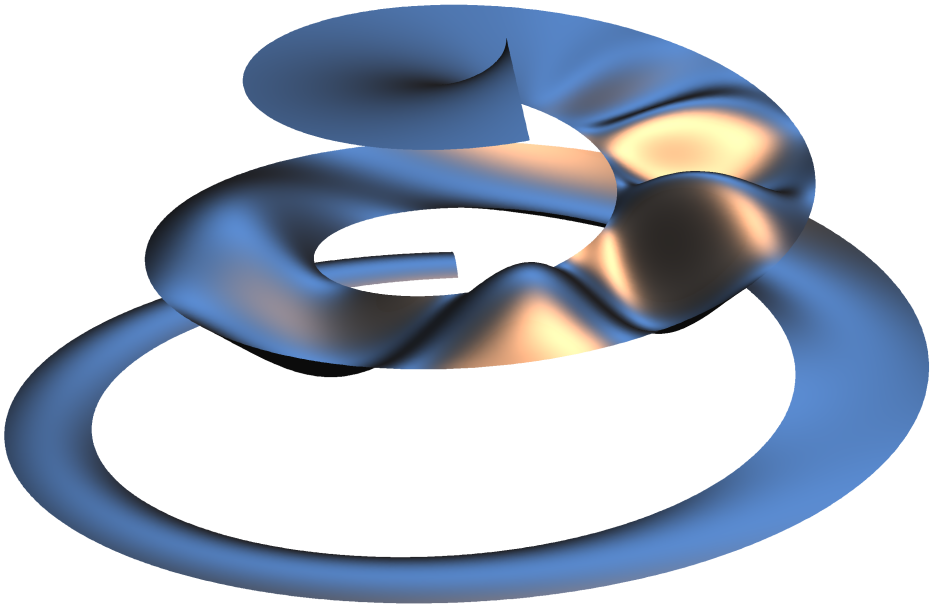
Human and Machine Hearing

—

Extracting Meaning from Sound

– – – –

(Author's **DRAFT** of a 2017 Cambridge University Press book)



Richard F. Lyon

January 8, 2017

Un beau visage est le plus beau de tous les spectacles ; & l'harmonie la plus douce est le son de voix de celle que l'on aime.

A fine Face is the finest of all Sights, and the sweetest Musick, the Sound of her Voice whom we love.

—Jean La Bruyère (1713) from 1691 French original.

This book is dedicated to my family: my beautiful, smart, cheerful, successful, inspiring, and sweet-voiced wife Peggy Asprey and our awesome children Susan and Erik—they are the loves of my life, and my fortune. Though this book has sometimes absorbed too much of my attention, they have all supported me in writing it, in so many ways. They are my finest of all sights, and sweetest music; they sustain me.

Contents

Foreword *Roy D. Patterson* **xi**

Preface **xv**

I Sound Analysis and Representation Overview **1**

1 Introduction **5**
1.1 On Vision and Hearing *à la* David Marr . . . 7
1.2 Top-Down versus Bottom-Up Analysis . . . 11
1.3 The Neuromimetic Approach . . . 12
1.4 Auditory Images . . . 13
1.5 The Ear as a Frequency Analyzer? . . . 14
1.6 The Third Sound . . . 15
1.7 Sound Understanding and Extraction of Meaning . . . 16
1.8 Leveraging Techniques from Machine Vision and Machine Learning . . . 17
1.9 Machine Hearing Systems "by the Book" . . . 17

2 Theories of Hearing **21**
2.1 A "New" Theory of Hearing . . . 21
2.2 Newer Theories of Hearing . . . 23
2.3 Active and Nonlinear Theories of Hearing . . . 24
2.4 Three Auditory Theories . . . 25
2.5 The Auditory Image Theory of Hearing . . . 26

3 On Logarithmic and Power-Law Hearing **31**
3.1 Logarithms and Power Laws . . . 31
3.2 Log Frequency . . . 34
3.3 Log Power . . . 35
3.4 Bode Plots . . . 36
3.5 Perceptual Mappings . . . 39
3.6 Constant-Q Analysis . . . 41
3.7 Use Logarithms with Caution . . . 42

4 Human Hearing Overview **43**
4.1 Human versus Machine . . . 43
4.2 Auditory Physiology . . . 43
4.3 Key Problems in Hearing . . . 45

4.4 Loudness . . . 47
4.5 Critical Bands, Masking, and Suppression . . . 50
4.6 Pitch Perception . . . 52
4.7 Timbre . . . 60
4.8 Consonance and Dissonance . . . 61
4.9 Speech Perception . . . 63
4.10 Binaural Hearing . . . 66
4.11 Auditory Streaming . . . 67
4.12 Nonlinearity . . . 68
4.13 A Way Forward . . . 69

5 Acoustic Approaches and Auditory Influence 71
5.1 Sound, Speech, and Music Modeling . . . 71
5.2 Short-Time Spectral Analysis . . . 72
5.3 Smoothing and Transformation of Spectra . . . 76
5.4 The Source–Filter Model and Homomorphic Signal Processing . . . 78
5.5 Backing Away from Logarithms . . . 80
5.6 Auditory Frequency Scales . . . 80
5.7 Mel-Frequency Cepstrum . . . 81
5.8 Linear Predictive Coding . . . 83
5.9 PLP and RASTA . . . 84
5.10 Auditory Techniques in Automatic Speech Recognition . . . 84
5.11 Improvements Needed . . . 85

II Systems Theory for Hearing 87

6 Introduction to Linear Systems 91
6.1 Smoothing: A Good Place to Start . . . 91
6.2 Linear Time-Invariant Systems . . . 92
6.3 Filters and Frequencies . . . 93
6.4 Differential Equations and Homogeneous Solutions . . . 96
6.5 Impulse Responses . . . 96
6.6 Causality and Stability . . . 98
6.7 Convolution . . . 99
6.8 Eigenfunctions and Transfer Functions . . . 99
6.9 Frequency Response . . . 103
6.10 Transforms and Operational Methods . . . 104
6.11 Rational Functions, and Their Poles and Zeros . . . 106
6.12 Graphical Computation of Transfer Function Gain and Phase . . . 108
6.13 Convolution Theorem . . . 111
6.14 Interconnection of Filters in Cascade, Parallel, and Feedback . . . 111
6.15 Summary and Next Steps . . . 113

7 Discrete-Time and Digital Systems **117**
7.1 Simulating Systems in Computers . . . 117
7.2 Discrete-Time Linear Shift-Invariant Systems . . . 117
7.3 Impulse Response and Convolution . . . 118
7.4 Frequency in Discrete-Time Systems . . . 118
7.5 Z Transform and Its Inverse . . . 118
7.6 Unit Advance and Unit Delay Operators . . . 119
7.7 Filters and Transfer Functions . . . 120
7.8 Sampling and Aliasing . . . 122
7.9 Mappings from Continuous-Time Systems . . . 126
7.10 Filter Design . . . 127
7.11 Digital Filters . . . 128
7.12 Multiple Inputs and Outputs . . . 131
7.13 Fourier Analysis and Spectrograms . . . 131
7.14 Perspective and Further Reading . . . 133

8 Resonators **135**
8.1 Bandpass Filters . . . 135
8.2 Four Resonant Systems . . . 139
8.3 Resonator Frequency Responses . . . 142
8.4 Resonator Impulse Responses . . . 145
8.5 The Complex Resonator and the Universal Resonance Curve . . . 147
8.6 Complex Zeros from a Parallel System . . . 150
8.7 Keeping It Real . . . 153
8.8 Digital Resonators . . . 155

9 Gammatone and Related Filters **159**
9.1 Compound Resonators as Auditory Models . . . 159
9.2 Multiple Poles . . . 159
9.3 The Complex Gammatone Filter . . . 161
9.4 The Real Gammatone Filter . . . 165
9.5 All-Pole Gammatone Filters . . . 167
9.6 Gammachirp Filters . . . 169
9.7 Variable Pole Q . . . 171
9.8 Noncoincident Poles . . . 173
9.9 Digital Implementations . . . 173

10 Nonlinear Systems **179**
10.1 Volterra Series and Other Descriptions . . . 179
10.2 Essential Nonlinearity . . . 181
10.3 Hopf Bifurcation . . . 181
10.4 Distributed Bandpass Nonlinearity . . . 183
10.5 Response Curves of Nonlinear Systems . . . 184
10.6 Two-Tone Responses . . . 187
10.7 Nonlinearity and Aliasing . . . 188
10.8 Cautions . . . 189

11 Automatic Gain Control **191**
11.1 Input–Output Level Compression . . . 191
11.2 Nonlinear Feedback Control . . . 192
11.3 AGC Compression at Equilibrium . . . 193
11.4 Multiple Cascaded Variable-Gain Stages . . . 197
11.5 Gain Control via Damping Control in Cascaded Resonators . . . 197
11.6 AGC Dynamics . . . 198
11.7 AGC Loop Stability . . . 201
11.8 Multiple-Loop AGC . . . 205

12 Waves in Distributed Systems **207**
12.1 Waves in Uniform Linear Media . . . 209
12.2 Transfer Functions from Wavenumbers . . . 217
12.3 Nonuniform Media . . . 218
12.4 Nonuniform Media as Filter Cascades . . . 221
12.5 Impulse Responses . . . 222
12.6 Group Velocity and Group Delay . . . 222

III The Auditory Periphery **225**

13 Auditory Filter Models **229**
13.1 What Is an Auditory Filter? . . . 230
13.2 From Resonance to Gaussian Filters . . . 232
13.3 Ten Good Properties for Auditory Filter Models . . . 233
13.4 Representative Auditory Filter Models . . . 235
13.5 Complications: Time-Varying and Nonlinear Auditory Filters . . . 239
13.6 Fitting Parameters of Filter Models . . . 241
13.7 Suppression . . . 244
13.8 Impulse Responses from Physiological Data . . . 245
13.9 Summary and Application to Cochlear Models . . . 248

14 Modeling the Cochlea **249**
14.1 On the Structure of the Cochlea . . . 250
14.2 The Traveling Wave . . . 250
14.3 1D, 2D, and 3D Hydrodynamics . . . 256
14.4 Long Waves, Short Waves, and 2D Models . . . 259
14.5 Active Micromechanics . . . 262
14.6 Scaling Symmetry and the Cochlear Map . . . 262
14.7 Filter-Cascade Cochlear Models . . . 263
14.8 Outer Hair Cells as Active Gain Elements . . . 267
14.9 Dispersion Relations from Mechanical Models and Experiments . . . 268
14.10 Inner Hair Cells as Detectors . . . 270
14.11 Adaptation to Sound via Efferent Control . . . 270
14.12 Summary and Further Reading . . . 273

15 The CARFAC Digital Cochlear Model **275**
15.1 Putting the Pieces Together . . . 275
15.2 The CARFAC Framework . . . 276
15.3 Physiological Elements . . . 276
15.4 Analog and Bidirectional Models . . . 277
15.5 Open-Source Software . . . 279
15.6 Detailing the CARFAC . . . 279

16 The Cascade of Asymmetric Resonators **281**
16.1 The Linear Cochlear Model . . . 281
16.2 Coupled-Form Filter Realization . . . 283

17 The Outer Hair Cell **293**
17.1 Multiple Effects in One Mechanism . . . 293
17.2 The Nonlinear Function . . . 294
17.3 AGC Effect of DOHC . . . 297
17.4 Typical Distortion Response Patterns . . . 299
17.5 Completing the Loop . . . 302

18 The Inner Hair Cell **305**
18.1 Rectification with a Sigmoid . . . 305
18.2 Adaptive Hair-Cell Models . . . 309
18.3 A Digital IHC Model . . . 312

19 The AGC Loop Filter **315**
19.1 The CARFAC's AGC Loop . . . 315
19.2 AGC Filter Structure . . . 317
19.3 Smoothing Filter Pole–Zero Analysis . . . 317
19.4 AGC Filter Temporal Response . . . 319
19.5 AGC Filter Spatial Response . . . 324
19.6 Time–Space Smoothing with Decimation . . . 324
19.7 Adapted Behavior . . . 325
19.8 Binaural or Multi-Ear Operation . . . 325
19.9 Coupled and Multistage AGC in CARFAC and Other Systems . . . 326

IV The Auditory Nervous System **329**

20 Auditory Nerve and Cochlear Nucleus **333**
20.1 From Hair Cells to Nerve Firings . . . 333
20.2 Tonotopic Organization . . . 336
20.3 Fine Time Structure in Cochleagrams . . . 336
20.4 Cell Types in the Cochlear Nucleus . . . 337
20.5 Inhibition and Other Computation . . . 338
20.6 Spike Timing Codes . . . 339

21 The Auditory Image **341**
21.1 Movies of Sound . . . 341
21.2 History . . . 342
21.3 Stabilizing the Image . . . 344
21.4 Triggered Temporal Integration . . . 344
21.5 Conventional Short-Time Autocorrelation . . . 349
21.6 Asymmetry . . . 349
21.7 Computing the SAI . . . 351
21.8 Pitch and Spectrum . . . 353
21.9 Auditory Images of Music . . . 353
21.10 Auditory Images of Speech . . . 355
21.11 Summary SAI Tracks: Pitchograms . . . 357
21.12 Cochleagram from SAI . . . 357
21.13 The Log-Lag SAI . . . 359

22 Binaural Spatial Hearing **365**
22.1 Rayleigh's Duplex Theory: Interaural Level and Phase . . . 365
22.2 Interaural Time and Level Differences . . . 366
22.3 The Head-Related Transfer Function . . . 371
22.4 Neural Extraction of Interaural Differences . . . 374
22.5 The Role of the Cochlear Nucleus and the Trapezoid Body . . . 377
22.6 Binaural Acoustic Reflex and Gain Control . . . 379
22.7 The Precedence Effect . . . 379
22.8 Completing the Model . . . 380
22.9 Interaural Coherence . . . 381
22.10 Binaural Applications . . . 381

23 The Auditory Brain **383**
23.1 Scene Analysis: ASA and CASA . . . 383
23.2 Attention and Stream Segregation . . . 385
23.3 Stages in the Brain . . . 389
23.4 Higher Auditory Pathways . . . 391
23.5 Prospects . . . 395

V Learning and Applications **397**

24 Neural Networks for Machine Learning **401**
24.1 Learning from Data . . . 401
24.2 The Perceptron . . . 402
24.3 The Training Phase . . . 403
24.4 Nonlinearities at the Output . . . 403
24.5 Nonlinearities at the Input . . . 407
24.6 Multiple Layers . . . 408
24.7 Neural Units and Neural Networks . . . 409
24.8 Training by Error Back-Propagation . . . 409
24.9 Cost Functions and Regularization . . . 412
24.10 Multiclass Classifiers . . . 412

24.11 Neural Network Successes and Failures . . . 414
24.12 Statistical Learning Theory . . . 416
24.13 Summary and Perspective . . . 417

25 Feature Spaces **419**
25.1 Feature Engineering . . . 420
25.2 Automatic Feature Optimization by Deep Networks . . . 421
25.3 Bandpass Power and Quadratic Features . . . 422
25.4 Quadratic Features of Cochlear Filterbank Outputs . . . 423
25.5 Nonlinearities and Gain Control in Feature Extraction . . . 423
25.6 Neurally Inspired Feature Extraction . . . 424
25.7 Sparsification and Winner-Take-All Features . . . 425
25.8 Which Approach Will Win? . . . 426

26 Sound Search **427**
26.1 Modeling Sounds . . . 428
26.2 Ranking Sounds Given Text Queries . . . 433
26.3 Experiments . . . 436
26.4 Results . . . 438
26.5 Conclusions and Followup . . . 440

27 Musical Melody Matching **441**
27.1 Algorithm . . . 443
27.2 Experiments . . . 449
27.3 Discussion . . . 452
27.4 Summary and Conclusions . . . 453

28 Other Applications **455**
28.1 Auditory Physiology and Psychoacoustics . . . 455
28.2 Audio Coding and Compression . . . 456
28.3 Hearing Aids and Cochlear Implants . . . 456
28.4 Visible Sound . . . 461
28.5 Diagnosis . . . 463
28.6 Speech and Speaker Recognition . . . 464
28.7 Music Information Retrieval . . . 465
28.8 Security, Surveillance, and Alarms . . . 466
28.9 Diarization, Summarization, and Indexing . . . 466
28.10 Have Fun . . . 467

Color Plates **469**

Bibliography **477**

Author Index **533**

Index **547**

Foreword

Roy D. Patterson

Human and Machine Hearing is a book for people who want to understand how the auditory system and the brain process sound, how to encapsulate aspects of our hearing knowledge in computer algorithms, and how to combine the algorithms into a machine that simulates the role of hearing in some aspect of everyday life—such as listening to the melody of a song or talking to a friend in a noisy restaurant. This is what Dick Lyon means by "Machine Hearing." The applications typically involve the segregation and identification of sound sources in everyday environments where there are competing sources and background noises—applications where there is reason to believe that the auditory form of sound analysis and feature extraction will be more effective and more robust than that provided by the traditional combination of the Fourier magnitude spectrum and MFCCs (mel-frequency cepstral coefficients). To construct a hearing machine and apply it to a real-world problem is an enormous undertaking; the latter half of the book documents the construction of a sophisticated auditory model and how it was integrated with machine learning algorithms to produce two hearing machines—an auditory search engine and an auditory melody matcher. The first half of the book describes the basic science that underpins machine hearing; it sets out the problems of constructing a stable, computationally efficient system, and it explains how to deal with each problem in turn. So the book is a comprehensive reference work for machine hearing with an ordered set of worked problems that culminate in two impressive demonstrations of machine hearing and its potential. This combination makes the book ideal both as a reference manual for experts working in the field of machine hearing and for graduate-level courses on machine hearing.

Lyon's idea of a machine hearing system has four "layers." The first two simulate auditory frequency analysis in the cochlea and auditory image construction in the brain stem. Together they form an auditory model that is intended to simulate all of the mechanical and neural processing required to produce your initial auditory image of a sound, that is, the internal auditory representation of sound that is thought to provide the basis for perception, streaming, auditory scene analysis, and all subsequent processing. The third layer applies application-dependent feature extraction to the auditory image and reduces the mass of features to a sparse form for the fourth layer, which extracts meaning with machine learning techniques. Together the third and fourth layers make the auditory model into a specific form of hearing machine, designed to perform a particular listening task.

