200
7.5 CNET Coders	201
7.5.1 CNET DFT Coder	201
7.5.2 CNET MDCT Coder 1	201
7.5.3 CNET MDCT Coder 2	202
7.6 Adaptive Spectral Entropy Coding	203
7.7 Differential Perceptual Audio Coder	204
7.8 DFT Noise Substitution	205
7.9 DCT with Vector Quantization	206
7.10 MDCT with Vector Quantization	207
7.11 Summary	208
Problems	208
Computer Exercises	210
8 SUBBAND CODERS	211
8.1 Introduction	211
8.1.1 Subband Algorithms	212
8.2 DWT and Discrete Wavelet Packet Transform (DWPT)	214
8.3 Adapted WP Algorithms	218
8.3.1 DWPT Coder with Globally Adapted Daubechies Analysis Wavelet	218
8.3.2 Scalable DWPT Coder with Adaptive Tree Structure	220
8.3.3 DWPT Coder with Globally Adapted General Analysis Wavelet	223
8.3.4 DWPT Coder with Adaptive Tree Structure and Locally Adapted Analysis Wavelet	223
8.3.5 DWPT Coder with Perceptually Optimized Synthesis Wavelets	224
8.4 Adapted Nonuniform Filter Banks	226
8.4.1 Switched Nonuniform Filter Bank Cascade	226
8.4.2 Frequency-Varying Modulated Lapped Transforms	227
8.5 Hybrid WP and Adapted WP/Sinusoidal Algorithms	227

8.5.1	Hybrid Sinusoidal/Classical DWPT Coder	228
8.5.2	Hybrid Sinusoidal/ M -band DWPT Coder	229
8.5.3	Hybrid Sinusoidal/DWPT Coder with WP Tree Structure Adaptation (ARCO)	230
8.6	Subband Coding with Hybrid Filter Bank/CELP Algorithms	233
8.6.1	Hybrid Subband/CELP Algorithm for Low-Delay Applications	234
8.6.2	Hybrid Subband/CELP Algorithm for Low-Complexity Applications	235
8.7	Subband Coding with IIR Filter Banks	237
	Problems	237
	Computer Exercise	240
9	SINUSOIDAL CODERS	241
9.1	Introduction	241
9.2	The Sinusoidal Model	242
9.2.1	Sinusoidal Analysis and Parameter Tracking	242
9.2.2	Sinusoidal Synthesis and Parameter Interpolation	245
9.3	Analysis/Synthesis Audio Codec (ASAC)	247
9.3.1	ASAC Segmentation	248
9.3.2	ASAC Sinusoidal Analysis-by-Synthesis	248
9.3.3	ASAC Bit Allocation, Quantization, Encoding, and Scalability	248
9.4	Harmonic and Individual Lines Plus Noise Coder (HILN)	249
9.4.1	HILN Sinusoidal Analysis-by-Synthesis	250
9.4.2	HILN Bit Allocation, Quantization, Encoding, and Decoding	251
9.5	FM Synthesis	251
9.5.1	Principles of FM Synthesis	252
9.5.2	Perceptual Audio Coding Using an FM Synthesis Model	252
9.6	The Sines + Transients + Noise (STN) Model	254
9.7	Hybrid Sinusoidal Coders	255
9.7.1	Hybrid Sinusoidal-MDCT Algorithm	256
9.7.2	Hybrid Sinusoidal-Vocoder Algorithm	257
9.8	Summary	258
	Problems	258
	Computer Exercises	259

10 AUDIO CODING STANDARDS AND ALGORITHMS	263
10.1 Introduction	263
10.2 MIDI <i>Versus</i> Digital Audio	264
10.2.1 MIDI Synthesizer	264
10.2.2 General MIDI (GM)	266
10.2.3 MIDI Applications	266
10.3 Multichannel Surround Sound	267
10.3.1 The Evolution of Surround Sound	267
10.3.2 The Mono, the Stereo, and the Surround Sound Formats	268
10.3.3 The ITU-R BS.775 5.1-Channel Configuration	268
10.4 MPEG Audio Standards	270
10.4.1 MPEG-1 Audio (ISO/IEC 11172-3)	275
10.4.2 MPEG-2 BC/LSF (ISO/IEC-13818-3)	279
10.4.3 MPEG-2 NBC/AAC (ISO/IEC-13818-7)	283
10.4.4 MPEG-4 Audio (ISO/IEC 14496-3)	289
10.4.5 MPEG-7 Audio (ISO/IEC 15938-4)	309
10.4.6 MPEG-21 Framework (ISO/IEC-21000)	317
10.4.7 MPEG Surround and Spatial Audio Coding	319
10.5 Adaptive Transform Acoustic Coding (ATRAC)	319
10.6 Lucent Technologies PAC, EPAC, and MPAC	321
10.6.1 Perceptual Audio Coder (PAC)	321
10.6.2 Enhanced PAC (EPAC)	323
10.6.3 Multichannel PAC (MPAC)	323
10.7 Dolby Audio Coding Standards	325
10.7.1 Dolby AC-2, AC-2A	325
10.7.2 Dolby AC-3/Dolby Digital/Dolby SR · D	327
10.8 Audio Processing Technology APT-x100	335
10.9 DTS – Coherent Acoustics	338
10.9.1 Framing and Subband Analysis	338
10.9.2 Psychoacoustic Analysis	339
10.9.3 ADPCM – Differential Subband Coding	339
10.9.4 Bit Allocation, Quantization, and Multiplexing	341
10.9.5 DTS-CA Versus Dolby Digital	342
Problems	342
Computer Exercise	342
11 LOSSLESS AUDIO CODING AND DIGITAL WATERMARKING	343
11.1 Introduction	343