The compact, authoritative introductions to auditory physiology, auditory perception, the acoustics of sound, and the mathematics of auditory filtering and auditory signal processing include the essential facts and functions, along with brief sketches of the people and experiments associated with milestones in the history of hearing research. This part of the book is a delightfully readable reference manual for machine hearing. Lyon's involvement with the field over the years gives the chapters real authority. The central chapters describe Lyon's preferred auditory model, which has two distinct stages: the first simulates the operation of the cochlea; the second simulates the conversion of the cochlear output into your initial auditory image of a sound in the neural centers between the cochlea and auditory cortex. The cochlear processing section is a transmission-

line filter bank that simulates basilar membrane motion with a "cascade of asymmetric resonators" (CAR). The gains of the resonators are continuously adjusted by a distributed AGC (automatic gain control) network whose action is applied separately to each CAR stage through the outer-hair-cell component of that stage. The resulting system exhibits the "fast-acting compression" (FAC) characteristic of auditory processing, as well as longer-term adaptation characteristic of mid-brain efferents. This stimulus-specific adaptation is intended to make machine hearing robust to interference in the way that human hearing is. The CARFAC model provides an accurate, stable simulation of cochlear processing across the full dynamic range of hearing—an enormous engineering achievement. These chapters are supported by some wonderful figures illustrating how the AGC network adjusts filter gain and shape across the complete set of CARFAC frequency responses as the level and content of a sound varies.

The neural processing section of the auditory model is relatively simple; it applies a form of "strobed" temporal integration (STI) separately to each channel of information flowing from the cochlear section of the model. STI automatically stabilizes sections of the neural activity that repeat, much as the trigger mechanism in an oscilloscope makes a stable picture from an ongoing time-domain waveform. The result for the complete set of cochlear channels is referred to as a stabilized auditory image (SAI)—a series of two-dimensional frames of real-valued data that form an "SAI movie" when presented in real time. Each frame is indexed by cochlear channel number on the vertical axis and "lag relative to strobe time" on the horizontal axis (see many examples in the figures in Chapter 21). The vocal sounds of animals (including speech) contain periodic segments that distinguish animate sources from environmental noises in the natural world, and the SAI presents a detailed, stable view of each repeating neural pattern for as long as it persists in the sound. In this way, STI and the SAI facilitate feature extraction and source segregation in everyday listening where the signals (speech, music, animal calls) are commonly mixed with interfering noises.

Together, the CARFAC cochlear model and the SAI encapsulate much of what we now know (and hypothesize) about auditory processing, and they provide a representation of sound that emphasizes the features and distinctions of everyday listening.

What is needed, then, is a digital version of the auditory brain that can put the auditory model to work in the service of machine hearing. This is the topic of the remaining chapters of the book. Lyon concludes that auditory scene analysis (ASA) and the algorithms used to perform computational ASA (CASA) are not, as yet, able to simulate the auditory brain, primarily because we do not understand the cortical processing behind the auditory brain. Similarly, he concludes that the neural networks commonly used in machine learning to train a nonlinear mapping from a large set of input patterns to outputs defined by a set of training data are unlikely to provide the basis for a successful model of the auditory brain, in this case because they are unlikely to be able to take SAI frames as input patterns due to the size of the frames and the frame rate. Some form of auditory feature extraction will have to be applied to the SAI frames to concentrate the auditory information in them and reduce the magnitude of the categorization problem for the machine learning systems used to implement machine hearing tasks. Lyon also believes that fine timing information is involved in the construction of human auditory features at a fundamental level, and that hearing machines will have to include fine temporal structure in some form or other.

This thinking leads to the intriguing idea of feature engineers and machine hearing engineers—people who use auditory knowledge, on the one hand, and knowledge about machine learning, on the other hand—designing mappings that convert auditory representations of sound with high dimensionality into forms that are suited to machine learning systems. Where possible, the engineers would identify auditory features that humans use and design algorithms to extract them from streams of SAIs. Lyon argues, however, that the development of machine hearing does not require the successful identification of the auditory features used by humans to solve listening problems. Rather, the engineer just needs to build a good interface between what we know about hearing and what we know about a machine learning system that might address the listening task. Indeed, it is argued that the mapping should not remove more information than absolutely necessary to

get the machine hearing task running. The machine learning algorithms might find nonintuitive features that actually perform better than the ones designed by a feature engineer to simulate human feature extraction. In summary, Lyon concludes that we will need to be careful about the problems we take on in the near future. We do not know enough about the auditory brain to simulate it. To make machine hearing a reality, we need intelligent mapping procedures to connect the very sophisticated CARFAC–SAI model of hearing to good learning machines—procedures that may, or may not, extract features the way humans do. This discussion of the options currently available to machine hearing engineers is fascinating, and his conclusions about how to proceed are very convincing.

Lyon is a great teacher and he has a deep understanding of the science and art of machine hearing. The reader will be greatly rewarded for engaging with any and all sections of the book.

— Roy D. Patterson, 2016
Cambridge, UK

Preface

If we understood more about how humans hear, we could make machines hear better, in the sense of being able to analyze sound and extract useful and meaningful information from it. Or so I claim. I have been working for decades, but more intensely in recent years, to add some substance to this claim, and to help engineers and scientists understand how the pieces fit together, so they can help move the art forward. There is still plenty to be done, and this book is my attempt to help focus the effort in this field into productive directions; to help new practitioners see enough of the evolution of ideas that they can skip to where new developments and experiments are needed, or to techniques that can already solve their sound understanding problems.

The book-writing process has been tremendous fun, with support from family, friends, and colleagues. They do, however, have a tendency to ask two annoying questions: "Is the book done yet?" and "Who is your audience?" The first eventually answers itself, but I need to say a few words about the second. I find that interest in sound and hearing comes from people of many different disciplines, with complementary backgrounds and sometimes incompatible terminology and concepts. I want all of these people as my audience, as I want to teach a synthesis of their various viewpoints into a more comprehensive framework that includes everything needed to work on machine hearing problems. That is, electrical engineers, computer scientists, physicists, physiologists, audiologists, musicians, psychologists, and others are all part of my audience. Students, teachers, researchers, product managers, developers, and hackers are, too.

The book's treatment of various aspects of hearing and engineering may be too deep for some, too shallow for others; many will find that something they know is missing, but hopefully all will also find useful things they didn't know. In particular, the system theory in Part II is taught with the aim of bringing this diverse audience to a common understanding of the math, physics, engineering, and signal-processing concepts needed to design, analyze, and understand the hearing models and applications taught in the later parts. Many aspects of the later parts of the book can be appreciated without mastering the system theory of Part II, but I recommend at least reading it through to get familiar with the terminology and to know where to refer later if more depth of understanding on particular points is desired.

Hearing has perhaps the most deep and elegant combination of linear and nonlinear aspects of any biological system. Readers will learn why the concepts of linear systems are so important in hearing, and also why these concepts are not nearly enough to explain hearing. Understanding nonlinear systems is always challenging, and we address that challenge by compartmentalizing the important nonlinearities of hearing into well-defined simple mechanisms that are individually not that hard to understand. We develop auditory models in terms of continuous-time systems, and implementations in terms of discrete-time systems with efficient implementations on computing machines; here again, having the nonlinearities compartmentalized is important.

The two aspects that best characterize the book's auditory models are ideas that I have pursued for many years, with many collaborators: the filter-cascade structure with embedded nonlinearities to model the cochlea; and the stabilized auditory image, or auditory correlogram, to capture and display the temporal fine structure in the signals that the cochlea sends to the brain. These two aspects are on opposite ends of the auditory nerve, and support my strategy to "respect the auditory nerve." We know so much from auditory

physiologists about the properties of sound representation on the auditory nerve, that to build models and systems that either do not produce or do not use the cochlear nerve's rich information about sound seems indefensible. The book shows some of the ways we have used such information productively.

The auditory models of Parts III and IV of the book are supported by open-source code, which should enable readers to get a good start on building machine hearing systems. Part V of the book introduces a very open-ended future of interesting applications, and I fully expect readers will become contributors to growth in this field of applications.

I mostly use the editorial "we" in the book, referring not only to myself as author but also to others who contribute to the ideas, including our readers. In a few places I switch to using "I," for more personal comments.

Though I paid my friends and colleagues a dollar for each bug or suggestion that I acted on, I owe them much more than that in thanks. Through their effort, the book has been much improved. I hope others will send suggestions for improving the next edition, and will earn a few dollars, too. I'm sure we have left some more errors for them to find.

On History and Connection Boxes

While there are historical comments, and comments on connections to related concepts in other fields, throughout many chapters, I have segregated some of them into boxes, both to highlight them and to keep them out of the way. In many cases, my aim is to honor the sources of the ideas we use, while trying to make the literature more accessible by saying a few words about how it connects. I trust that my mention of old technologies such as vacuum tube (valve) amplifiers and Helmholtz resonators and flame manometers will be received as intended: as clues to a very interesting heritage from generations of giants whose shoulders we stand upon, in both human and machine hearing.

My own EE training was in the era of transistors and early integrated circuits, when courses like "Circuits, Signals, and Systems" were all about analog continuous-time technology. In modern times, signals and systems are taught from the beginning with discrete-time concepts, for good reasons having to do both with pedagogy and the modern medium of implementation in digital computers. Although modern engineers may view sound naturally as the kind of discrete-time sampled data that they work with in computers, I have chosen to stick with continuous time as the primary conceptual domain in this work, since sound and the ear really exist in that domain. I hope that readers will not view the continuous-time domain as something out of history—it is the real world.

Online materials

Find errata, and links to code and other resources, at machinehearing.org.

Thanks

There are many people who have cared enough about this work to spend time helping and encouraging me. First among them is Roy Patterson, without whose encouragement I could never have even started, and who has continued to inspire me through the slow process.

Among my readers who have given me actionable feedback, Ryan "Rif" Rifkin stands out; he found me more bugs than everyone else combined. Others who contributed, whether by carefully reading chapters or giving feedback on overall impressions, include: Jont Allen, Peggy Asprey, Fred Bertsch, Alex Brandmeyer, Peter Cariani, Wan-Teh Chang, Sourish Chaudhuri, Brian Clark, Lynn Conway, Achal Dave, Bertrand Delgutte, Dick Duda, Diek Duifhuis, Dan Ellis, Doug Eck, Dylan Freedman, Jarret Gaddy, Daniel Galvez, Dan Geisler, Pascal Getreuer, Chet Gnegy, Alex Gutkin, Yuan Hao, Thad Hughes, Aren Jansen, James Kates, Nelson Kiang, Ross Koningstein, Harry Levitt, Carver Mead, Ray Meddis, Harold Mills, Channing Moore, Stephen Neely, Eric Nichols, Fritz Obermeyer, Ratheet Pandya, Brian Patton, Justin Paul, Manoj Plakal, Jay Ponte, Rocky Rhodes, David Ross, Mario Ruggero, R. J. Ryan, Bryan Seybold, Shihab Shamma, Phaedon Sinis, Jan Skoglund, Malcolm Slaney, Daisy Stanton, Rich Stern, John L. Stewart, Ian Sturdy, Jeremy Thorpe, George Tzanetakis, Marcel van der Heijden, Tom Walters, Yuxuan Wang, W. Bruce Warr, Lloyd Watts, Ron Weiss, Kevin Wilson, Kevin Woods, Ying Xiao, Bill Yost, Tao Zhang, and probably others that I have missed. Many thanks to all!

And finally, huge thanks to Lauren Cowles, my editor at Cambridge University Press, for her years of patience in helping to make this book happen.

Part I

Sound Analysis and Representation Overview

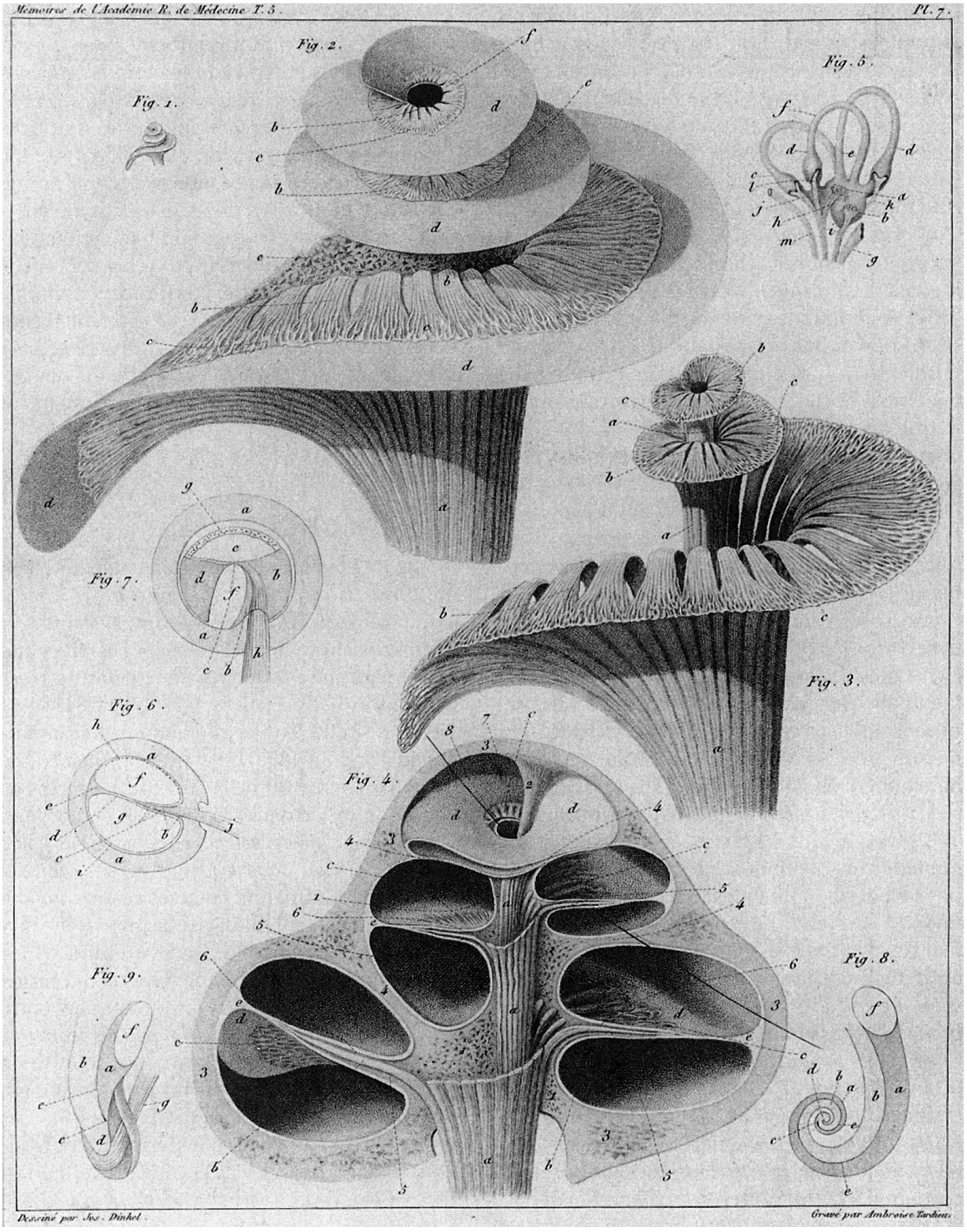
Part I Dedication: John Pierce

This part is dedicated to the memory of John Robinson Pierce (1910–2002). John was a dear friend and mentor for many years, beginning in my undergraduate years at Caltech. He gave me a summer job doing lab work on electronic musical instruments, and then on digital codecs that led to my first journal article. He persuaded his colleagues at Bell Labs to take me on as an intern, even after they had objected to my "less than an A in some important subjects." I owe my knowledge of digital signal processing to this great start with the early researchers and practitioners there. Pierce's work with George Zweig and Richard Lipes at Caltech, after I had left, became one of the most important influences on my thinking in hearing: the wave analysis that led to my filter-cascade approach to modeling the cochlea (Zweig, Lipes, and Pierce, 1976).

Pierce was better known for his work outside of hearing: from his early work in traveling-wave tubes and communication satellites at Bell Labs, his coining of the word *transistor*, his chief technologist role at the Jet Propulsion Laboratory, his science fiction writing under the pen name J. J. Coupling, through his enormous influence on computer music starting at Bell Labs and continuing at Stanford's Center for Computer Research in Music and Acoustics (CCRMA) in the 1980s and 1990s. His regular attendance at CCRMA's weekly hearing seminar provided a huge benefit to many of us in the hearing field. He continued to conduct and publish hearing research at Stanford even in his 80s, for example providing clarity on important issues in pitch perception (Pierce, 1991).

In Part I, we survey our concept of what the machine hearing field is, and how it relates to conventional acoustic approaches to sound processing and to a range of theories of hearing. We include a brief overview of human hearing from the conventional psychoacoustics and physiology points of view, which provide the data and some of the models that we build on.

Throughout the book, but especially in Part 1, I strive to make my point of view clear, describing the relationship of my conceptual framework and models to other concepts, old and new. Partly, this approach is to raise awareness about some older concepts that are still "hanging around," causing unneeded distraction and confusion. Equally importantly, it is to draw attention to ideas that still need more research and exploration, to see how well they hold up when experiments are designed specifically to test them. My hope is that this approach will help others find useful directions in which to extend, or to challenge, what I have gathered here.



Engraving of the structures of the cochlea and spiral ganglion by Gilbert Breschet (1836), before the discovery and description of the microscopic organ of Corti at the interface between the cochlea's ducts in 1851 by Alfonso Giacomo Gaspare Corti.

Chapter 1

Introduction

> . . . things inanimate have mov'd,
> And, as with living Souls, have been inform'd,
> By Magick Numbers and persuasive Sound.
>
> — William Congreve (1697) *The Mourning Bride*

> The ear is a most complex and beautiful organ. It is the most perfect
> *acoustic*, or hearing instrument, with which we are acquainted,
> and the ingenuity and skill of man would be in vain exercised to imitate it.
>
> — John Frost (1838), *The Class Book of Nature: Comprising Lessons on the Universe, the Three Kingdoms of Nature, and the Form and Structure of the Human Body*

Would it truly be in vain to exercise our ingenuity to imitate the ear? It would have been, in the 1800s—but now we are beginning to do so, using the "magick" of numbers. Machines imitating the ear already perform useful services for us: answering our queries, telling us what music is playing, locating gunshots, and more. By imitating ears more faithfully, we will be able to make machines hear even better. The goal of this book is to teach readers how to do so.

Understanding how humans hear is the primary strategy in designing machines that hear. Like the study of vision, the study of human hearing is ancient, and has enjoyed impressive advances in the last few centuries. The idea of *machines* that can see and hear also dates back more than a century, though the computational power to build such machines has become available only in recent decades. It is now, as they say in the computer business, a simple matter of programming. Well, not quite—there is still work to be done to firm up our understanding of sound analysis in the ear, and yet more to be done to understand the enormous capabilities of the human brain, and to abstract these understandings to better support machine hearing. So let's get started.