11.2	Lossless Audio Coding (L ² AC)	344
11.2.1	L ² AC Principles	345
11.2.2	L ² AC Algorithms	346
11.3	DVD-Audio	356
11.3.1	Meridian Lossless Packing (MLP)	358
11.4	Super-Audio CD (SACD)	358
11.4.1	SACD Storage Format	362
11.4.2	Sigma-Delta Modulators (SDM)	362
11.4.3	Direct Stream Digital (DSD) Encoding	364
11.5	Digital Audio Watermarking	368
11.5.1	Background	370
11.5.2	A Generic Architecture for DAW	374
11.5.3	DAW Schemes – Attributes	377
11.6	Summary of Commercial Applications	378
	Problems	382
	Computer Exercise	382
12	QUALITY MEASURES FOR PERCEPTUAL AUDIO CODING	383
12.1	Introduction	383
12.2	Subjective Quality Measures	384
12.3	Confounding Factors in Subjective Evaluations	386
12.4	Subjective Evaluations of Two-Channel Standardized Codecs	387
12.5	Subjective Evaluations of 5.1-Channel Standardized Codecs	388
12.6	Subjective Evaluations Using Perceptual Measurement Systems	389
12.6.1	CIR Perceptual Measurement Schemes	390
12.6.2	NSE Perceptual Measurement Schemes	390
12.7	Algorithms for Perceptual Measurement	391
12.7.1	Example: Perceptual Audio Quality Measure (PAQM)	392
12.7.2	Example: Noise-to-Mask Ratio (NMR)	396
12.7.3	Example: Objective Audio Signal Evaluation (OASE)	399
12.8	ITU-R BS.1387 and ITU-T P.861: Standards for Perceptual Quality Measurement	401
12.9	Research Directions for Perceptual Codec Quality Measures	402
REFERENCES		405
INDEX		459

PREFACE

Audio processing and recording has been part of telecommunication and entertainment systems for more than a century. Moreover bandwidth issues associated with audio recording, transmission, and storage occupied engineers from the very early stages in this field. A series of important technological developments paved the way from early phonographs to magnetic tape recording, and lately compact disk (CD), and super storage devices. In the following, we capture some of the main events and milestones that mark the history in audio recording and storage.¹

Prototypes of phonographs appeared around 1877, and the first attempt to market cylinder-based gramophones was by the Columbia Phonograph Co. in 1889. Five years later, Marconi demonstrated the first radio transmission that marked the beginning of audio broadcasting. The Victor Talking Machine Company, with the little nipper dog as its trademark, was formed in 1901. The “telegraphone”, a magnetic recorder for voice that used still wire, was patented in Denmark around the end of the nineteenth century. The Odeon and His Masters Voice (HMV) label produced and marketed music recordings in the early nineteen hundreds. The cabinet phonograph with a horn called “Victrola” appeared at about the same time. Diamond disk players were marketed in 1913 followed by efforts to produce sound-on-film for motion pictures. Other milestones include the first commercial transmission in Pittsburgh and the emergence of public address amplifiers. Electrically recorded material appeared in the 1920s and the first sound-on-film was demonstrated in the mid 1920s by Warner Brothers. Cinema applications in the 1930s promoted advances in loudspeaker technologies leading to the development of woofer, tweeter, and crossover network concepts. Juke boxes for music also appeared in the 1930s. Magnetic tape recording was demonstrated in Germany in the 1930s by BASF and AEG/Telefunken. The Ampex tape recorders appeared in the US in the late 1940s. The demonstration of stereo high-fidelity (Hi-Fi) sound in the late 1940s spurred the development of amplifiers, speakers, and reel-to-reel tape recorders for home use in the 1950s both in Europe and



Apple iPod®. (Courtesy of Apple Computer, Inc.) Apple iPod® is a registered trademark of Apple Computer, Inc.

the US. Meanwhile, Columbia produced the 33-rpm long play (LP) vinyl record, while its rival RCA Victor produced the compact 45-rpm format whose sales took off with the emergence of rock and roll music. Technological developments in the mid 1950s resulted in the emergence of compact transistor-based radios and soon after small tape players. In 1963, Philips introduced the compact cassette tape format with its EL3300 series portable players (marketed in the US as Norelco) which became an instant success with accessories for home, portable, and car use. Eight track cassettes became popular in the late 1960s mainly for car use. The Dolby system for compact cassette noise reduction was also a landmark in the audio signal processing field. Meanwhile, FM broadcasting, which had been invented earlier, took off in the 1960s and 1970s with stereo transmissions. Helical tape-head technologies invented in Japan in the 1960s provided high-bandwidth recording capabilities which enabled video tape recorders for home use in the 1970s (e.g., VHS and Beta formats). This technology was also used in the 1980s for audio PCM stereo recording. Laser compact disk technology was introduced in 1982 and by the late 1980s became the preferred format for Hi-Fi stereo recording. Analog compact cassette players, high-quality reel-to-reel recorders, expensive turntables, and virtually all analog recording devices started fading away by the late 1980s. The launch of the digital CD audio format in

the 1980s coincided with the advent of personal computers, and took over in all aspects of music recording and distribution. CD playback soon dominated broadcasting, automobile, home stereo, and analog vinyl LP. The compact cassette formats became relics of an old era and eventually disappeared from music stores. Digital audio tape (DAT) systems enabled by helical tape head technology were also introduced in the 1980s but were commercially unsuccessful because of strict copyright laws and unusually large taxes.

Parallel developments in digital video formats for laser disk technologies included work in audio compression systems. Audio compression research papers started appearing mostly in the 1980s at IEEE ICASSP and Audio Engineering Society conferences by authors from several research and development labs including, Erlangen-Nuremberg University and Fraunhofer IIS, AT&T Bell Laboratories, and Dolby Laboratories. Audio compression or audio coding research, the art of representing an audio signal with the least number of information bits while maintaining its fidelity, went through quantum leaps in the late 1980s and 1990s. Although originally most audio compression algorithms were developed as part of the digital motion video compression standards, e.g., the MPEG series, these algorithms eventually became important as stand alone technologies for audio recording and playback. Progress in VLSI technologies, psychoacoustics and efficient time-frequency signal representations made possible a series of scalable real-time compression algorithms for use in audio and cinema applications. In the 1990s, we witnessed the emergence of the first products that used compressed audio formats such as the MiniDisc (MD) and the Digital Compact Cassette (DCC). The sound and video playing capabilities of the PC and the proliferation of multimedia content through the Internet had a profound impact on audio compression technologies. The MPEG-1/-2 layer III (MP3) algorithm became a defacto standard for Internet music downloads. Specialized web sites that feature music content changed the ways people buy and share music. Compact MP3 players appeared in the late 1990s. In the early 2000s, we had the emergence of the Apple iPod® player with a hard drive that supports MP3 and MPEG advanced audio coding (AAC) algorithms.

In order to enhance cinematic and home theater listening experiences and deliver greater realism than ever before, audio codec designers pursued sophisticated multichannel audio coding techniques. In the mid 1990s, techniques for encoding 5.1 separate channels of audio were standardized in MPEG-2 BC and later MPEG-2 AAC audio. Proprietary multichannel algorithms were also developed and commercialized by Dolby Laboratories (AC-3), Digital Theater System (DTS), Lucent (EPAC), Sony (SDDS), and Microsoft (WMA). Dolby Labs, DTS, Lexicon, and other companies also introduced 2:N channel upmix algorithms capable of synthesizing multichannel surround presentation from conventional stereo content (e.g., Dolby ProLogic II, DTS Neo6). The human auditory system is capable of localizing sound with greater spatial resolution than current multi-channel audio systems offer, and as a result the quest continues to achieve the ultimate spatial fidelity in sound reproduction. Research involving spatial audio, real-time acoustic source localization, binaural cue coding, and application of

head-related transfer functions (HRTF) towards rendering immersive audio has gained interest. Audiophiles appeared skeptical with the 44.1-kHz 16-bit CD stereo format and some were critical of the sound quality of compression formats. These ideas along with the need for copyright protection eventually gained momentum and new standards and formats appeared in the early 2000s. In particular, multichannel lossless coding such as the DVD-Audio (DVD-A) and the Super-Audio-CD (SACD) appeared. The standardization of these storage formats provided the audio codec designers with enormous storage capacity. This motivated *lossless* coding of digital audio.