Humans tend to take hearing for granted. We are so aware of what's going on around us, largely by extracting information from sound, yet so unable to describe or appreciate how we do it. Can we make machines do as well at interpreting their world, and ours, through sound? We can, if we leverage scientific knowledge of how humans process sound.

Being able to produce and analyze sound waves is a prerequisite to developing a better understanding of hearing. Early progress in the field was made with the help of analytical instruments such as Helmholtz's resonators and recording devices, like the waveform drawing device in Figure 1.1, and controlled sound production instruments such as Seebeck's siren, shown in Figure 1.2. Representing such waves as electrical signals has been routine since the invention of the telephone. We now have a myriad of machines that help us generate, compress, communicate, store, reproduce, and modify sound signals, in ways tuned to how we hear.

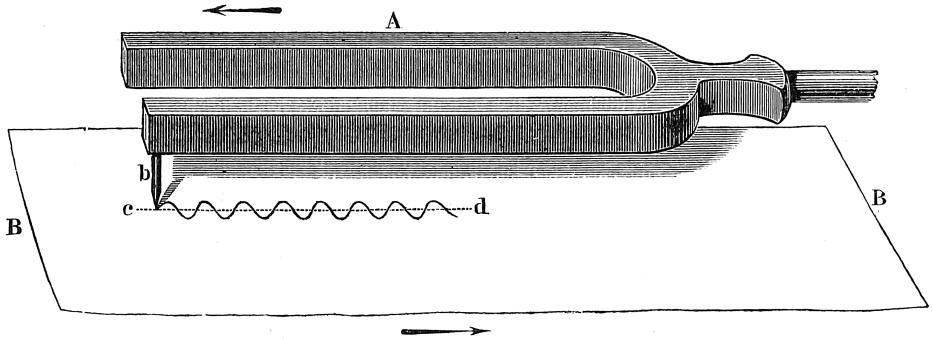


Figure 1.1: Helmholtz explained the idea of a sound's waveform via this diagram of a tuning fork with a stylus point attached, drawing its vibration on a moving piece of paper.

For most of these applications, though, the machines remain "deaf," in that they get very little meaning out of the sounds they process.

What if you had a device at home, always listening to what's going on? Could it tell what interesting things it heard while you were out? Could it tell you the refrigerator sounds like it's wearing out? Would it understand if you asked it a question? Could it find you some music to listen to if you described your mood? Could it listen to you and determine your mood itself? Could it say where a mouse might be hiding because it heard it run there? Could it distinguish between normal household sounds and an anomaly in the dead of night? Could it also be your intelligent answering machine, and tell you who called, and why, based on hearing their voice? Of course it could.

Who might make such a machine? What crazy functionality might they give a machine that could hear and understand sounds? Have we chosen the best path through the complex web of theories about hearing? Can we do better on some tasks by modifying the approach? What advances in the study of human hearing might we discover while trying to put our theories to the test of real use? These are the kinds of ideas and questions about sound and hearing that have been going around in my head for decades—and that we are getting some answers on recently. I've worked on spatial effects in music and games, and on machines to synthesize and recognize speech and music, and on other fun things to do with sound. Where most others deal with sounds by various conventional or ad hoc methods, I keep coming back to how the ear would do it—and this approach has proved fruitful.

There is enough known about how the ear and hearing work that we have gotten serious about putting this knowledge to practical uses. Starting with the anatomy, we model the structure and function of the ear and the auditory nervous system; using physiological and psychophysical techniques, we figure out what the brain gets from the ear, and how it deals with the information to perform meaningful tasks. Then we program computing machines to do similarly, based on this knowledge. In essence, we mimic the biology.

Today we have access to massive quantities of sound, to analyze, organize, index, and learn from. The soundtracks of YouTube videos alone have hundreds of millions of hours of sound, and so far our computers are rather ignorant of what those soundtracks are trying to communicate. Imagine what value there might be in having our machines just listen to them and understand. Speech, music, laughing babies, sounds of interesting events, activities, places, and personalities—it's all there to be discovered, categorized, indexed, summarized, remembered, and retrieved.

The full scope of machine hearing will reveal itself as people discover that it is relatively easy to have machines understand sounds of all sorts, and people find imaginative uses for such machines. Elephant infrasound hearing and bat ultrasound hearing and echolocation suggest that the same basic strategies have been put to many purposes by other mammals. We might include other sonic applications—such as medical

imaging—that use sound waves but don't rely on anything about sound perception. At Schlumberger Research in the 1980s, we experimented with hearing techniques applied to the analysis of underground sonic waves. Any far-out infrasound through ultrasound applications that can benefit from the use of techniques like those evolved by humans fall within the scope of what we're trying to teach via this book.

As we get more people engaged in machine hearing, there will be more good ideas and more things we can take on. The potential is enormous, and the scope broad.

1.1 On Vision and Hearing *à la* David Marr

The pioneering vision scientist David Marr was a big influence on my approach to modeling hearing. When I visited him at MIT in 1979 to show him what I was working on, he was very encouraging of the approach. Twisting his words, from vision to hearing, illustrates how his thinking influenced mine:

> What does it mean to hear? The plain man's answer (and Aristotle's, too) would be, to know what is where by listening. In other words, hearing is the *process* of discovering from sounds what is present in the world and where it is.
>
> Hearing is therefore, first and foremost, an information processing task, but we cannot think of it just as a process. For if we are capable of knowing what is where in the world, our brains must somehow be capable of representing this information—in all its profusion of color and form, beauty, motion, and detail.
>
> — modified from *Vision*, David Marr (1982)

I honor Marr's introduction to his ground-breaking book *Vision* in the quotation above, having changed *see* to *hear*, *looking* to *listening*, *vision* to *hearing*, and *images* to *sounds.* I've left the last phrase unchanged, as I believe that *"color and form, beauty, motion, and detail"* is a much more apt description of what our brains extract and represent about sound than the usual more pedestrian properties of *loudness, pitch,* and *timbre*.

Marr's computational and representational approach to vision helped to define the vibrant field of computer vision, or machine vision as it's also called, more than thirty years ago. My book is motivated by the feeling that something along these lines is still needed in the hearing field. It's a daunting challenge to try to live up to David Marr, even if I've had a few extra decades to prepare, but it's time to give it a shot.

Compared to other mammals, humans have put vision to some very special applications, like reading written language, and analogously have put hearing to use in spoken language and in music. These pinnacle applications should not exclusively drive the study of vision and hearing, however, and perhaps are best addressed only after low-level preliminaries are well understood, and more general applications are under control. Therefore, we focus on these more general and lower-level aspects, and on broader applications of hearing, as Marr focused on the more general aspects of vision. At the end, we come back and touch on applications in speech and music.

David Mellinger (1991) should be credited with helping drive this approach via his dissertation, pointing out that "Advances in machine vision have long stemmed from a physiological approach where researchers have been heavily influenced by Marr's computational theory. Perhaps the same transfer will begin to happen more in machine hearing." But this transfer has been incomplete, so we need to drive it some more.

Martin Cooke (1993) has provided an excellent review of Marr's approach to vision and its influence on work in speech and hearing. Marr's identification of three levels at which the sensory system is to be understood—*function, process,* and *mechanism*, also described as *computation*, *algorithm*, and *implementation*—certainly does help us organize our study of hearing. In an interesting twist, Peter Dallos (1973) used a

Figure 1.2: A make-it-yourself acoustic siren, much like August Seebeck's, as shown by Alfred M. Mayer (1878). The spinning disk, driven from a crank via string and pulleys, interrupts a stream of air from the tube to make waves of sound pressure that we hear as a tone. Different tones can be made by moving the tube to a different row of holes, or by changing the disk to one with a different pattern of holes. August Seebeck and Hermann von Helmholtz were among the nineteenth-century scientists who used such devices in their research that contributed to connecting the physical and perceptual properties of musical tones to the mechanisms of human hearing—though their theories were somewhat in opposition to each other.

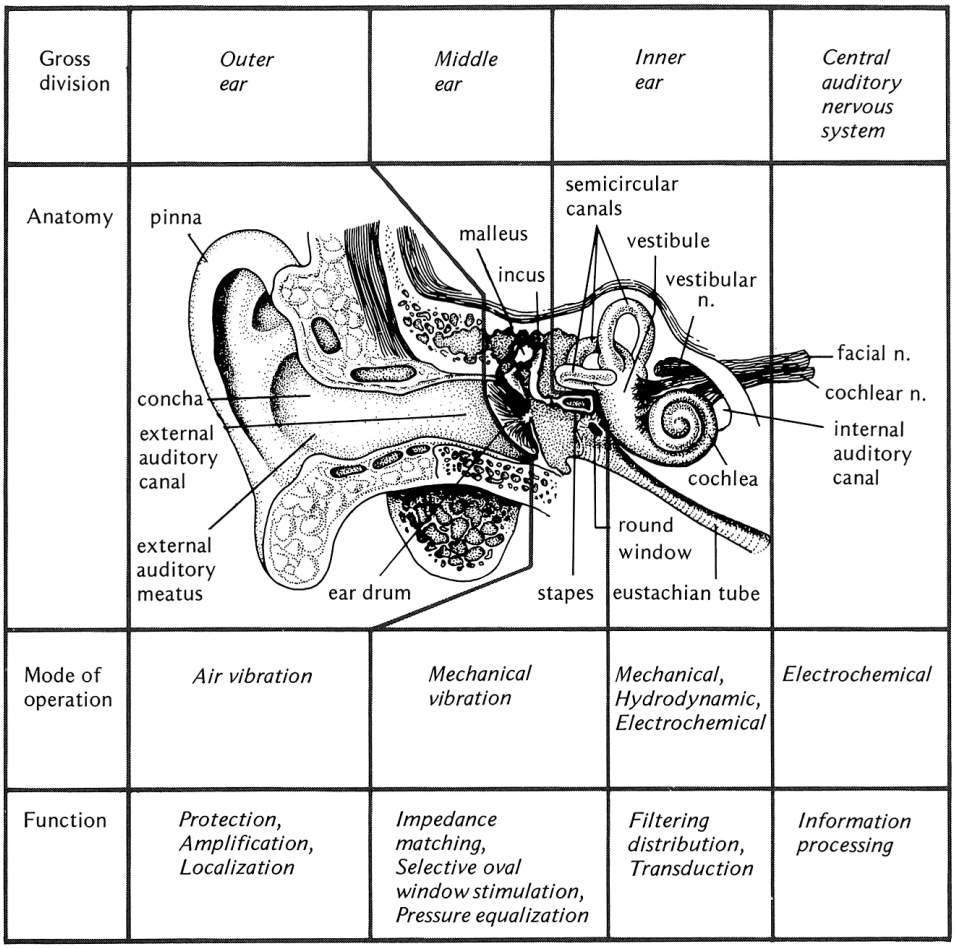


Figure 1.3: Ear diagram by Yost (2007). While the anatomy and modes of operation are important, we are most interested in emulating the *function*, described in the bottom row. The *information processing* in the central nervous system—the bit where meaning is extracted—is the part that remains most open to exploration and speculation. [Figure 6.1 (Yost, 2007) reproduced with permission of Wiliam A. Yost.]

similar division of concerns into function, mode of operation, and anatomy to describe the auditory periphery, before Marr's work. His scheme is still used this way and credited in current hearing books (Yost, 2007), as shown in Figure 1.3.

Cooke reviews several applications of Marr's levels and principles to speech processing, but provides relatively little connection to hearing. The repurposing of Marr's *primal sketch* concept into a *speech sketch*, by Green and Wood (1986), points up a disconnect: Marr didn't go from primitive images directly to reading, and we shouldn't go from primitive sound representations straight to speech; *primal* should imply a much lower level. A sketch is a "sparsified" version of an image, which may be used as part of a feature extraction strategy at the input to a learning system, as described in Section 25.7.

In vision, objects and images must be analyzed at many different scales. Referring to Marr, Andy Witkin (1983) said, "The problem of scale has emerged consistently as a fundamental source of difficulty, because the events we perceive and find meaningful vary enormously in size and extent. The problem is not so much to eliminate fine-scale noise, as to separate events at different scales arising from distinct physical processes." In hearing, we have the same issue, especially in the temporal dimension, where sounds have periodicities

and structure on all time scales.

The idea of an "auditory primal sketch" has been introduced by Neil Todd (1994) as a way to represent the rhythm and temporal structure of music and speech. I had published a related idea on multiscale temporal analysis, as part of a speech recognition approach (Lyon, 1987). Both of these are based on Witkin's scale-space filtering, which was descended from Marr. Both fall far short of a comprehensive framework for machine hearing, but help to inspire some of the sorts of representations that we will be working with.

Albert Bregman (1990), in his book *Auditory Scene Analysis: The Perceptual Organization of Sound*, discusses how aspects of hearing are valued from an evolutionary perspective, yielding certain advantages of hearing over vision. The auditory system evolved in a context in which better understanding of meaning from an auditory scene—better answers to *what* and *where*—led to a better chance of survival. When I refer to *human hearing* in my title, I mean to include the cortical-level processing systems that have evolved to handle speech, music, and other big-brain functions; but I do not mean to diminish the importance of the lower levels of auditory processing—in the ear, the brainstem, and the midbrain—that underlie the exquisite hearing capabilities of our pets (and pests), and that form the basis for robust representations of sound from which actionable information can be extracted. Even animals that don't normally use speech can learn to reliably recognize their own names, and discriminate them against other speech sounds; for example, Shepherd (1911) taught four raccoons that their names were Jack, Jim, Tom, and Dolly.

We can question Marr's insistence that a symbolic representation or *description* be generated (Hacker, 1991). Some approaches to machine hearing systems successfully use representations that remain completely abstract and nameless until the final output—the information that the system is trained to extract—with intermediate steps being subsumed in the learning system. Other approaches will use explicit and named concepts, such as objects, events, musical instruments, notes, talkers, and so forth, that artificial intelligence systems can reason with. Different theories of mind, or different computational frameworks that we have available, will bias our machine hearing applications one way or the other. We are not yet in a position to say which way is likely to be more fruitful for any given area, and hope to encourage exploration in all such directions.

Comments on hearing's analogy with vision are not new. For example, in 1797, the effect of auditory masking on sensitivity was observed and compared to visual masking effects in "annotations" on Perrole's "Philosophical Memoir" on sound transmission (Perrole, 1797):

> Sounds seem more intense, and are heard to a greater distance, by night than by day. . . . It is a practical question of some importance to ascertain whether this difference may arise from the different state of the air, the greater acuteness of the organ, or the absence of the ordinary noises produced in the day. By attentive listening to the vibrations of a clock in the night, and remarking the difference between the time when no other noise was heard, and when a coach passed along, it has appeared clear to me that this difference arises from the greater or less stillness only, and that no voluntary effort or attention can render the near sound much more audible, while another noise acts upon the organ. In this situation the ear is nearly in the state of the eye, which cannot perceive the stars in the day time, nor an object behind a candle.

In that memoir, Perrole also introduced the term *timbre* from the French to explain what he meant by *tone* in English: "The tone (*timbre*) was changed in the water in a striking manner." This "catch-all" term, as it has been called, captures everything about what a sound "sounds like," except for its pitch and loudness—sort of like *texture* in vision, which captures much of what shape, size, and brightness don't. It is the job of our machine hearing systems to map timbre (along with pitch and loudness and direction, and their evolution and rhythm over time) into useful information about what the sound represents, be it speech, music, environmental noises, or evidence of mundane or exceptional events.

Nomenclature: What to Call This Endeavor

The terms *computer vision* and *machine vision* are in wide use, not quite interchangeably, the former having a more computer-science connotation, and the latter a more industrial or applications connotation. Terms like *computer hearing*, *computational hearing*, and *computer listening* seem awkward to me, especially since I spent a lot of years building analog electronic models of hearing, probably not qualifying as computers. And what about *listening* or *audition* as a better analogy to *vision*? Several of these terms have overloaded meanings: we can convene a hearing, or perform in an audition, or plant listening devices. The term *machine listening* is sometimes used, but mostly in connection with music listening and performance.

The term *machine hearing* has a strong history at Stanford's computer music lab, CCRMA. In their 1992 progress report, Bernard Mont-Reynaud (1992) wrote a section on machine hearing, which noted that "The purpose of this research is to design a model of Machine Hearing and implement it in a collection of computer programs that capture essential aspects of human hearing including source formation and selective attention to one source (the 'cocktail party problem') without tying the model closely to speech, music, or other domain of sound interpretation."

We hope that by calling the space of computer applications of sound analysis *machine hearing*, following Mont-Reynaud, we will leverage this good name and good direction, and help the field build around a good framework, as Marr did with what we refer to as *machine vision*.

1.2 Top-Down versus Bottom-Up Analysis

Top-down processing evaluates sensory evidence in support of hypothesized interpretations (meaning), while bottom-up processing converts sensory input to ever-higher-level representations that drive interpretation. Real systems are not necessarily at either extreme, but the distinction can be useful.

Marr says, with respect to general-to-specific (or coarse-to-fine) stereo matching approaches (Marr, 1982),

> This type of approach is typical of the so-called top-down school of thought, which was prevalent in machine vision in the 1960s and early 1970s, and our present approach was developed largely in reaction to it. Our general view is that although some top-down information is sometimes used and necessary, it is of only secondary importance in early visual processing.

Here we totally agree. Although I have nothing but respect for the strong case for the power of top-down information and expectations in human hearing (Slaney, 1998; Huron, 2006), and though there are prominent "descending" pathways at all levels of the auditory nervous system (Schofield, 2010), my understanding is that the more extensive and complex feedback is within the cortical levels of the central nervous system, and that early audition, like early vision, is best conceived as a modular set of mostly feed-forward bottom-up processing modules. There is feedback, to be sure, but its function can often be treated as secondary, as Marr says. At some levels, feedback may be about parameter learning and optimization; from cortex to thalamus, top-down projections may be about attention. These are important, but not where we start, especially in "early" layers as Marr says.

In the mammalian brain, these early hearing modules include the periphery (the ear) as well as auditory structures in the brainstem and midbrain, and maybe even some stages of cortical processing, such as primary auditory cortex. These levels were successful stable subsystems long before the evolution of the big neocortex that led to speech and music. The "near decomposability" condition (Simon, 1981) is what allows complex systems to evolve. That's why we rely so much on data from bottom-up experiments in animals to help us understand human hearing; we accept that the amazing abilities of humans evolved on top of these stable mammalian subsystems, which are themselves not so different from reptilian, bird, and even fish auditory systems.