The purpose of this book is to provide an in-depth treatment of audio compression algorithms and standards. The topic is currently occupying several communities in signal processing, multimedia, and audio engineering. The intended readership for this book includes at least three groups. At the highest level, any reader with a general scientific background will be able to gain an appreciation for the heuristics of perceptual coding. Secondly, readers with a general electrical and computer engineering background will become familiar with the essential signal processing techniques and perceptual models embedded in most audio coders. Finally, undergraduate and graduate students with focuses in multimedia, DSP, and computer music will gain important knowledge in signal analysis and audio coding algorithms. The vast body of literature provided and the tutorial aspects of the book make it an asset for audiophiles as well.

Organization

This book is in part the outcome of many years of research and teaching at Arizona State University. We opted to include exercises and computer problems and hence enable instructors to either use the content in existing DSP and multimedia courses, or to promote the creation of new courses with focus in audio and speech processing and coding. The book has twelve chapters and each chapter contains problems, proofs, and computer exercises. Chapter 1 introduces the readers to the field of audio signal processing and coding. In Chapter 2, we review the basic signal processing theory and emphasize concepts relevant to audio coding. Chapter 3 describes waveform quantization and entropy coding schemes. Chapter 4 covers linear predictive coding and its utility in speech and audio coding. Chapter 5 covers psychoacoustics and Chapter 6 explores filter bank design. Chapter 7 describes transform coding methodologies. Subband and sinusoidal coding algorithms are addressed in Chapters 8 and 9, respectively. Chapter 10 reviews several audio coding standards including the ISO/IEC MPEG family, the cinematic Sony SDDS, the Dolby AC-3, and the DTS-coherent acoustics (DTS-CA). Chapter 11 focuses on lossless audio coding and digital audio watermarking techniques. Chapter 12 provides information on subjective quality measures.

Use in Courses

For an undergraduate elective course with little or no background in DSP, the instructor can cover in detail Chapters 1, 2, 3, 4, and 5, then present select

sections of Chapter 6, and describe in an expository and qualitative manner certain basic algorithms and standards from Chapters 7-11. A graduate class in audio coding with students that have background in DSP, can start from Chapter 5 and cover in detail Chapters 6 through Chapter 11. Audio coding practitioners and researchers that are interested mostly in qualitative descriptions of the standards and information on bibliography can start at Chapter 5 and proceed reading through Chapter 11.

Trademarks and Copyrights

Sony Dynamic Digital Sound, SDDS, ATRAC, and MiniDisc are trademarks of Sony Corporation. Dolby, Dolby Digital, AC-2, AC-3, DolbyFAX, Dolby Pro-Logic are trademarks of Dolby laboratories. The perceptual audio coder (PAC), EPAC, and MPAC are trademarks of AT&T and Lucent Technologies. The APT-x100 is trademark of Audio Processing Technology Inc. The DTS-CA is trademark of Digital Theater Systems Inc. Apple iPod® is a registered trademark of Apple Computer, Inc.

Acknowledgments

The authors have all spent time at Arizona State University (ASU) and Prof. Spanias is in fact still teaching and directing research in this area at ASU. The group of authors has worked on grants with Intel Corporation and would like to thank this organization for providing grants in scalable speech and audio coding that created opportunities for in-depth studies in these areas. Special thanks to our colleagues in Intel Corporation at that time including Brian Mears, Gopal Nair, Hedayat Daie, Mark Walker, Michael Deisher, and Tom Gardos. We also wish to acknowledge the support of current Intel colleagues Gang Liang, Mike Rosenzweig, and Jim Zhou, as well as Scott Peirce for proof reading some of the material. Thanks also to former doctoral students at ASU including Philip Loizou and Sassan Ahmadi for many useful discussions in speech and audio processing. We appreciate also discussions on narrowband vocoders with Bruce Fette in the late 1990s then with Motorola GEG and now with General Dynamics.

The authors also acknowledge the National Science Foundation (NSF) CCLI for grants in education that supported in part the preparation of several computer examples and paradigms in psychoacoustics and signal coding. Also some of the early work in coding of Dr. Spanias was supported by the Naval Research Laboratories (NRL) and we would like to thank that organization for providing ideas for projects that inspired future work in this area. We also wish to thank ASU and some of the faculty and administrators that provided moral and material support for work in this area. Thanks are extended to current ASU students Shibani Misra, Visar Berisha, and Mahesh Banavar for proofreading some of the material. We thank the Wiley Interscience production team George Telecki, Melissa Yanuzzi, and Rachel Witmer for their diligent efforts in copyediting, cover design, and typesetting. We also thank all the anonymous reviewers for

their useful comments. Finally, we all wish to express our thanks to our families for their support.

The book content is used frequently in ASU online courses and industry short courses offered by Andreas Spanias. Contact Andreas Spanias (spanias@asu.edu / <http://www.fulton.asu.edu/~spanias/>) for details.

¹ Resources used for obtaining important dates in recording history include web sites at the University of San Diego, Arizona State University, and Wikipedia.