Like Marr, we are partly reacting to an overreliance on top-down information in sound processing systems. For example, automatic speech recognition (ASR) systems have been gradually improved over the years by reliance on larger and more complex language models and by statistical models that can capture complex prior distributions, while their front-end processing remains relatively stagnant, stuck with spectro-temporal approaches that have no way to improve in terms of robustness to noise and interference, since they don't represent the aspects of sound that help our auditory systems tease sound mixtures apart. Such problems demand that we understand hearing better, and build systems that can hear and understand multiple sounds at once; how else can we expect a speech recognizer to give us a transcript of a boisterous meeting? Of course, good prior distributions from top-down information will continue to play an important role, too.

Is the auditory system *complex*? Herb Simon (1981) characterizes a complex system this way:

> In such systems, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.

I think this applies to the auditory system as a whole, when the cortex is included, especially in a living organism in which the auditory system is interacting with visual, motor, and other systems, with strong top-down and feedback effects. But for the various bottom-up modules of lower-level auditory processing, perhaps the system is merely *complicated*, but not so complex that we can't describe its function, and its process, in terms of its mechanisms. I think this is how Marr saw early vision, too. Otherwise, it would be hard to be optimistic about our ability to assemble machines to do similar jobs.

1.3 The Neuromimetic Approach

A strategic element of our machine hearing approach is to respect the representation of sounds on the auditory nerve, which involves both a *tonotopic* (arranged by frequency) organization and detailed temporal structure, as extracted by the rather nonlinear inner ear. At this level, the approach can be said to be *neuromimetic* (Jutten et al., 1988), or *neuromorphic* (Mead, 1990), in the sense that we may be building a copy of a complicated neural system, mimicking its function—or mimicking its structure when we can't quite describe the function. In the neuromorphic case, copying the structure of the neural system, the expectation is that the structure will have an appropriate *emergent behavior* and therefore a useful information-processing function. Here *emergent* means that the behavior is not explicitly designed in, but *emerges* from the simpler behaviors of the lower-level elements as a consequence of the structural pattern of interconnection of those elements (Bar-Yam, 1997).

This neuromimetic approach is somewhat distinct from the Marr approach, but sometimes a useful supplement. When a system built this way is found to have a useful function due to its emergent behavior, it can sometimes be further analyzed, and the important parts of its function abstracted, described, and reengineered more efficiently. I believe we are part of the way through this process with neuromimetic hearing front ends. At the level of the cochlea, for example, the function is largely understood, but the description is still as much structural as functional. We do not have the clean separation of function, process, and mechanism that Marr recommended, but we do have a structure for which we can understand the function.

Beyond the cochlea, we still have a mixed structural and functional view, though it is somewhat speculative, of what the function is—the little "information processing" box in the lower right corner of Bill Yost's diagram, Figure 1.3, is where we ultimately extract meaning. We have pretty good ideas from physiological data about what kinds of auditory images are formed in the brainstem. The main thing we use that is neuromorphic is the very idea of an auditory image: a neural pathway with two spatial dimensions, like the optic nerve from the retina, projecting a time-varying pattern to a two-dimensional sheet of cortical tissue, the primary auditory cortex, for further processing.

An early proponent of a neuromimetic, or *bionic*, approach to machine hearing systems was John L. Stewart (1963), who published a number of reports, papers, patents, and a book on the topic in the 1960s and 1970s. He explains the reasoning behind this approach (Stewart, 1979):

> The model becomes an intermediary—a surrogate reality. ... It is my belief that effective explanations for the traits of living organisms demand the construction of models which behave as do their living counterparts. For, in no other way can the research be disciplined to produce an effective holistic theory!

Stewart (1979) anticipated much of our current approach, including a cochlear transmission-line analog with nonlinearities, a "neural-like analyzer" stage following the cochlea (Stewart, 1966), and the idea of efferent (feedback) adaptation to conditions, via coupled frequency-dependent gain control (Stewart, 1967).

1.4 Auditory Images

In our approach to hearing, we incorporate the notion of an *auditory image*: a presumed representation developed in the subcortical parts of the auditory nervous system (cochlea, brainstem, and midbrain), projecting to primary auditory cortex in the same way that the retinal image projects to primary visual cortex. This approach brings together the strategies of Marr with the two-dimensional neural circuits of the *place theory of sound localization* of Jeffress (1948) and the *duplex theory of pitch perception* of Licklider (1951).

A *spectrogram* is a picture of sound on a time–frequency plane. But this two-dimensional image is not what we call an auditory image, as it has too few dimensions to be analogous to the image that the eye sends to cortex. In the spectrogram, one axis is time, and there is only one other axis (frequency, mapped to spatial location). To make auditory images, we develop one more dimension, to map to a spatial axis orthogonal to the frequency axis, resulting in a movie-like representation, an image that changes with time. This added spatial dimension can represent direction (lateral or azimuth direction of arrival of a sound) in a binaural auditory image like those of Jeffress, or can represent pitch period and other temporal texture as in Licklider's duplex images. But these are just examples, not the limit of what an auditory image might be.

A possible next (cortical processing) step is to reduce the auditory image to a *sketch*, or line drawing, as Marr does, but that is not the only approach.

Our study of hearing will necessarily involve a lot of function, process, and mechanism to arrive at auditory images, corresponding mostly to levels below primary auditory cortex. This complicated architecture is a bit different from the vision case, where the information starts as an optical image that makes a 2-D response image on the retina, and further processing is mostly in cortex. Even in secondary and later levels of auditory and visual cortex, much of the mammalian brain's processing is about what and where, and only humans, with huge areas of more highly evolved cortex, implement the much higher levels of interpretation that support language and music (Rijntjes et al., 2012).

Marr was very much in touch with the developing sciences of visual psychology and visual neurophysiology, which informed his approach, especially at the level of multiscale edge analysis in visual cortex, on which he modeled his primal sketch. Similarly, our approach to machine hearing draws on the fields of auditory psychology and physiology, where so much is known about many levels of hearing, and where I've been so lucky to know and interact with so many of the great scientists over the last several decades. Part of our goal with this book is to help these fields in return, by providing a conceptual framework in which much of their detailed knowledge can find a place, and be better understood and promulgated in terms of signal processing, information extraction, and sound understanding.

The physiological data informing this approach are from animal studies, in mammals, birds, reptiles, and other groups. Most of the auditory brainstem and midbrain was already stable before the mammals split

off from the reptiles, so studies in many animals contribute to our understanding of human hearing, and are included in our scope. For example, the notion of auditory images as a representation of objects in space, as extracted from binaural (two-ear) signals, has been well developed to describe the function and organization of the auditory nervous system in the barn owl (Konishi, 1995). We humans may not swoop down and catch mice in the dark, but we do have an auditory spatial sense that's not so different from that of the barn owl, using very similar structures in our brains.

1.5 The Ear as a Frequency Analyzer?

At the functional level of description, it can be hard to say what the ear is doing. A traditional view is that the *cochlea* in the inner ear acts as a *Fourier analyzer* or *frequency analyzer* (Gold and Pumphrey, 1948; Plomp, 1964). We believe that as a top-level functional description, that's often misleading. One goal of this book is to help displace this view with a better description of the kind of information the ear sends to the brain.

In the late nineteenth century, it was not unusual to find statements such as "the function of the cochlea is to determine the pitch of the sound" (Draper, 1883), or "the function of the cochlea is to receive and appreciate musical sounds" (Murché, 1884). Generally, the cochlea was interpreted as a frequency analyzer. A few interpretations were a bit broader, with statements like "the function of the cochlea is to appreciate the *qualities* of sounds" (Bale, 1879).

The simple frequency view was largely derived from Helmholtz (1863), though his book on the subject was much more thoughtful than these simplifications. He did address function head-on, but his book was about connecting hearing to music, so he can't be faulted for describing the function in relation to musical tones:

> Hence the ear does not distinguish the different forms of wave in themselves, as the eye distinguishes the different vibrational curves. The ear must be said rather to decompose every wave form into simpler elements according to a definite law. It then receives a sensation from each of these simpler elements as from an harmonious tone. By trained attention the ear is able to become conscious of each of these simpler tones separately. And what the ear distinguishes as different qualities of tone are only different combinations of these simpler sensations.

This phase-blind frequency-analysis view of hearing had originally been articulated by Georg Ohm (1843), inspired by Joseph Fourier's 1822 finding that periodic functions could be described as sums of sinusoids. While the idea does have some merit as a model of hearing, it is also easily found to disagree with various experiments, so has often been regarded as a half-truth, or sometimes worse, as in this statement by W. Dixon Ward (1970):

> For years musicians have been told that the ear is able to separate any complex signal into a series of sinusoidal signals—that it acts as a Fourier analyzer. This quarter-truth, known as Ohm's Other Law, has served to increase the distrust with which perceptive musicians regard scientists, since it is readily apparent to them that the ear acts in this way only under very restricted conditions.

Ohm's and Helmholtz's view of hearing as Fourier analysis, and the confusion of frequency with pitch, continued to permeate, if not dominate, thinking about hearing in the early twenty-first century, even though problems with the approach had been repeatedly demonstrated, and arguments against it published continually over a century and a half.

August Seebeck (1841), using his acoustic siren, demonstrated several effects that were hard to explain in Ohm's model. In fact, Ohm published his law in response to Seebeck's first paper in 1841, and they engaged

in a back-and-forth in print for a number of years. Helmholtz later sided with Ohm, and tried to explain Seebeck's results in his book (Helmholtz, 1863) in a way that would resuscitate Ohm's point of view. These disputes have been frequently recounted (Scripture, 1902; Jungnickel and McCormmach, 1986; Cahan, 1993; Beyer, 1999), so we don't need to go into detail here. Heller (2013) has a particularly cogent discussion of the evolution of the thinking of Seebeck, Ohm, and Helmholtz, as influenced by Fourier's mathematics (and it is a great undergraduate-level book on sound and hearing in general).

Many modern papers and books sidestep the description at a functional level, with sections entitled "the function of the cochlea" typically describing lots of phenomena, process, and mechanism, but with very little commitment to an idea of function. Statements of function are sometimes made, but are kept very general and conservative, such as "The primary function of the cochlea is hearing" (Van De Water and Staecker, 2006), and "The function of the cochlea is to convert the vibration of sound into nerve impulses in the auditory nerve" (Cook, 2001), and "the essential function of the cochlea can be conceptualized as a transduction process" (Phillips, 2001). Some invoke the traditional Fourier analyzer concept, as in "Its principal role is to perform a real-time spectral decomposition of the acoustic signal in producing a spatial frequency map" (Dallos, 1992).

In a very few cases, we find a bit about capturing the quality of sound and something about temporal properties, as in "The main function of the cochlea is to translate auditory events into a pattern of neural impulses that precisely reflects the nature and timing of the sound stimulus" (Probst et al., 2006). This concept is better, especially in being tied to general properties of the sound instead of to narrower musical properties based on frequency. We need this kind of more general functional thinking if we're going to process arbitrary real-world sounds—the kinds of sounds for which hearing evolved, long before music and speech came along.

An important function of the cochlea that is often missed in functional characterizations has recently been given first-class status: loudness compression. Jont Allen (2001) says:

> The two main roles of the cochlea are to separate the input acoustic signal into overlapping frequency bands, and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell.

Allen's conceptualization of function is a much better starting place, and explains part of why nonlinearities are so important in hearing. A proper focus on function will be key to our progress in machine hearing. In support of the function "to separate the input acoustic signal into overlapping frequency bands," we discuss the progression from Fourier analysis, to short-time Fourier analysis, to linear bandpass filterbanks; and in support of the function "to compress the large acoustic intensity range," compressive nonlinear filterbanks. We further connect filterbanks to filter-cascade structures, to make a more realistic relationship of the filtering function to the underlying mechanisms. Part II of the book develops the necessary systems theory, and Part III applies these concepts to develop good computational models of cochlear function.

1.6 The Third Sound

The importance of nonlinearity is not yet well integrated into the typical understanding of the functions and processes of hearing. One of the earliest phenomena to bring the problem to the attention of scientists was the *third sound*, observed by Sorge (1745) (*den dritten Klang*) and by Tartini (1754) (*un terzo suono*). This third sound is a low-pitch tone heard when two other tones are sustained, for example by two horn players; pitches of such tones are illustrated in Figure 1.4. It turns out to be usually a pitch equal to the difference of the pitches of the first two tones or of some of their harmonics, and is what we call a *combination tone*, a *difference tone*, or a *distortion product*.

We'll see that there are good reasons for the existence of several types of nonlinearities in hearing, and for modeling them in machine hearing systems. But before we tackle nonlinearity, we have to understand what

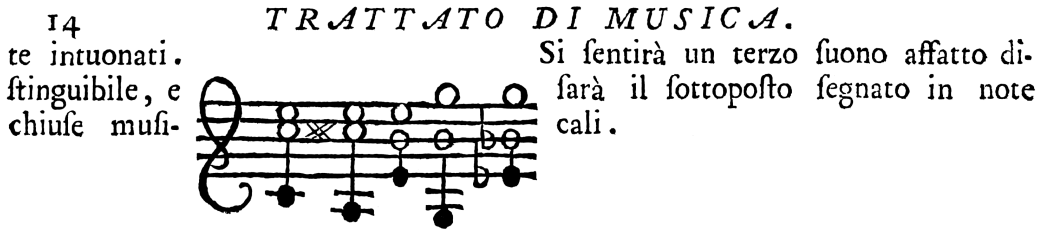



Figure 1.4: Tartini's 1754 publication of his observation of *un terzo suono*, a third sound, shown as filled notes below the first two sounds playing on violins or horns—among the earliest recognitions of a nonlinear effect in hearing. The note pitches that Tartini illustrated represent the ratios 4:5:2, 5:6:2, 3:4:2, 5:8:2, and 3:5:2 ($f_1 : f_2 : f_3$, for f_1 being the pitch of the lower played sound and f_2 being the pitch of the upper one, and f_3 being the pitch of the low third tone). The third-tone pitch corresponds to the quadratic intermodulation product $f_2 - f_1$, or the cubic intermodulation product $2f_1 - f_2$, and/or an octave above one of those. As Helmholtz (1863) remarked of these observations, "It is very easy to make a mistake of an octave. This has happened to the most celebrated musicians and acousticians. Thus it is well known that Tartini, who was celebrated as a violinist and theoretical musician, estimated all combinational tones an octave too high." Sorge's 1745 observation of c″ and a″ making an f would be 3:5:1, with *den dritten Klang*, a third-order (cubic) distortion product, at $2f_1 - f_2$.

linear systems are, and how such systems give rise to sinusoidal analysis. We'll cover the theory of linear and nonlinear systems in Part II, and apply them in subsequent parts of the book.

1.7 Sound Understanding and Extraction of Meaning

We conceptualize the machine hearing space as *sound understanding*, or *information extraction*, or *extraction of meaning*, in a very general sense. Here *understanding* signifies extraction of actionable information, as is sometimes implied in *speech understanding* systems as distinguished from *speech recognition* systems. That is, it means that from a sound we are able to provide useful information for some practical application.

It's not just humans and machines that do this—my dog is pretty good at processing sounds, too. If her practical application is to greet someone at the front door, she gets the information she needs from the sound of either a knock or the doorbell. For the application of when to eat, she recognizes the sound of her dish being set down. She's pretty clever about learning the sound cues for when she'll be taken for a walk, and other things she cares about. Does she *understand* sounds? Yes—in the same sense that humans do, and that machine hearing systems do: from sounds, she extracts what she needs to know.

If we can make machines hear half as well as my dog does, that will be progress. Humans are involved because we want to build up to where we can replicate a human's ability to extract information from speech, music, video soundtracks, and the everyday environment that humans live in. And humans provide a wealth of psychophysical experimental data that can be leveraged in the design of machine hearing systems.

Winnie-the-Pooh has introspected on the extraction of meaning from sound (Milne, 1926):

> "That buzzing-noise means something. You don't get a buzzing-noise like that, just buzzing and buzzing, without it meaning something. If there's a buzzing-noise, somebody's making a buzzing-noise, and the only reason for making a buzzing-noise that *I* know of is because you're a bee. ... And the only reason for being a bee that I know of is making honey. ... And the only reason for making honey is so as *I* can eat it."

How did Pooh interpret a "buzzing-noise" as indicating the availability of honey? We interpret this question as having two main parts: first, analyzing and representing sound in such a way that this "buzzing-sound" is distinguishable from other sounds; and second, learning one or more decision functions that address the question of when and where food might be available, based on the sound present. The connection from "buzzing-noise" to food is probably the result of a fairly opaque learned decision function, in a brain or a hearing machine; Pooh's semilogical chain of reasoning should probably be regarded as a *post-hoc* rationalization of the decision, not an explanation of how the decision was arrived at. It seems likely that at this level of abstraction, humans and other mammals probably perform such functions about the same way as Milne's anthropomorphized fictional characters do.

When decisions are reached, and those decisions are useful, then we can say that meaning, or information, has been extracted from the sound. Sometimes the meaning is more indirect, as by inference from the linguistic content carried by words in the sound of speech. In speech recognition, we can say that meaning has been extracted when the recovered word sequence serves to further the successful execution of a task.

1.8 Leveraging Techniques from Machine Vision and Machine Learning

At the applications end of machine hearing, there are many overlaps of problems, and techniques, with other fields. Therefore, we have many opportunities to leverage techniques that have been developed in those fields. In particular, machine vision and machine learning, especially as applied to problems in situations involving both images and sound, whether live or recorded, give us a good set of tools to apply. Leveraging these much larger fields is a key part of our strategy in trying to bring the field of machine hearing forward.

The machine vision field gives us a number of successful feature extraction approaches, and trainable system structures, some of which will map well into hearing problems. In systems such as video analysis, or surveillance, where both vision and hearing can be applied together, we have opportunities to *fuse* information from the different senses, on the way to the extraction of meaning. Even the simple concatenation of sound features onto image features has already been shown to improve the performance of video classification systems (Gargi and Yagnik, 2008); they may still be half blind and "hard of hearing," but they're no longer completely deaf.

1.9 Machine Hearing Systems "by the Book"

After we survey a range of conventional and novel sound analysis and representation techniques in Part I, we review in Part II the linear system theory that explains why the idea of analyzing sounds into frequencies, or overlapping frequency bands, makes sense, and how important nonlinear concepts such as compression need to be integrated into that view.

In Part III, we go on to apply that concept at other levels of description, culminating in a model of the cochlea that runs as an efficient machine algorithm for processing sounds into a representation that respects what we know about signals on the auditory nerve.

Part IV of the book attempts to do the same for the next levels of processing, in the lower parts of the auditory nervous system: to provide a functional concept, and an efficient process and mechanism that will extract the "auditory image" sound representations needed by the higher levels of hearing, to connect to the information that applications need to understand sound.

In Part V, we get into applications, which we can think of as paralleling the uses to which humans apply the information they extract from sound. We may not yet know enough about the function of neocortex to really leverage that knowledge for building intelligent machines, so at the application level we turn mostly to techniques we understand better, from the field of machine learning. We use various methods to convert the

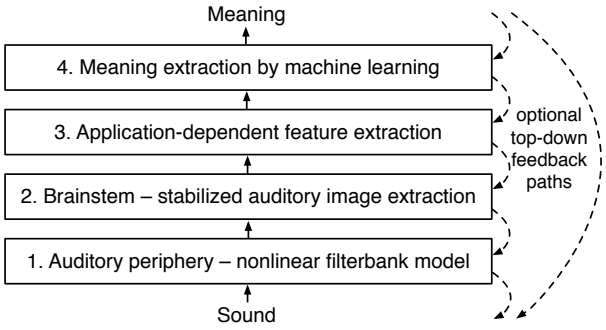


Figure 1.5: The *four-layer model* of machine hearing systems developed in this book—from sound to meaning, and sometimes back the other way. The big feedback loop from meaning to sound is for a system that can make sound and hear itself, for example, a speech conversation system.

sound representations from the subsystems in Parts I and III into the kinds of features that machine learning systems can easily use, and from there we train transformations that extract the information we want. None of this has much to do with frequency analysis, so we should be careful to not let that concept dominate our thinking about the ear.

Our book develops the idea of a machine hearing system made of four modules, or layers; from the bottom up, as illustrated in Figure 1.5 and detailed here:

1. A model of the cochlea, or auditory periphery, built as a cascade of nonlinear filters, as developed in Part III;

2. A model of the auditory brainstem, extracting one or more auditory images appropriate to the range of sounds and tasks to be addressed as developed in Part IV;

3. A feature extraction layer to convert auditory images to a form more suited to the particular application and tailored to the machine learning system chosen, as developed in Part V;

4. A machine learning system that is trained to extract the kind of decisions or *meaning* needed for the target application, as addressed also in Part V.

This layering will focus us on a known-working and factored structure, based closely on human hearing where possible, not specific to the higher-level properties of speech and music, that is open-ended enough to allow expansion into arbitrary applications. From the point of view of many applications, such as speech recognition, most of the action is at the top, in level 4, and the lower three levels just make a black-box front end. The challenge there will be to make sure that the features that come out of level 3 are what the recognizer needs.

Our machine hearing systems are characterized by several special features, in the first two modular layers: the cascade filterbank structure with nonlinearities, and the auditory image approach. Hence, much of our emphasis is on developing an understanding of these hearing-based ideas and their historical precedents, in the corresponding book parts.

These special features are not new or radical, but are not yet widely enough appreciated and used in hearing systems. Both were discussed in the middle of the twentieth century. The notion of a cascade as an alternative to the more common parallel-resonator filterbank was presented by Licklider (1956) as a model of cochlear filtering. He also adopted what we now call auditory images in his “duplex theory of pitch

perception" and combined this approach with Jeffress's "place theory of sound localization," to form his "triplex theory" of pitch perception (Licklider, 1956):

> ... It outlines a mechanism that accounts for the three ways in which acoustic stimulation can give rise to subjective pitch and, at the same time, brings into mutual relation a number of facts from other parts of auditory experience. ... If the aim is to understand the process of perception, the inquiry must extend into the higher centres of the brain. At the present time, this is sure to lead one into speculation. However, if there is a lack of anatomical and physiological facts, there is an abundance of psychophysical ones.

It took a few more decades for the understanding of the auditory nervous system, and of the cochlear nonlinearity, to evolve. Auditory image maps in the auditory nervous system are now known and being actively investigated (Knudsen, 1982; Sullivan and Konishi, 1986; Schreiner, 1991; Langner et al., 1997; Velenovsky et al., 2003), as discussed in Part IV. Incorporating appropriate nonlinearities into the cascade filterbank, with results reflected in the auditory image, is straightforward, once the different types of nonlinearity are understood, as emphasized in Part III.

Pierce and David (1958) commented on the fact that different types of meaning extraction involve different types of processing:

> Undoubtedly the nervous system uses a multiplicity of methods in dealing with the range of auditory stimuli presented to it. We don't "perceive" vowels in the same way as gunshots. A machine to emulate the nervous system in these functions would be an intelligent machine indeed. Can we ever understand enough to make such a machine? Before science can answer unequivocally it must look farther, directly or indirectly, into both the problems involved in such recognition and the way in which human beings manage to solve them.

These differences, which we now know more about from studies of psychoacoustics and of the nervous system, would be reflected in the application-dependent feature extraction layer, where we extract different features to localize a gunshot than to classify a vowel, and in the trainable decision system in the final layer.

On the prospects of building machines to do it, Pierce and David (1958) knew it would be a long hard road:

> We have already taken the first few faltering steps toward building machines which will respond to and correctly interpret the sounds of speech. Through further diligent work it may indeed become possible to construct devices which will respond to and reply in human speech, perhaps even to make useful voice typewriters, and maybe, later on, to build that staple of science fiction, a device which translates spoken words of one language into spoken words of another. Whether such machines are ever actually built will depend upon how complex they need be; and, in essence, how much human time, effort, and money man is willing to expend in simulating human functions.

But this was more than a half century ago; the value of building such systems is now generally known to exceed the costs in many application areas. It would be good to reflect on the various dimensions of progress since then, as well as on the remaining difficulties, as we set out to build more such machines.

In proceeding to build such machines, we will learn much about hearing. I hope the first lesson has sunk in already: sounds are not just sums of tones of different frequencies, and the ear is not a frequency analyzer.

Chapter 2

Theories of Hearing

> In respect of the theory of hearing, it seems to me that we need fewer theories and more theorizing. Of theories, focused upon some new finding and seeking to align the entire body of auditory fact with the new principle, we have more than a plenty.
>
> — "Auditory theory with special reference to intensity, volume, and localization," Edwin G. Boring (1926)

> The principle of diversity suggests that a simple description of the auditory process may not be possible because the process may not be simple. Theories that appear at first thought to be alternatives may in fact supplement one another.
>
> — "Place mechanisms of auditory frequency analysis," William H. Huggins and Licklider (1951)

Many theories and models have influenced thinking in this field; here we survey some of these, including those modern theories on which we base machine hearing systems.

2.1 A "New" Theory of Hearing

Books and papers entitled "A New Theory of Hearing" or something to that effect were once almost commonplace (Rutherford, 1887; Hurst, 1895; Ewald, 1899; Meyer, 1899; Békésy, 1928; Fletcher, 1930; Wever and Bray, 1930b; Wever, 1949). Like many ideas from a few generations back, some of these theories seem a bit quaint from our modern perspective. But in many cases they really did represent some of the most insightful scientific thinking and freshest experimental observations of their times. We review some of these ideas here, emphasizing those that left a lasting mark on our thinking about how hearing works.

Hermann von Helmholtz's *Tonempfindungen* (Helmholtz, 1863) presented the first major influential theory of hearing. His theory that structures in the cochlea vibrate sympathetically, each place resonating with its own narrow range of frequencies to stimulate a specific nerve, was the foundation for the long-lasting concept of the ear as a frequency analyzer. The arrangement of nerves in the cochlea was associated with individual just-distinguishable tone frequencies, adapting Müller's *doctrine of specific nerve energies* (Müller, 1838) and applying Fourier's finding that any periodic signal is equal to a sum of sinusoids of harmonically related frequencies (Fourier, 1822). The idea that the nerve signal could represent the intensity of each resolved sinusoid didn't leave a place to represent their relative phases, but that was OK with Helmholtz, because Georg Ohm had already articulated his law that the sound of a tone depends only on the amplitudes (also known as magnitudes) of the components, irrespective of the phases (Ohm, 1843).

This theory essentially says that a perceived pitch corresponds to a *place* of maximal resonant response, and that all other aspects of tone quality and more complex sounds are captured in the *spectrum*. Such theories are called *resonance theories*, or *place theories*.

But many disliked Helmholtz's conception, and were not afraid to say so (Perrett, 1919):

> When ... I foretold a great fall for Helmholtz and his book I little suspected that the prophecy would be so soon fulfilled, by the publication of Sir Thomas Wrightson's *Inquiry into the Analytical Mechanism of the Internal Ear*, 1918. Now the case is altered. The wilderness in which I whispered to the reeds the oppressive secret, "— hath —'s Ears," has suddenly, through a feat of invisible engineering long since planned, become populous, and no less an anatomical authority than Professor Arthur Keith has proclaimed the crudity and impossibility of the Helmholtz theory of hearing. ... The present chapter underlines that proclamation, bringing linguistic proof that there cannot be any resonators in the internal ear acting like "a kind of practical Fourier's theorem." The physicists (some of them) must be less superstitious.

The inquiry that Perrett refers to (Wrightson, 1918) develops an elaborate theory based on reflection and coincidence of waveforms along the cochlear partition; it is hardly remembered today, but was one of many attempts to find a better explanation of how the ear analyzes and represents sounds.

Pitch is the one aspect of sound that since ancient times was already widely used and understood at some level, from its role in music, including melody, consonance and dissonance, and the construction of musical instruments. By the time of the nineteenth-century theorizing about hearing, pitch had long been associated with rates of vibration, not just as ratios but even calibrated to vibrations per second. Marin Mersenne had estimated the speed of sound and corresponding frequencies of musical pitches of organ pipes in the early seventeenth century, and Joseph Sauveur had improved on his estimates in the early eighteenth century (Beyer, 1999). It was natural that investigations of hearing focused on pitch.

William Rutherford, a Scottish physiologist, was one of many who had a hard time believing that the cochlea could have thousands of distinct resonators for all the distinguishable pitches, and proposed instead a new theory of hearing based on the working of a telephone, a then-recent technological hit (Rutherford, 1887):

> The theory which I have to propose may be termed the Telephone Theory of the Sense of Hearing. The theory is that the cochlea does not act on the principle of sympathetic vibration, but that the hairs of all its auditory cells vibrate to every tone just as the drum of the ear does; that there is no analysis of complex vibrations in the cochlea or elsewhere in the peripheral mechanism of the ear; that the hair cells transform sound-vibrations into nerve-vibrations similar in frequency and amplitude to the sound-vibrations; that simple and complex vibrations of nerve-molecules arrive in the sensory cells of the brain, and there produce, not sound again of course, but the sensations of sound, the nature of which depends not upon the stimulation of different sensory cells, but on the frequency, amplitude, and form of the vibrations coming into the cells, probably through all the fibres of the auditory nerve. On such a theory the physical cause of harmony and discord is carried into the brain, and the mathematical principles of acoustics find an entrance into the obscure region of consciousness.

Somehow, Rutherford's *telephone theory* came to be called the *frequency theory* of hearing, which seems odd for a theory that contains no frequency analysis. The earliest instance that I find of this renaming is in a discussion of "tonal volume and pitch," about the perceptual dimension of "volume" as distinct from pitch and loudness (Dunlap, 1916). The term *frequency theory* has also been used the other way, as an alternative name for a resonance or place theory, contrasted with *periodicity theory* in describing pitch perception mediated by

time patterns (Rossing, 2007; O'Callaghan, 2007). Some authors treat both *periodicity theory* and *frequency theory* as synonymous names for Rutherford's telephone theory (Gelfand, 1990; Schiffman, 1990).

Part of the confusion is explained by Peter Cariani (1994):

> "Frequency" has two meanings, one associated with a rate of events, the other associated with a particular periodicity of events. Frequency Coding implies the former meaning.

More recently, theories related to Rutherford's telephone theory are sometimes called *temporal theories* (Moore, 2003; Gelfand, 2004), avoiding this confusion.

Another part of the confusion is that theories of hearing were really theories of pitch, or of the coding of perceived pitch frequency, and the dichotomy was often seen as between coding pitch by place, as Helmholtz theorized, versus coding pitch by frequency, or periodic time patterns, of nerve firings. Georg von Békésy (1956), who studied the sound-evoked vibration of the cochlea's *basilar membrane*, remarked on this situation:

> The words "theories of hearing" as commonly used are misleading. We know little about the functioning of the auditory nerve, and even less about the auditory cortex, and most of the theories of hearing do not make any statements about their functioning. Theories of hearing are usually concerned only with answering the question, how does the ear discriminate pitch? We must know how the vibrations produced by a sound are distributed along the length of the basilar membrane before we can understand how pitch is discriminated, and therefore theories of hearing are basically theories concerning the vibratory pattern of the basilar membrane and the sense organs attached to it.

This percept of *pitch* has a long and often confused or confusing history in auditory science. The basic problem is that the pitch of a sinusoid is equal to its frequency (by definition, essentially), but that the same pitch may be heard from sounds lacking that frequency in their Fourier decompositions. Treating pitch as a time-domain repetition provides an often better result than treating it via a Fourier decomposition, but a theory that gets pitch right needs to account for the frequency analysis in the cochlea, too, not just a periodicity analysis of the original sound waveform. None of these older "new" theories come close.

2.2 Newer Theories of Hearing

The observation by Békésy (1928, 1960) of traveling waves on the basilar membrane led to a big improvement in the understanding of mechanisms that partially separate sounds by frequency, but left the functional view of Helmholtz's resonance or place theory essentially unchanged. In his "Theorie des Hörens," the ear was still thought of as essentially a Fourier analyzer.

Harvey Fletcher (1930) referred to the frequency theories as "time pattern theories," which makes more sense, and saw the need to combine these with the Helmholtz-style resonance or "space pattern" theories, to explain more than just pitch:

> Two general types of hearing theories have been put forth from time to time to explain these effects. One might be called a space pattern theory and the other a time pattern theory. In the first theory, it is assumed that the time pattern of the wave motion in the air is transferred into a space pattern in the inner ear so that the nerve impulses reaching the brain give us information concerning the time pattern of the wave motion by means of the location of the nerves which are stimulated. In the second theory, it is assumed that the time sequences are transmitted directly to the brain. It is the opinion of the author that both of these effects are operating in aiding one

> to interpret the sounds which one hears. The term "A Space–Time Pattern Theory of Hearing" therefore best expresses this conception.

Except for his conclusion that "the term 'A Space–Time Pattern Theory of Hearing' best expresses this conception," modern scientists agree—sounds are coded on the auditory nerve by patterns of nerve firings with important spatial and temporal aspects. But Fletcher's terminology seems to have been lost in the ages. In the same year as he published it, Wever and Bray (1930b) referred to Fletcher's conception as a *resonance–volley theory*. Their *volley* concept of how nerves communicate time patterns has survived, and their theory is often called a *place–volley* theory (Freeman, 1948). There is still no widely shared conception of how the brain handles these patterns, or what to call them; Fletcher's "space–time pattern" is a term we can use, if we clarify that the space is the one-dimensional cochlear place. Patterns of time and two-dimensional space, introduced below as "auditory images," are also important to us in theories of what happens beyond the cochlea.

Time pattern theories need a description of how nerves can carry waveform information with bandwidth of over 1000 Hz, even though each neuron has a very limited firing rate—up to only a few hundred hertz. Auditory neurons do this by acting together in groups, extending the time pattern capability to a few kHz. Harvard psychologist Leonard Troland (1929, 1930) proposed such an idea:

> If it should turn out that single auditory nerve fibres are physiologically incapable of carrying the higher range of frequencies which yield pitch variation, we may still suppose that such frequencies can be conveyed by a *group of fibres* acting together.

It was Wever and Bray (1930b) at Princeton who invoked the terms *volley principle* and *volley theory* and were remembered for it: "it is possible for a high rate to be established by slowly acting fibers going off in volleys." This high rate of precisely timed firings on groups of neurons allowed the transmission—and reconstruction from electrically picked-up nerve signals—of sound information with frequencies at least up to 4500 Hz, according to their reported observations with cats:

> The transmission process is one of great fidelity. A tone sounded into the cat's ear is represented in the nerve response so that the effect as heard in the receiver is indistinguishable in pitch from the stimulus tone as heard directly. Complex sounds, including speech, are communicated readily.

Their concept of "great fidelity" is probably a great exaggeration, as it doesn't take much to reproduce pitch exactly, or speech intelligibly.

The volley idea has been further adapted to explain how the apparently random firings of neurons in quiet can have their timings modulated by signals that are too weak to noticeably increase the firing rates (Rose et al., 1967, 1971; Greenberg, 1980; Davis, 1983). In this way, the volleys of firings can represent waveforms even when the rate-versus-place representation shows nothing. Rose et al. (1967) reported synchrony in the range of 10–25 dB SPL in a squirrel monkey auditory nerve fiber with rate threshold of 25 dB SPL. Neural rate thresholds in cats are reported in the 6 to 24 dB SPL range (Greenberg et al., 1986), while behavioral threshold in the same species are in the −20 to −10 dB SPL range (Sokolovski, 1974). The observation that cats and monkeys (and we) can detect tones about 20 dB below the firing rate thresholds of auditory neurons can be understood as a volley or timing effect.

2.3 Active and Nonlinear Theories of Hearing

Thomas Gold (1948) proposed an "active" theory of hearing, in which the cochlea behaves like a regenerative radio receiver, using positive feedback to amplify weak signals. It took quite a few decades for this idea to

gain any acceptance, but after evoked oto-acoustic emissions ("Kemp echoes") and spontaneous oto-acoustic emissions were observed (Kemp, 1978; Zurek, 1981), the idea of an active cochlea finally caught on. It took a while longer to integrate it with traveling-wave models (Neely and Kim, 1983). But the concept of active mechanics didn't provide much more than an idea of how Helmholtz's place model might be realized; it was not a major rethinking of what the cochlea sends to the brain or how the brain interprets it.

Gold's theory, combined with various observations on strongly nonlinear amplitude response in cochlea mechanics, led to *active traveling-wave* theories and models of cochlea function (Johnstone et al., 1986), which continue to be evolved and improved today. But before this happened, there was another side trip to explore the *second-filter theories*. These were attempts to reconcile the relatively unsharp frequency tuning of passive traveling-wave models with the apparently much sharper frequency resolution seen through electrophysiology experiments on the cochlear nerve (Evans and Wilson, 1973). By the early 1980s, experiments had conclusively shown that the mechanical tuning in a healthy cochlea was just as sharp as the neural tuning, when viewed in a comparable way, so the second-filter work dropped by the way; as Cooper et al. (2008) summarizes:

> The original idea of a "second filter" in the auditory periphery turned out to be something of a red herring, but took over a decade to be replaced by our present concept of a "cochlear amplifier."