CHAPTER 1

INTRODUCTION

Audio coding or *audio compression* algorithms are used to obtain compact digital representations of high-fidelity (wideband) audio signals for the purpose of efficient transmission or storage. The central objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent signal reproduction, i.e., generating output audio that cannot be distinguished from the original input, even by a sensitive listener (“golden ears”). This text gives an in-depth treatment of algorithms and standards for transparent coding of high-fidelity audio.

1.1 HISTORICAL PERSPECTIVE

The introduction of the compact disc (CD) in the early 1980s brought to the fore all of the advantages of digital audio representation, including true high-fidelity, dynamic range, and robustness. These advantages, however, came at the expense of high data rates. Conventional CD and digital audio tape (DAT) systems are typically sampled at either 44.1 or 48 kHz using pulse code modulation (PCM) with a 16-bit sample resolution. This results in uncompressed data rates of 705.6/768 kb/s for a monaural channel, or 1.41/1.54 Mb/s for a stereo-pair. Although these data rates were accommodated successfully in first-generation CD and DAT players, second-generation audio players and wirelessly connected systems are often subject to bandwidth constraints that are incompatible with high data rates. Because of the success enjoyed by the first-generation

systems, however, end users have come to expect “CD-quality” audio reproduction from any digital system. Therefore, new network and wireless multimedia digital audio systems must reduce data rates without compromising reproduction quality. Motivated by the need for compression algorithms that can satisfy simultaneously the conflicting demands of high compression ratios and transparent quality for high-fidelity audio signals, several coding methodologies have been established over the last two decades. Audio compression schemes, in general, employ design techniques that exploit both *perceptual irrelevancies* and *statistical redundancies*.

PCM was the primary audio encoding scheme employed until the early 1980s. PCM does not provide any mechanisms for redundancy removal. Quantization methods that exploit the signal correlation, such as differential PCM (DPCM), delta modulation [Jaya76] [Jaya84], and adaptive DPCM (ADPCM) were applied to audio compression later (e.g., PC audio cards). Owing to the need for drastic reduction in bit rates, researchers began to pursue new approaches for audio coding based on the *principles of psychoacoustics* [Zwic90] [Moor03]. Psychoacoustic notions in conjunction with the basic properties of signal quantization have led to the theory of *perceptual entropy* [John88a] [John88b]. Perceptual entropy is a quantitative estimate of the fundamental limit of transparent audio signal compression. Another key contribution to the field was the characterization of the auditory filter bank and particularly the time-frequency analysis capabilities of the inner ear [Moor83]. Over the years, several *filter-bank* structures that mimic the critical band structure of the auditory filter bank have been proposed. A filter bank is a parallel bank of bandpass filters covering the audio spectrum, which, when used in conjunction with a perceptual model, can play an important role in the identification of perceptual irrelevancies.

During the early 1990s, several workgroups and organizations such as the International Organization for Standardization/International Electro-technical Commission (ISO/IEC), the International Telecommunications Union (ITU), AT&T, Dolby Laboratories, Digital Theatre Systems (DTS), Lucent Technologies, Philips, and Sony were actively involved in developing perceptual audio coding algorithms and standards. Some of the popular commercial standards published in the early 1990s include Dolby’s Audio Coder-3 (AC-3), the DTS Coherent Acoustics (DTS-CA), Lucent Technologies’ Perceptual Audio Coder (PAC), Philips’ Precision Adaptive Subband Coding (PASC), and Sony’s Adaptive Transform Acoustic Coding (ATRAC). Table 1.1 lists chronologically some of the prominent audio coding standards. The commercial success enjoyed by these audio coding standards triggered the launch of several multimedia storage formats.

Table 1.2 lists some of the popular multimedia storage formats since the beginning of the CD era. High-performance stereo systems became quite common with the advent of CDs in the early 1980s. A compact-disc–read only memory (CD-ROM) can store data up to 700–800 MB in digital form as “microscopic-pits” that can be read by a laser beam off of a reflective surface or a medium. Three competing storage media – DAT, the digital compact cassette (DCC), and the

Table 1.1. List of perceptual and lossless audio coding standards/algorithms.

Standard/algorithm	Related references
1. ISO/IEC MPEG-1 audio	[ISOI92]
2. Philips' PASC (for DCC applications)	[Lokh92]
3. AT&T/Lucent PAC/EPAC	[John96c] [Sinh96]
4. Dolby AC-2	[Davi92] [Fiel91]
5. AC-3/Dolby Digital	[Davis93] [Fiel96]
6. ISO/IEC MPEG-2 (BC/LSF) audio	[ISOI94a]
7. Sony's ATRAC; (MiniDisc and SDDS)	[Yosh94] [Tsut96]
8. SHORTEN	[Robi94]
9. Audio processing technology – APT-x100	[Wyli96b]
10. ISO/IEC MPEG-2 AAC	[ISOI96]
11. DTS coherent acoustics	[Smyt96] [Smyt99]
12. The DVD Algorithm	[Crav96] [Crav97]
13. MUSICompress	[Wege97]
14. Lossless transform coding of audio (LTAC)	[Pura97]
15. AudioPaK	[Hans98b] [Hans01]
16. ISO/IEC MPEG-4 audio version 1	[ISOI99]
17. Meridian lossless packing (MLP)	[Gerz99]
18. ISO/IEC MPEG-4 audio version 2	[ISOI00]
19. Audio coding based on integer transforms	[Geig01] [Geig02]
20. Direct-stream digital (DSD) technology	[Reef01a] [Jans03]

Table 1.2. Some of the popular audio storage formats.

Audio storage format	Related references
1. Compact disc	[CD82] [IECA87]
2. Digital audio tape (DAT)	[Watk88] [Tan89]
3. Digital compact cassette (DCC)	[Lokh91] [Lokh92]
4. MiniDisc	[Yosh94] [Tsut96]
5. Digital versatile disc (DVD)	[DVD96]
6. DVD-audio (DVD-A)	[DVD01]
7. Super audio CD (SACD)	[SACD02]

MiniDisc (MD) – entered the commercial market during 1987–1992. Intended mainly for back-up high-density storage (~ 1.3 GB), the DAT became the primary source of mass data storage/transfer [Watk88] [Tan89]. In 1991–1992, Sony proposed a storage medium called the MiniDisc, primarily for audio storage. MD employs the ATRAC algorithm for compression. In 1991, Philips introduced the DCC, a successor of the analog compact cassette. Philips DCC employs a compression scheme called the PASC [Lokh91] [Lokh92] [Hoog94]. The DCC began

as a potential competitor for DATs but was discontinued in 1996. The introduction of the digital versatile disc (DVD) in 1996 enabled both video and audio recording/storage as well as text-message programming. The DVD became one of the most successful storage media. With the improvements in the audio compression and DVD storage technologies, multichannel surround sound encoding formats gained interest [Bosi93] [Holm99] [Bosi00].

With the emergence of streaming audio applications, during the late 1990s, researchers pursued techniques such as combined speech and audio architectures, as well as joint source-channel coding algorithms that are optimized for the packet-switched Internet. The advent of ISO/IEC MPEG-4 standard (1996–2000) [ISOI99] [ISOI00] established new research goals for high-quality coding of audio at low bit rates. MPEG-4 audio encompasses more functionality than perceptual coding [Koen98] [Koen99]. It comprises an integrated family of algorithms with provisions for scalable, object-based speech and audio coding at bit rates from as low as 200 b/s up to 64 kb/s per channel.

The emergence of the DVD-audio and the super audio CD (SACD) provided designers with additional storage capacity, which motivated research in *lossless* audio coding [Crav96] [Gerz99] [Reef01a]. A lossless audio coding system is able to reconstruct perfectly a bit-for-bit representation of the original input audio. In contrast, a coding scheme incapable of perfect reconstruction is called *lossy*. For most audio program material, lossy schemes offer the advantage of lower bit rates (e.g., less than 1 bit per sample) relative to lossless schemes (e.g., 10 bits per sample). Delivering real-time lossless audio content to the network browser at low bit rates is the next grand challenge for codec designers.

1.2 A GENERAL PERCEPTUAL AUDIO CODING ARCHITECTURE

Over the last few years, researchers have proposed several efficient signal models (e.g., transform-based, subband-filter structures, wavelet-packet) and compression standards (Table 1.1) for high-quality digital audio reproduction. Most of these algorithms are based on the generic architecture shown in Figure 1.1.

The coders typically segment input signals into quasi-stationary frames ranging from 2 to 50 ms. Then, a time-frequency analysis section estimates the temporal and spectral components of each frame. The time-frequency mapping is usually matched to the analysis properties of the human auditory system. Either way, the ultimate objective is to extract from the input audio a set of time-frequency parameters that is amenable to quantization according to a *perceptual distortion metric*. Depending on the overall design objectives, the time-frequency analysis section usually contains one of the following:

- Unitary transform
- Time-invariant bank of critically sampled, uniform/nonuniform bandpass filters

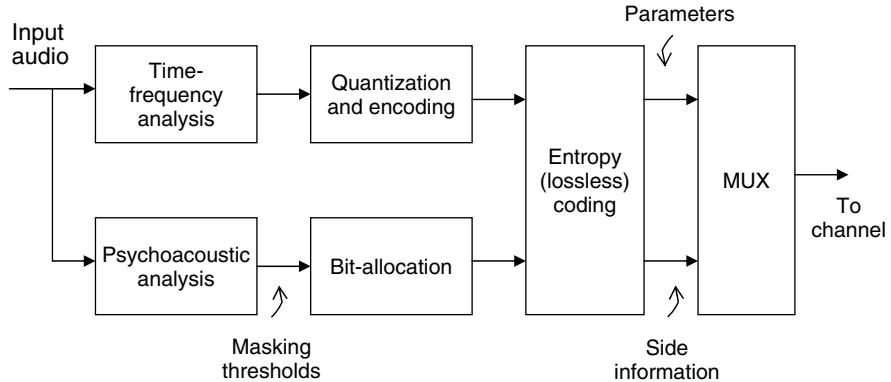


Figure 1.1. A generic perceptual audio encoder.