An important leap from the theories of hearing based on peripheral auditory function is represented in two theories that explicitly include central neural processing as well. The *place theory of sound localization* of Jeffress (1948) and the *duplex theory of pitch perception* of Licklider (1951) broke new ground, as early instances of what we now call auditory image theories. The duplex theory is discussed in Section 4.6, and more in Chapter 21, but before we focus on that, let's look at what else Licklider said.

2.4 Three Auditory Theories

In his chapter "Three Auditory Theories," Licklider (1959) made an attempt to describe some of the partial theories of hearing that had been formulated, since more complete theories "exist, if at all, only at a level below verbal formulation in a few brains." This chapter is well worth reading.

The three theories concerned signal detection, speech intelligibility, and pitch perception. He says:

> There is no systematic, over-all theory of hearing. No one since Helmholtz has tried to handle anything like all the known problems within a single framework. Each of the several theories of hearing that are extant deals with a restricted set of questions.

And it's not as if he meant that Helmholtz had got it right. The main parts of Helmholtz's theory were that the cochlea has an array of independent resonators, and that phases are ignored. Licklider goes on:

> Helmholtz's resonance–place theory of auditory frequency analysis and pitch perception was for years the main force in the field of hearing. The fact that both main parts of it were largely wrong did not lessen its influence. Békésy's direct observations of the inner ear in action altered the whole structure of the field.

Unfortunately, this altered structure was incomplete, and somewhat ephemeral, as too many scientists, and most nonspecialists, continued to accept Helmholtz's basic frequency–place idea as a model for what the cochlea sends to the brain, and to ignore that fact that the auditory nerve sends actual waveform detail to the brain in the form of the timing of nerve firings, known as *action potentials*. Everyone agrees that the

spatio-temporal pattern of these discrete action events is used to represent and compute signals in the brain, but too often the pattern is conceptually trivialized as just a local average rate of action potentials on each nerve, ignoring the information that can be carried by fine temporal patterns within and between the nerve fibers. This deficit is like encoding Fourier component magnitudes and ignoring phase relationships.

The pitch perception part of the three theories was Licklider's own duplex theory, now elaborated into a "triplex" theory that includes some binaural effects. His attitudes about pitch perception, and its place in the nervous system, led him to formulate this reaction to Fourier analysis:

> In the theory of pitch perception, frequency analysis is fundamental. Probably the most important conceptual operations—analysis of waves into elementary sinusoidal components, which we have already encountered, and synthesis of waves from these components—are derived from the physicist–mathematician Fourier. Fourier's ideas got into the field of hearing in time to influence Helmholtz. They have been, and are, basic and essential for handling the mechanical part of the auditory process. But I think they have been applied beyond their realm of applicability. It seems to me that the power of the Fourier transformations, and the tractability of the assumption of linearity (not applicable to the later stages of the process) trapped auditory research into a long and unfortunate preoccupation with pure tones as auditory stimuli.

Békésy (1974) said essentially the same thing, in much stronger terms, when he ranked Fourier analysis right up with "dehydrated cats" as among the main impediments to progress in hearing research.

Are we in a position yet to combine something like Licklider's three theories, and others, into a single framework? It seems that it ought to be possible. Signal detection and speech intelligibility and pitch perception should all be expressed in terms of the same signals from the periphery via the auditory nerve, and might as well be expressed in terms of common representations at the next few levels, too. This duplex theory is a good place to start; we call it the auditory image.

2.5 The Auditory Image Theory of Hearing

Rather than propose yet another new theory of hearing, we propose a framework, and a name—the *auditory image theory*—within which modern approaches can be unified and conceptualized.

A modern theory must go way beyond trying to explain pitch. The idea of this approach is to incorporate theories, knowledge, and experimental data up through processing in auditory cortex—that two-dimensional sheet of gray matter that seems more well matched for processing images. We don't necessarily stop at primary auditory cortex, but leverage the analogy of auditory cortex to visual cortex, including secondary and subsequent areas, motivating the idea that representations that project to cortex are "images," or "maps," including "sketches." Both visual and auditory senses, along with touch and possibly others, map sensory dimensions into two-dimensional sheets of cortex, with a temporal resolution too slow to follow the time patterns of even very low pitches. The job of the lower parts of the auditory brain is to "demodulate" the fine time structure that comes in on the cochlear nerve, to lay it out spatially for projection to cortex.

This approach does not constrain what theories we might rely on at lower levels, such as theories of what the cochlea sends to the lower brain stages via the auditory nerve. Whether we conceptualize the cochlea via one of the older simpler theories, or represent its function via the wealth of detailed modern knowledge and models, we can build on that level, using models of the intermediate brain stages. Such models produce one or more image-like representations, of the sort that might be projected to cortex. From there, we generate further derived representations, to put hearing to use.

The earliest theories that leverage the second place dimension afforded by the sheet-like structure of cortex are probably the Jeffress and Licklider theories mentioned above, illustrated in Figure 2.1, which make activity maps that capture much more than just pitch and direction of sounds.

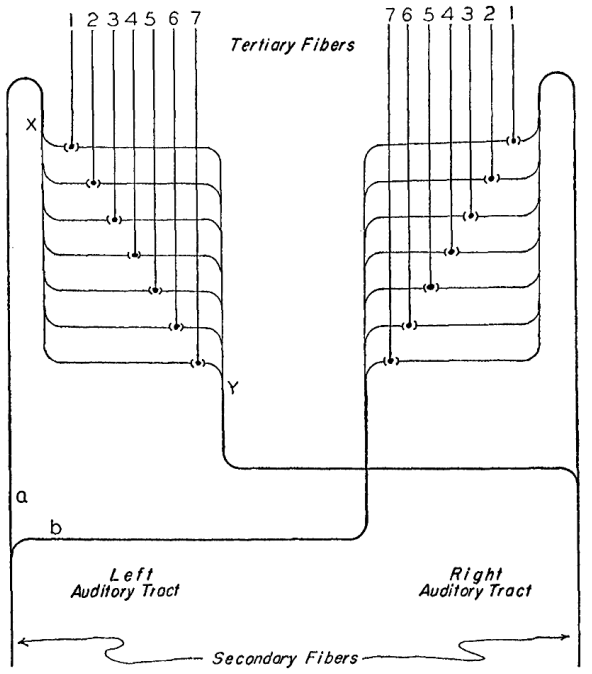


FIGURE 1. HYPOTHETICAL MID-BRAIN MECHANISM FOR THE LOCALIZATION OF LOW FREQUENCY TONES

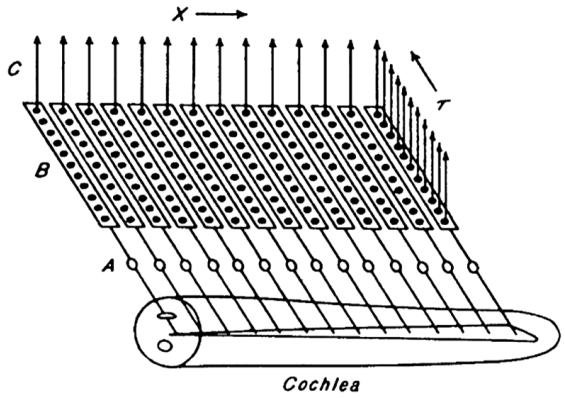


Fig. 2. – *Schematic diagram of overall analyzer.* At the bottom is the uncoiled cochlea. Its lengthwise dimension and the corresponding dimension in the neural tissue above it is designated the x-dimension. The cochlea performs a crude frequency analysis of the stimulus time function, distributing different frequency bands to different x-positions. In the process of exciting the neurons of the auditory nerve, the outputs of the cochlear filters are rectified and smoothed. The resulting signals are carried by the groups of neurons A to the autocorrelators B, whose delay- or τ-dimension is orthogonal to x. The outputs of the autocorrelators are fed to higher centers over the matrix of channels C, a cross-section through which is called the (x, τ)-plane. (Output arrows arise from all the dots; some are omitted in the diagram to avoid confusion.) The time-varying distribution of activity in the (x, τ)-plane provides a progressive analysis of the acoustic stimulus, first in frequency and then in periodicity.

Figure 2.1: Jeffress's (left) and Licklider's (right) drawings of their binaural and pitch models of the neural formation of auditory images (Jeffress, 1948; Licklider, 1951). Coincidence detection between differently delayed neural events, or in Licklider's between delayed and nondelayed events, generates the time-difference dimension of a map. Jeffress does not show a tonotopic axis, but his scheme has generally been interpreted as one frequency slice of a two-dimensional structure like Licklider's (Lyon, 1983; Shackleton et al., 1992; Hartung and Trahiotis, 2001). Jeffress guessed that such a structure might be found in the superior olivary complex—where a mapping of interaural delay was actually found years later. [Reproduced (Jeffress, 1948) with permission of the American Psychological Association; (Licklider, 1951) with permission of Springer.]

As Licklider (1959) explains, his duplex theory essentially merges competing views, in the spirit of Fletcher's space–time pattern theory:

> This duplex theory reconciles place and frequency theories in the sense that both appear as partly correct. It makes clear the futility of trying to disprove one by proving the other.

Similar ideas were developed into electrical waveform analyzers—early machine hearing systems—in the 1960s. These systems by John L. Stewart and his colleagues illustrate an evolution of thinking from a cochlear place model (Caldwell, Glaesser, and Stewart, 1962) to a two-dimensional "auditory image" model resembling Licklider's model, with a "neural analyzer" on each place channel (see Figure 2.2), adding another dimension to make an image-like output (Stewart, 1966).

More recently, many auditory neurophysiologists have been trying to find out what signal attributes may be mapped in the second dimension, in brain structures where one dimension is usually tonotopic, that is, in monotonic correspondence with frequency or with the place dimension of the cochlea (Schreiner, 1991; Cariani, 1994; Langner et al., 1997). By 1980, at least six different such two-dimensional maps in auditory cortex had been identified in rhesus monkeys (Merzenich and Kaas, 1980; Cook, 1986); the codes of these auditory images are still not well understood, and are still being actively investigated (Schulze et al., 2002; Langner, 2005).

The Jeffress and Licklider models represent auditory images that are calculated well below the level of cortex. The Jeffress interaural time difference (ITD) map is calculated in the medial superior olive (MSO) of the brainstem (Joris et al., 1998), and the Licklider periodicity map, or something roughly equivalent, is likely calculated in the central nucleus of inferior colliculus (ICC) of the midbrain (Ehret and Merzenich, 1985; Langner et al., 2002). In echolocating bats, maps of echo delay, formed via correlators as in Licklider's model, are prominent in cortex (Knudsen et al., 1987). Much is still unknown about what images are computed by what brain structure, and what transformations they undergo on the way to auditory cortex and within different cortical areas. The auditory image framework is intended to embrace all of these levels, giving us a way to conceptualize and visualize rich representations of important sound properties, as computed by the nervous system from the space–time patterns on the two auditory nerves.

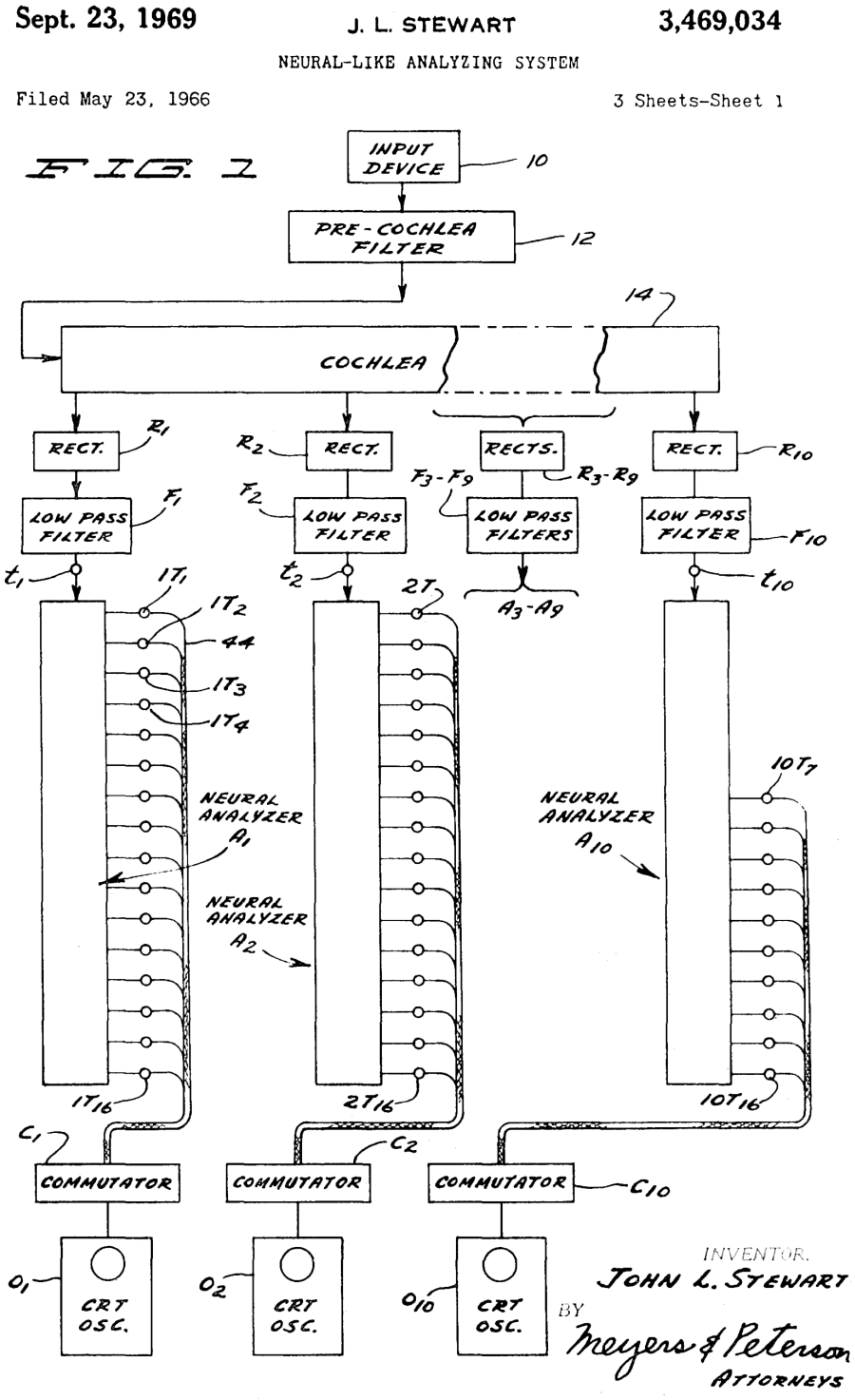


Figure 2.2: A patent drawing from the John L. Stewart (1966) "neural-like analyzer." The "neural analyzer" stages at each rectified output of the "cochlea" generate a second dimension, mapping the cochlea's space–time pattern to a space–space pattern, an image-like map, much as in Jeffress's and Licklider's theories.

Chapter 3

On Logarithmic and Power-Law Hearing

> The task of clearing the scientific bench top of the century-long preoccupation with the *jnd* [just-noticeable difference], and the consequent belief in logarithmic functions, demands the cleansing power of a superior replacement. My optimism on this score has been recorded in other places, but I would like here to suggest that, if I seem to feel a measure of enthusiasm for the power law relating sensation magnitude to stimulus intensity, it is only because that law seems to me to exhibit some highly desirable features.
>
> — Stevens (1961), "To honor Fechner and repeal his law: a power function, not a log function, describes the operating characteristic of a sensory system"

Logarithms, exponentials, and power laws appear frequently in signal analysis, and especially in hearing-motivated techniques. It is important to understand the reasons for their use, and to be able to recognize when they are inappropriate, and how to modify such mappings to make them more practical and robust.

3.1 Logarithms and Power Laws

Engineers like to describe signals and their spectra—and systems that process them—in logarithmic units. Our hearing is sometimes described as logarithmic, along both the loudness dimension and the pitch (or frequency) dimension. So we need to understand what this means, what's powerful and useful about logarithms, and what their limitations are as a conceptual model for perception of loudness and pitch in hearing.

As the Britannica (1797) cryptically explains, logarithms are "the indices of the ratios of numbers to one another ; being a series of numbers in arithmetical progression, corresponding to others in geometrical progression ; by means of which, arithmetical calculations can be made with much more ease and expedition than otherwise." That is, logarithms were an invented way to make multiplication not much harder than addition, long before the logarithm was understood as a mathematical function. The logarithm function is also of great importance as the inverse of the exponential function, as we discuss in a later section.

A power law, on the other hand, is a remapping through a power, or exponentiation, such as a square, or a square root. Power laws also come in function/inverse pairs: the square and square root, or cube and cube root, or Nth power and Nth root ($1/N$ power) in general, are such pairs. Such relationships are at least as useful in describing sensory systems as exponential or logarithmic relationships are. The exponential and logarithm functions can be considered to represent the limiting cases of power laws for N very far from 1, as shown in Figure 3.1.

The Mathematics of Logarithms and Power Laws

The algebraic definition of logarithm leads to several useful relationships. Given a value x, and a base b, the base-b logarithm of the value x is the number y that satisfies the equation:

$$x = b^y$$

The logarithm is essentially a functional inverse of this exponentiation operation. In terms of the logarithm function, we write the "solution" of the above equation as:

$$y = log_b(x)$$

That is, exponentiation maps y to x, and the logarithm function maps x to y, as long as both of them use the same base b. Any positive number other than 1 will work for b, but special numbers like 2 for *binary* logarithmss, e for *natural* logarithms, and 10 for so-called *common* logarithms are most often encountered as bases. The value e is the unique number (about 2.71828) such that the exponential curve e^x has unit slope at $x = 0$ (more generally, $\frac{d}{dx}e^x = e^x$, for this and no other value of e).

Properties of logarithms and different bases are easy to derive from the properties of exponents.

A power law looks similar, but the variables are not in the exponents. Here we base the formulas on an exponent parameter α, usually between 0 and 1, instead of the integer power N and its reciprocal $1/N$ mentioned earlier:

$$y = x^\alpha$$

$$x = y^{1/\alpha}$$

As α approaches 1, we approach the identity relationship between x and y. The other extreme, as α approaches 0, is more interesting, but we'll need to rewrite the relations in a way that makes the power law functions converge to a consistent mapping in that limit. Let's scale and offset x and y to pick the case of converging on the identity function near the point (1, 1)—that is, such that all functions pass through the point (1, 1) with unit slope—while keeping the point of infinite slope at $x = 0$:

$$y = \left(x^\alpha - 1\right)/\alpha + 1$$

$$x = (\alpha y - \alpha + 1)^{1/\alpha}$$

In the limit of small α, these modified power-law functions approach exponential/logarithm relationships that have been similarly shifted to be tangent to the identity function at (1, 1), as illustrated in Figure 3.1:

$$y = \log_e(x) + 1$$

$$x = \exp(y - 1)$$

In this sense, the power-law functions are good intermediate mappings for many purposes—not linear, but not as extreme as logarithms and exponentials.