- Time-varying (signal-adaptive) bank of critically sampled, uniform/nonuniform bandpass filters
- Harmonic/sinusoidal analyzer
- Source-system analysis (LPC and multipulse excitation)
- Hybrid versions of the above.

The choice of time-frequency analysis methodology always involves a fundamental tradeoff between time and frequency resolution requirements. Perceptual distortion control is achieved by a psychoacoustic signal analysis section that estimates signal masking power based on psychoacoustic principles. The psychoacoustic model delivers masking thresholds that quantify 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 psychoacoustic model therefore allows the quantization section to exploit perceptual irrelevancies. This section can also exploit statistical redundancies through classical techniques such as DPCM or ADPCM. Once a quantized compact parametric set has been formed, the remaining redundancies are typically removed through noiseless run-length (RL) and entropy coding techniques, e.g., Huffman [Cove91], arithmetic [Witt87], or Lempel-Ziv-Welch (LZW) [Ziv77] [Welc84]. Since the output of the psychoacoustic distortion control model is signal-dependent, most algorithms are inherently variable rate. Fixed channel rate requirements are usually satisfied through buffer feedback schemes, which often introduce encoding delays.

1.3 AUDIO CODER ATTRIBUTES

Perceptual audio coders are typically evaluated based on the following attributes: audio reproduction quality, operating bit rates, computational complexity, codec delay, and channel error robustness. The objective is to attain a high-quality (transparent) audio output at low bit rates (<32 kb/s), with an acceptable

algorithmic delay (~ 5 to 20 ms), and with low computational complexity (~ 1 to 10 million instructions per second, or MIPS).

1.3.1 Audio Quality

Audio quality is of paramount importance when designing an audio coding algorithm. Successful strides have been made since the development of simple near-transparent perceptual coders. Typically, classical objective measures of signal fidelity such as the signal to noise ratio (SNR) and the total harmonic distortion (THD) are inadequate [Ryde96]. As the field of perceptual audio coding matured rapidly and created greater demand for listening tests, there was a corresponding growth of interest in perceptual measurement schemes. Several subjective and objective quality measures have been proposed and standardized during the last decade. Some of these schemes include the noise-to-mask ratio (NMR, 1987) [Bran87a] the perceptual audio quality measure (PAQM, 1991) [Beer91], the perceptual evaluation (PERCEVAL, 1992) [Pail92], the perceptual objective measure (POM, 1995) [Colo95], and the objective audio signal evaluation (OASE, 1997) [Spor97]. We will address these and several other quality assessment schemes in detail in Chapter 12.

1.3.2 Bit Rates

From a codec designer's point of view, one of the key challenges is to represent high-fidelity audio with a minimum number of bits. For instance, if a 5-ms audio frame sampled at 48 kHz (240 samples per frame) is represented using 80 bits, then the encoding bit rate would be $80 \text{ bits}/5 \text{ ms} = 16 \text{ kb/s}$. Low bit rates imply high compression ratios and generally low reproduction quality. Early coders such as the ISO/IEC MPEG-1 (32–448 kb/s), the Dolby AC-3 (32–384 kb/s), the Sony ATRAC (256 kb/s), and the Philips PASC (192 kb/s) employ high bit rates for obtaining transparent audio reproduction. However, the development of several sophisticated audio coding tools (e.g., MPEG-4 audio tools) created ways for efficient transmission or storage of audio at rates between 8 and 32 kb/s. Future audio coding algorithms promise to offer reasonable quality at low rates along with the ability to scale both *rate* and *quality* to match different requirements such as time-varying channel capacity.

1.3.3 Complexity

Reduced computational complexity not only enables real-time implementation but may also decrease the power consumption and extend battery life. Computational complexity is usually measured in terms of millions of instructions per second (MIPS). Complexity estimates are processor-dependent. For example, the complexity associated with Dolby's AC-3 decoder was estimated at approximately 27 MIPS using the Zoran ZR38001 general-purpose DSP core [Vern95]; for the Motorola DSP56002 processor, the complexity was estimated at 45 MIPS [Vern95]. Usually, most of the audio codecs rely on the so-called asymmetric encoding principle. This means that the codec complexity is not evenly

shared between the encoder and the decoder (typically, encoder 80% and decoder 20% complexity), with more emphasis on reducing the decoder complexity.

1.3.4 Codec Delay

Many of the network applications for high-fidelity audio (streaming audio, audio-on-demand) are delay tolerant (up to 100–200 ms), providing the opportunity to exploit long-term signal properties in order to achieve high coding gain. However, in two-way real-time communication and voice-over Internet protocol (VoIP) applications, low-delay encoding (10–20 ms) is important. Consider the example described before, i.e., an audio coder operating on frames of 5 ms at a 48 kHz sampling frequency. In an ideal encoding scenario, the minimum amount of delay should be 5 ms at the encoder and 5 ms at the decoder (same as the frame length). However, other factors such as analysis-synthesis filter bank window, the look-ahead, the bit-reservoir, and the channel delay contribute to additional delays. Employing shorter analysis-synthesis windows, avoiding look-ahead, and re-structuring the bit-reservoir functions could result in low-delay encoding, nonetheless, with reduced coding efficiencies.

1.3.5 Error Robustness

The increasing popularity of streaming audio over packet-switched and wireless networks such as the Internet implies that any algorithm intended for such applications must be able to deal with a noisy time-varying channel. In particular, provisions for error robustness and error protection must be incorporated at the encoder in order to achieve reliable transmission of digital audio over error-prone channels. One simple idea could be to provide better protection to the error-sensitive and priority (important) bits. For instance, the audio frame header requires the maximum error robustness; otherwise, transmission errors in the header will seriously impair the entire audio frame. Several error detecting/correcting codes [Lin82] [Wick95] [Bayl97] [Swee02] [Zara02] can also be employed. Inclusion of error correcting codes in the bitstream might help to obtain error-free reproduction of the input audio, however, with increased complexity and bit rates.

From the discussion in the previous sections, it is evident that several tradeoffs must be considered in designing an algorithm for a particular application. For this reason, audio coding standards consist of several tools that enable the design of scalable algorithms. For example, MPEG-4 provides tools to design algorithms that satisfy a variety of bit rate, delay, complexity, and robustness requirements.

1.4 TYPES OF AUDIO CODERS – AN OVERVIEW

Based on the signal model or the analysis-synthesis technique employed to encode audio signals, audio coders can be broadly classified as follows:

- Linear predictive
- Transform

- Subband
- Sinusoidal.

Algorithms are also classified based on the lossy or the lossless nature of audio coding. Lossy audio coding schemes achieve compression by exploiting perceptually irrelevant information. Some examples of lossy audio coding schemes include the ISO/IEC MPEG codec series, the Dolby AC-3, and the DTS CA. In lossless audio coding, the audio data is merely “packed” to obtain a bit-for-bit representation of the original. The meridian lossless packing (MLP) [Gerz99] and the direct stream digital (DSD) techniques [Brue97] [Reef01a] form a class of high-end lossless compression algorithms that are embedded in the DVD-audio [DVD01] and the SACD [SACD02] storage formats, respectively. Lossless audio coding techniques, in general yield high-quality digital audio without any artifacts at high rates. For instance, perceptual audio coding yields compression ratios from 10:1 to 25:1, while lossless audio coding can achieve compression ratios from 2:1 to 4:1.

1.5 ORGANIZATION OF THE BOOK

This book is organized as follows. In Chapter 2, we review basic signal processing concepts associated with audio coding. Chapter 3 provides introductory material to waveform quantization and entropy coding schemes. Some of the key topics covered in this chapter include scalar quantization, uniform/nonuniform quantization, pulse code modulation (PCM), differential PCM (DPCM), adaptive DPCM (ADPCM), vector quantization (VQ), bit-allocation techniques, and entropy coding schemes (Huffman, Rice, and arithmetic).

Chapter 4 provides information on linear prediction and its application in narrow and wideband coding. First, we address the utility of LP analysis/synthesis approach in speech applications. Next, we describe the open-loop analysis-synthesis LP and closed-loop analysis-by-synthesis LP techniques.

In Chapter 5, psychoacoustic principles are described. Johnston’s notion of perceptual entropy is presented as a measure of the fundamental limit of transparent compression for audio. The ISO/IEC 11172-3 MPEG-1 psychoacoustic analysis model 1 is used to describe the five important steps associated with the global masking threshold computation. Chapter 6 explores filter bank design issues and algorithms, with a particular emphasis placed on the modified discrete cosine transform (MDCT) that is widely used in several perceptual audio coding algorithms. Chapter 6 also addresses pre-echo artifacts and control strategies.

Chapters 7, 8, and 9 review established and emerging techniques for transparent coding of FM and CD-quality audio signals, including several algorithms that have become international standards. Transform coding methodologies are described in Chapter 7, subband coding algorithms are addressed in Chapter 8, and sinusoidal algorithms are presented in Chapter 9. In addition to methods based on uniform bandwidth filter banks, Chapter 8 covers coding methods that