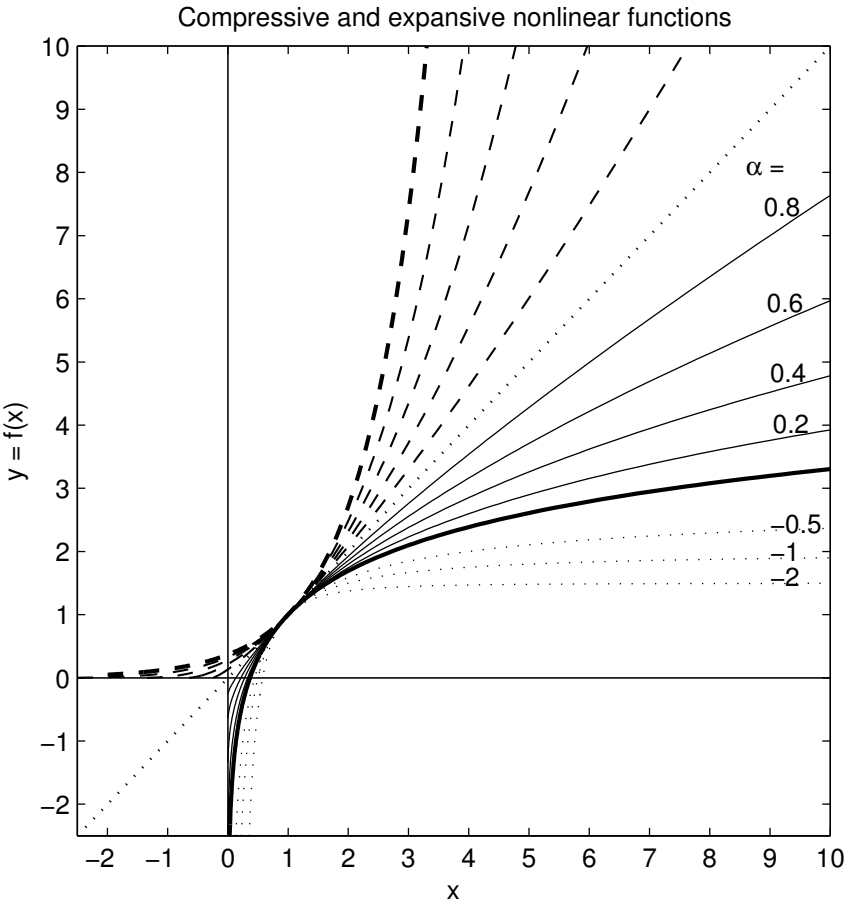


Figure 3.1: Some compressive nonlinearities (solid curves) and the expansive nonlinearities (dashed curves) that are their inverses (compressive means with "diminishing return," that is, slope decreasing as input increases, while expansive means the opposite). The heavy solid curve is a logarithmic compression, and the heavy dashed curve is an exponential expansion; the lighter curves are based on power-law relationships, with exponents $0 < \alpha < 1$ as annotated. As explained in the text, the functions are all adjusted to be tangent to the identity function (dotted) at the point (1, 1), such that the α exponent interpolates the compressive functions between log and linear, for $x > 0$. Also shown (dotted curves) are the even more compressive functions that result from negative exponents—linear tranformations of the reciprocal square root, reciprocal, and reciprocal square. Tukey (1957) discussed the logarithm as the natural limit between these curves with positive and negative exponents.

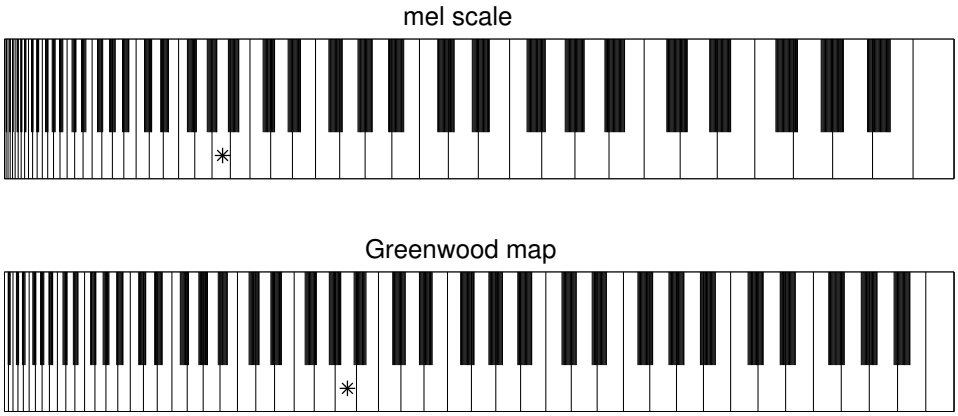


Figure 3.2: A normal piano keyboard has a logarithmic mapping of frequency to place, but these distorted ones have auditory mappings. The top keyboard is based on the mel scale (see Section 5.6), which severely squashes the low frequencies. The bottom one is based on the Greenwood map, a more accurate reflection of auditory physiology and psychophysics. In both, the x coordinates of the keys are *stabilized logarithms* of the corresponding note frequencies, with different stabilizing offsets. That is, positions on the distorted keyboard are linear functions of $\log(f + f_{\text{break}})$ for a stabilizing offset f_{break} that we refer to as the break frequency—the approximate breakpoint between a low-frequency linear limit and a high-frequency logarithmic limit. The keys marked "*" are A440, a 440 Hz pitch that is on the linear side of the mel scale's 700 Hz break frequency, but on the logarithmic side of the Greenwood map's 165 Hz break frequency.

3.2 Log Frequency

The 88 keys of a piano are pretty nearly equally spaced across the keyboard, and the pitches of the musical notes that they produce are in approximately equal ratios from one to the next. The ratio between adjacent notes is called one *semitone*, which is a constant ratio in some tuning systems, but can vary a bit from note to note in other systems. Assuming the semitone is a constant ratio, we can say that the key number is the logarithm of the pitch, because the key numbers (1, 2, 3, … 87, 88) are in arithmetical progression, that is, with a constant difference between successive numbers, and these key numbers correspond to pitches (27.5, 29.1, 30.9, … 3951, 4186 Hz) in geometric progression, that is, with a constant ratio between successive pitch values.

The octave number is also a logarithm. Keys that are an *octave* apart on the keyboard produce notes with a pitch ratio of 1:2. The note names A0, A1, …, A7 correspond to the successively doubling frequencies 27.5, 55, 110, 220, 440, 880, 1760, 3520 Hz. The formula $f/27.5 = 2^m$ gives the ratio of the note pitch to the starting pitch, from the octave number m, for the notes Am. Here we say that the *base* of the logarithm is 2, since 2 is the number being raised to a power as specified by the logarithm. That is, the logarithm tells us what power of the base (2) is needed to give the pitch in question—or rather, its ratio to a specified starting pitch, 27.5 Hz in this example.

A semitone corresponds to 1/12 of an octave, or a frequency ratio of $2^{1/12} = 1.059$, the twelfth root of 2, so 1.059 would be the base implied in calling key number a logarithm of pitch.

Musicians know that certain pitch ratios have certain characteristic sounds. A ratio 3:2 is a perfect fifth, and 4:3 a perfect fourth, no matter what pitch range they are in. These musical intervals correspond to moving 7 or 5 keys or semitones to the right, respectively. These differences of logarithms, 7 and 5, represent approximately the ratios 3:2 and 4:3, and seem to have some important relationship to how we hear pitches. So it is often said that humans perceive pitch on a logarithmic scale. It is more true that musical instruments

produce notes on a logarithmic scale.

For the piano pitch examples, the x in $x = b^y$ would be the ratio of pitch to the starting pitch, the pitch that corresponds to a logarithm of 0: x = PitchRatio = $f/27.5$ for the octave example, relative to the pitch of the lowest piano note, A0.

$$\text{OctaveNumber} = \log_2(\text{PitchRatio})$$

In spite of these pure logarithmic relationships in music, a human's perceptual scaling of frequency is not quite what we might think from the fact that musical intervals depend only on the frequency ratio. Below a few hundred hertz, equal perceptual pitch intervals for sine waves approach an equal number of hertz, instead of a constant percentage, as we illustrate by warping the piano keyboards in Figure 3.2. As Pierce and David (1958) explained in *Man's World of Sound*, contrasting the perceptual pitch scale, known as *mel scale* for *melody*, to a logarithmic musical scale:

> We can only conclude that for sine waves, at least, "equal" musical intervals do not represent equal intervals of subjective pitch. ... This baffled me to the extent that I nearly left the mel scale out of this book. However, it represents real psychoacoustic data. Moreover, it is related to other important psychoacoustic data. The limen or just noticeable difference in pitch is a nearly equal number of mels. ... I am now inclined to believe that the mel scale reflects a "place" mechanism in the ear ..., while the scale of musical pitch is associated with another, a time-comparison phenomenon ...

This dichotomy between different aspects of musical pitch often shows up as a complicating factor in machine hearing and music analysis, as it does in psychophysics. A relationship based on a just-noticeable difference (jnd, also known as a difference limen) is not predictive of the relationships between more widely separated pitches.

3.3 Log Power

It is also often said that we perceive loudness on a logarithmic scale. Doubling the power (or *intensity*) of a sound wave corresponds to what engineers call a 3 decibel (dB) increase. Equally spaced dB values such as 0, 3, 6, 9, ... correspond to a starting power and successive doublings, such as 1, 2, 4, 8, ..., so the dB values are logarithms of the sound power. Each doubling of power sounds to us more like a certain *increment* in loudness than a certain *factor* in loudness—or so some would claim, even though it's far from a good fit to reality.

Formal definitions of the decibel as a unit seem to confuse more than help, so we revert to the engineer's informal definition. The decibel scale is defined such that an increment of 10 dB corresponds to a power ratio of 10, or an amplitude (e.g. peak voltage or pressure) ratio of $\sqrt{10}$. An increment of 1 dB represents the tenth root of these ratios; therefore, a number in dB is a base 1.259 logarithm of a power ratio or a base 1.122 logarithm of an amplitude ratio. But we never think of dB quite that way, as those would be ridiculous numbers to have to remember and work with. Instead we say that a relationship expressed in dB is 10 times the common logarithm of a power ratio, or 20 times the common logarithm of an amplitude ratio, where *common* means base 10:

$$A_{\text{dB}} = 10\log_{10}(\text{PowerRatio}) = \log_{1.259}(\text{PowerRatio})$$

$$A_{\text{dB}} = 20\log_{10}(\text{AmplitudeRatio}) = \log_{1.122}(\text{AmplitudeRatio})$$

This dual definition reflects the fact that power is proportional to the square of amplitude (in the context of linear systems). Doubling the factor in front has the same effect as squaring the ratio, so these definitions

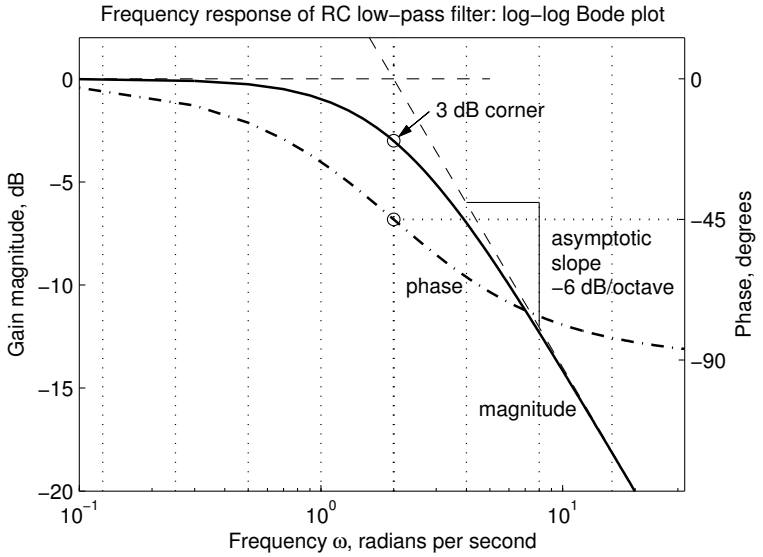


Figure 3.3: Bode plot, or log–log frequency response, of a simple lowpass filter. This is a typical way in which engineers combine logarithmic frequency and amplitude scales to characterize a filter, usually resulting in straight-line asymptotes, such as shown (dashed). The phase response of the filter is also shown (dash-dot). Phase can also be thought of as a logarithmic scale, the imaginary part of the complex logarithm of the filter's transfer function.

are equivalent, at least in linear systems.

An amplitude ratio of $\sqrt{2}$, or a power ratio of 2, corresponds to $10\log_{10}(2) = 3.0103$ dB, which is conventionally just called 3 dB, since that's more accuracy than engineers usually need. The actual power ratio represented by 3 dB is thus $10^{3/10} = 1.995$, a doubling, to within a fraction of a percent.

3.4 Bode Plots

Engineers like to combine log-frequency with log-power to make linear-system descriptions known as *Bode plots* or *Bode diagrams* (Siebert, 1986), named for their originator, the American engineer Hendrik Wade Bode (1945). These are plots of response magnitude (output-to-input ratio) in dB (and sometimes also phase shift) as the ordinate (vertical coordinate), versus frequency transformed to a logarithmic abscissa (horizontal coordinate).

There are several good reasons for the mapping of both magnitude (power or amplitude) and frequency to their logarithms. In terms of engineering, there are two important considerations: first, the wide dynamic range of intensity and frequency are easier to describe after logarithmic compression, so that details in the low part of the range are not lost; second, many functional relationships that are encountered in system analysis yield straight lines, or straight-line asymptotes, in the log–log domain, making plots of system characteristics easier to use and describe—which is why Bode plots, such as the one shown in Figure 3.3, are sometimes called *asymptotic phase and magnitude plots*.

Complex Numbers and Euler's Formula

The reader will need basic familiarity with complex numbers (numbers such as $3 + 2i$ that have an imaginary part using the pure imaginary number i, defined by $i^2 = -1$) and with $e = 2.71828\ldots$, the base of the natural logarithms, for any kind of work with sound, waves, hearing, or linear systems. It will also be important to understand some basic properties of the exponential function, $\exp(x) = e^x$, with complex argument. In particular, one needs to be familiar with *Euler's formula* (or Euler's identity, not to be confused with Euler's polyhedron formula or the numerous other mathematical concepts named for Leonard Euler):

$$\exp(i\theta) = \cos\theta + i\sin\theta$$

The formula says that the exponential of a pure-imaginary number, $\exp(i\theta)$, is a point on the unit circle in the complex plane (that is, having an absolute value of 1, or at Euclidean distance 1 from the origin), at an angle from the real axis equal to the value θ in the exponent. Here, angles are measured in natural units, known as radians, equivalent to arc length on the unit circle, counterclockwise from the real axis in the real–imaginary plane; see Figure 3.4. Since the arc length around a complete cycle of the unit circle is 2π, that factor shows up frequently, for example in converting between cycles and radians.

Euler's formula is key to how we express oscillations, or tones. Its oft-quoted amusing special case $e^{i\pi} = -1$ obscures its importance.

The angle θ often represents a phase increasing with time, in which case the derivative $d\theta/dt$ is the frequency, in radians per second, and the exponential represents a rotating complex function of time—an important complex generalization of the *sine wave*.

Notice the notation for the exponential function: we prefer to treat e^x as a function, at least as important conceptually as things like $\cos(x)$ and $\log(x)$, rather than emphasize the number e.

It would work almost as well to use other numbers, such as 10, instead of e—with $\exp_{10}$ as the inverse of $\log_{10}$—representing oscillating signals as $10^{i\theta}$. But that would introduce unnatural conversion factors: rather than the 2π units of phase per cycle that we get with $e^{i\theta}$, we would have about 2.729 units of phase per cycle, and that number, $2\pi/\log_e(10)$, couldn't be expressed exactly without invoking e, or natural logarithms. So *naturally,* we use e, and the natural phase unit, radians, and the corresponding natural definitions of the sine and cosine functions, and frequency in the natural units of radians per second.

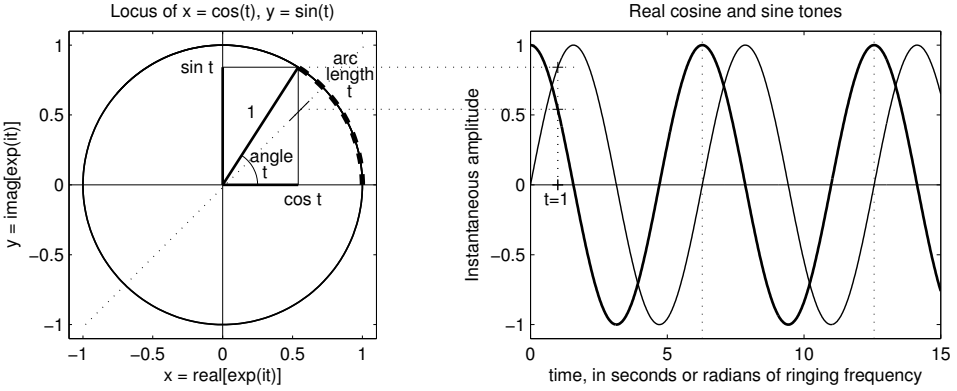


Figure 3.4: Applying Euler's formula to tones. On the left, the points on the unit circle represent pairs $(\cos t, \sin t)$ with parameter t; by Euler's formula, each such point can be interpreted as a complex-plane plot of the value $\exp(it)$. The marked example point and coordinates represent $t = 1$, corresponding to an arc length of 1 and an angle of 1 radian, as measured from the $\exp(0) = 1$ point on the positive real axis. On the right, the cosine function $\cos(t)$ (dark curve) and sine function $\sin(t)$ (light curve) are the x and y coordinates of points on that unit circle, plotted as functions of the parameter t (the dotted diagonal on the left reflects the x coordinate to make the $\cos(t)$ ordinate on the right).

Complex Logarithm History

Before Euler articulated the formula that bears his name, circa 1740, Roger Cotes had already observed the logarithmic form of it in 1714 (Stillwell, 2010):

$$\log(\cos\theta + i\sin\theta) = i\theta$$

where by log he meant the natural logarithm, base e. In so doing, he invented the value e, and was among the inventors of the natural logarithm.

But Cotes's version has a problem that was not appreciated at the time of this first attempt to extend logarithms to the domain of complex numbers: in order to treat the complex logarithm as a function, we need to define a principal value, by picking one *branch*. There are multiple distinct values of θ that give the same results for $\cos\theta$ and $\sin\theta$, and hence for the log, so the equation can't be true for more than one of them. For the others, $i\theta$ is not within the range of the chosen branch. We can extend Cotes's formula to say

$$\log(\cos\theta + i\sin\theta) = i(\theta + n2\pi)$$

for some n that depends on θ, chosen to bring the result into the range of principal values (typically defined as the range where the imaginary part, the angle in radians of the argument of the log function in the complex plane, is greater than $-\pi$ and less than or equal to π).

Euler used the exponential function, the inverse of the logarithm, to sidestep this complication and give a simple and always correct equation. The interpretation involving points on a circle in the Cartesian complex plane was not known until it was discovered and published by Caspar Wessel in 1797 and by Jean-Robert Argand in 1806 (Wessel's article wasn't translated from Danish into French until 1899, so it didn't have much impact) (Fine, 1903). The Cartesian complex plane as a mechanism for visualizing complex numbers geometrically is sometimes called the *Argand plane*.

Electrical engineers know that the breakthrough in analysis of AC circuits, which led to widespread electrical power generation and distribution, was Charles Steinmetz's application of complex numbers to circuit analysis (Steinmetz, 1893); complex logarithms are key to analyzing propagation of telephone and telegraph signals over long wires. This application of Euler's formula also contributed to the techniques of filter design and analysis that led to a revolution in multiplexed wired and wireless telephone and telegraph communications, across continents, and across oceans. These steps have been part of what has so accelerated progress in the technical arts in the last century.

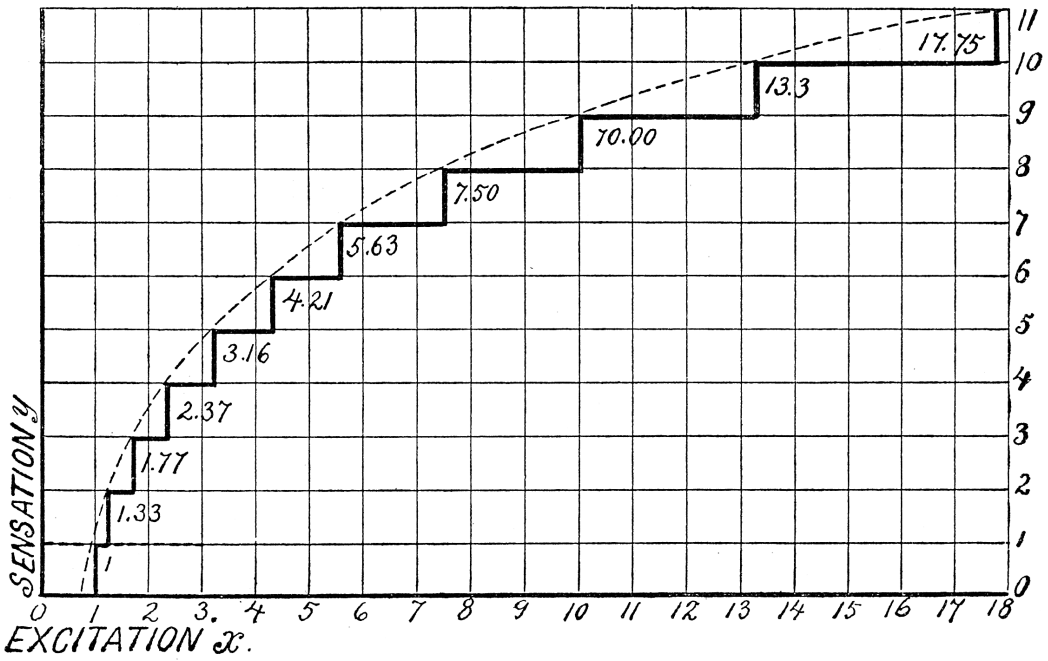


Figure 3.5: The Weber–Fechner psychophysical law: intensities in geometric progression evoke sensations in arithmetic progression; plot from Howell (1915). It is apparent that an intensity (excitation) approaching zero cannot be accommodated in this scheme, and that the zero point of sensation is arbitrary. Logarithmic scales always have such problems.

3.5 Perceptual Mappings

In scaling of perceptual dimensions, there is a good reason that logs are used: the perceptual distances between stimuli tend to agree better with ratios or logarithms than with linear differences.

The famous Weber–Fechner psychophysical law (Howell, 1915; Hecht, 1924) says that there is a logarithmic relationship between the intensity of a stimulus and its perceived strength—or equivalently that a given perceived difference corresponds to the same intensity ratio over a wide range of intensities. And our experience with music supports a strong connection of a logarithmic frequency scale to musical interval perception. But the logarithm has a nasty singularity (that is, it "blows up") at zero, which would suggest that there's no lower limit to the intensity range over which we can perceive a given ratio, and no lower limit to the frequency range over which we can distinguish musical intervals. So the logarithmic mappings, though convenient for some engineering purposes, should be regarded as only a first approximation for perceptual attributes, not valid at the low end of the range.

Weber and others before him had observed that for an increment in intensity to be detectable, a certain ratio of intensities was required; Bouguer (1760) had found that a light intensity ratio of 1 part in 64 was just noticeable, but did not explore the range of intensities over which that observation held. Fechner actually paid more attention to the failure, at low intensities, of Weber's proposed constancy of the ratio $\Delta I/I$, and proposed an alternative with a low-end stabilizing offset (Hecht, 1924) representing essentially the noise floor or absolute threshold in the expression for a constant increment of sensation ΔS:

$$\Delta S = \frac{\Delta I}{I_0 + I}$$

Unfortunately, this good idea got lost, and the idea that an increment in sensation corresponds to constant ratio in intensity came to be cast as a "law," confusing generations to come. Though Fechner's law came in for a lot of early criticism for its bad behavior at the low end (Trotter, 1878), it lived on for its simplicity. Others repeated Fechner's observation that the Weber fraction $\Delta I/I$ would have to be bigger at the low end; for example, Guernsey (1922) wrote:

> Weber's law as applied to audition apparently holds true with a fraction of about one-third throughout the middle range of intensities. The fraction is larger for low tones and for very near the limen, decreasing universally in the third, fourth, or fifth step. Whether this result is consistent through the upper range of intensities it has not been possible to determine until our apparatus is modified to produce greater intensities.

where the "fraction of about one-third" for $\Delta I/I$, meaning an intensity ratio of 4/3, corresponds to about a 1.2 dB difference limen, or jnd. Careful studies with other methods, about the same time, found a Weber fraction near 1/10, or difference limen near 0.4 dB, and also mapped out the dependence on frequency and level, up to very high levels (Knudsen, 1923). Modern values can be higher or lower, depending on the exact task, and tend to be less level-dependent for broadband signals than for tones (Houtsma et al., 1980).

The absurdity of Fechner's law near zero stimulus level, as shown in Figure 3.5, was apparent to everyone, so he had plenty of detractors from his attempt to make psychology mathematical. James (1890) commented on reactions to Fechner:

> Those who desire this dreadful literature can find it; it has a "disciplinary value;" but I will not even enumerate it in a footnote. The only amusing part of it is that Fechner's critics should always feel bound, after smiting his theories hip and thigh and leaving not a stick of them standing, to wind up by saying that nevertheless to him belongs the *imperishable glory*, of first formulating them and thereby turning psychology into an *exact science*.

Following Fechner's good idea of modifying the difference-limen ratio, the log function can be stabilized at the low end by a small offset, pushing its singularity outside the domain of nonnegative intensity or frequency:

$$f(x) = \log(x + \epsilon)$$

where ϵ is near the low end of the range in which constant ratios are perceived as constant increments.

Alternatively, the log function can be replaced by a power law, as explained by S. S. Stevens (1961) in his paper "To honor Fechner and repeal his law":

$$f(x) = x^{\alpha}$$

for $0 < \alpha < 1$. Since the exponent used to describe sensation is typically less than 1, some might be more comfortable thinking of this as being more like a root than a power, that is, a square-root relationship for $\alpha = 1/2$ or a cube-root relationship for $\alpha = 1/3$. For sound intensity, an exponent of about $\alpha = 0.3$ works well to represent the mapping from intensity (the physical power of the sound wave or of the electrical signal driving a speaker or headphone) to loudness (the perceptual correlate of intensity). This exponent was chosen by Stevens (1936) as the basis of a better perceptual unit of loudness, the *sone*: a doubling of perceived loudness, as quantified by a doubling of the sones, corresponds to a factor of 10 increase in power, by his definition.

The *Stevens power law* mappings with exponents between 0 and 1 are more moderate and well behaved than a logarithmic mapping, and power laws with exponents greater than 1 are more moderate than exponential, or geometric, relationships. Such mappings are widely applicable, not just in perceptual functions. Unlike the logarithm, the power-law function doesn't have a singularity at zero—but its derivative does. Sometimes

the power law is also stabilized by an offset, to push the slope singularity out of the domain of intensities. Tukey (1957) used explicit offsets in his family of log and power-law transformations, and considered separate cases depending on whether the value after offset could be zero or not.

The actual nonlinear mappings encountered in sensory and perception systems are a bit more complicated than these simple functions, and arise from the mechanisms that such systems use to adapt to a wide dynamic range of stimulus intensities. While treating the result as either power-law or logarithmic is often useful, it is also often misleading if taken too seriously and applied outside the range where the approximation fits. Billock and Tsou (2011), attempting "to honor Fechner and obey Stevens," explain the psychophysical functions as emergent from physiological mechanisms, much as they are in the models that we develop in this book:

> There is a class of models in which two nonlinear neural mechanisms (e.g., a sensory channel and the cortical numerosity mechanism tapped by magnitude estimation) are coupled through feedback, yielding power law behavior as an emergent property of the system, with an exponent that is a ratio of neural coupling strengths. Rather than a discrepancy between psychophysics and physiology, these models suggest complementarity between inner and outer psychophysics, because the Weber constants required for outer psychophysics modeling can be derived from the sigmoid nonlinearities of inner psychophysics.

3.6 Constant-Q Analysis

We often analyze signals through *filterbanks*, banks (ordered collections) of *bandpass filters*. Each *channel* of a filterbank (that is, each bandpass filter in the collection) has a *center frequency* and a *bandwidth*. The bandwidth might be measured as the width of the band of frequencies that the filter channel passes with less than (typically) 3 dB of loss relative to its center frequency. The ratio of center frequency to bandwidth is called the Q of the filter. For example, a filter centered at 1 kHz that passes frequencies from 900 to 1100 Hz with less than 3 dB of loss has a bandwidth of 200 Hz and a Q of 5.

If all the channels have the same shape and the same Q, and they are spaced so that each one's lower band edge matches the previous one's upper band edge, then the center frequencies will form a geometric progression. The center frequencies will be an exponential function of channel number, and the channel number will be a logarithmic function of center frequency. Therefore, the idea of constant-Q filterbank analysis, or a constant-Q transform as it is sometimes called, is essentially equivalent to working on a log-frequency scale.

Various constant-Q techniques for sound analysis were developed in the 1970s and 1980s, motivated by simple models of hearing (Kates, 1983; Petersen and Boll, 1983; Schwede, 1983; Roads, 1996). Even earlier, banks of one-third-octave filters (Q of about 4.3) were a common signal analysis tool; they were so commonly used that there were even efficient digital implementations of them in the 1960s (Otnes and Enochson, 1968). Due to the nature of the log, or the nature of geometric progressions, an arbitrary lower limit on the sequence of center frequencies is needed. The result is that the lower octaves of sound either don't get included, or they get included with more channels and more resolution than they deserve, relative to human hearing.

As we discuss in Section 5.6, the constant-Q idea has been mostly superseded by the idea of an auditory frequency scale, such as the *mel scale*. The nonlinearity used to map channel number to frequency is typically an offset exponential; the result is that the filter Q values are nearly constant in the higher-frequency channels, but decrease to near 1 at the lowest frequencies, such that zero frequency can be approached without too many channels.

3.7 Use Logarithms with Caution

In this chapter, we have reviewed the mathematical properties of logarithms, and explained why they are so popular and useful in various engineering and mathematical contexts. We have also looked at how they are sometimes applied to hearing and other perceptual systems, where their mathematical properties are more problematic. Stabilized logarithms and power laws can be useful alternatives, or at least workarounds to avoid the worst problems of logarithms, in representing either physical attributes such as frequency and intensity or perceptual attributes such as pitch and loudness.

In Part II we will use logarithms, especially complex logarithms, extensively in describing the behavior of linear systems, and especially distributed systems such as wave propagation in the cochlea.

Chapter 4

Human Hearing Overview

> On the zigzagging road towards wisdom about the human auditory system we collect knowledge from two entirely different sources of experimental information. First, from anatomy and physiology ... Secondly, from perception and psychoacoustics ... Our ever-wondering mind tries to combine and to explain these findings in terms of some model, law, hypothesis or theory.
>
> — "The residue revisited," J. F. Schouten (1970)

4.1 Human versus Machine

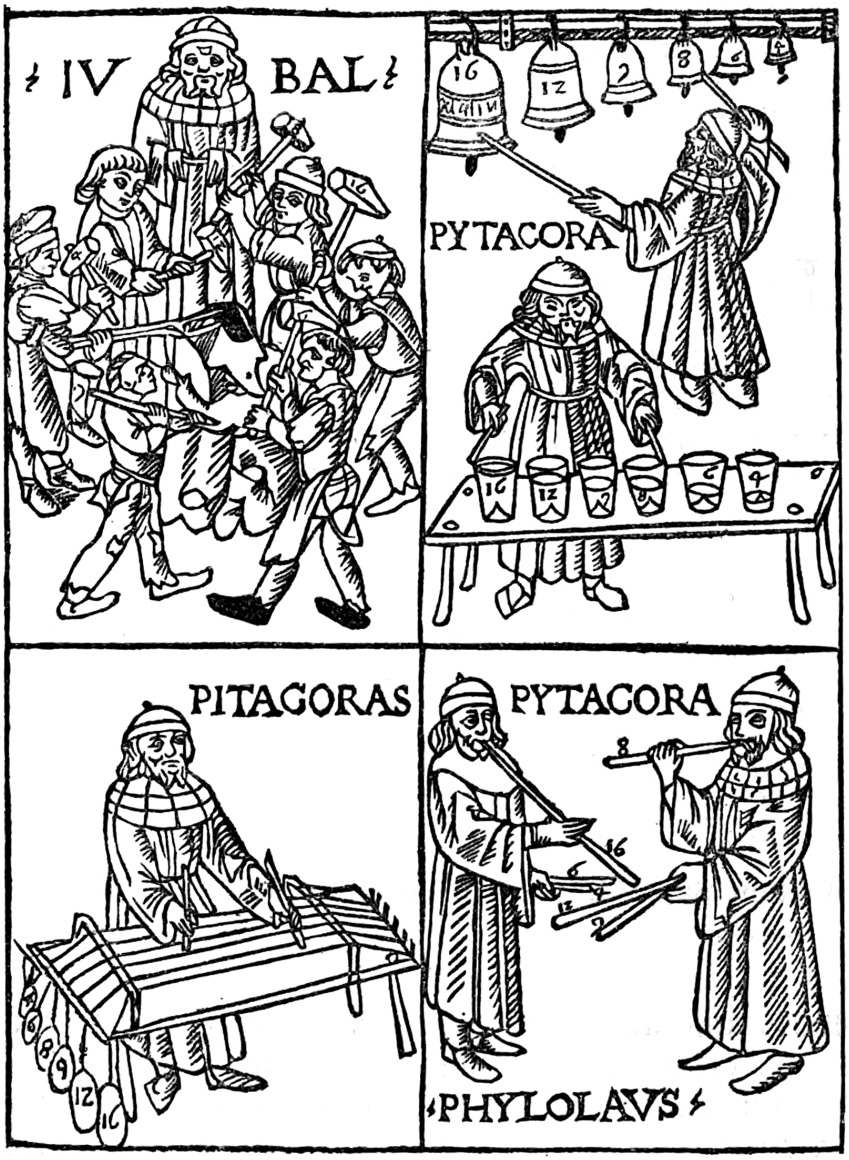
Our approach throughout the book is to describe human hearing via machine models. Much of the hearing research field takes a different approach. A huge amount of experimental data has been amassed by auditory psychologists and physiologists, along with a variety of theories, hypotheses, descriptions, and explanations of the data. In this chapter, we attempt to frame this knowledge, to summarize some of the history of attempts to model and explain it, and to connect it to our machine models. We believe that the true test of our "wisdom about the human auditory system" will be the success of our models in replicating important features and functions of hearing, not just in controlled experiments, but in successful applications that process real-world sound mixtures.

The particular problem that Schouten focused on was pitch perception, which has been one of several key problems in hearing, including early studies of music as illustrated in Figure 4.1. Pitch is probably the single most important problem in leading to the auditory image approach to machine hearing. We review pitch and several other aspects of human hearing in this chapter.

4.2 Auditory Physiology

Most of what is known about human hearing is either from psychophysical experiments, or extrapolated from physiology experiments on animals. Cats, gerbils, guinea pigs, ferrets, and other laboratory animals have been extensively studied, and we have reason to believe that what we learn from them extrapolates well across other mammals.

Early studies of *auditory evoked potentials*, electrical signals picked up by electrodes in or near the cochlea or various nerves or brain structures, played a big role in theories of hearing. When it was discovered that intelligible speech could be reproduced from the evoked potentials near the auditory nerve, the idea that nerves couldn't carry frequencies above a few hundred hertz had to be revised. Tones up to 4 kHz were being reproduced as such in the cat's auditory nerve evoked potentials (Wever and Bray, 1930a).



GAFORUS. *Theorica musicae.* Mediolani 1492.

Figure 4.1: The observation by Pythagoras that small integer ratios of string lengths or tensions, or of reed flute lengths, led to consonant notes, and that bells and glasses of water could be tuned to corresponding pitch ratios, was celebrated in this woodcut from Franchino Gafurio's *Theorica Musice* of 1492. Stillingfleet (1771) points out that the traditional story about Pythagoras and the hammer weights that he used to tension the strings is incorrect, since the string tensions would have to be set in the squares of those ratios to get the consonant tone intervals described.

Another huge breakthrough with cats was the ability to record action potentials, discrete firing events, from single fibers of the auditory nerve, in response to a variety of stimuli at a wide range of levels (Kiang, 1965). The exquisite waveform detail that the cochlea sends to the brain was apparent in Kiang's recordings, as summarized by *peri-stimulus-time histograms*, which approximate the probability of nerve firings as a function of time within a repeating stimulus.

A few years later, using squirrel monkeys, Rhode (1971) developed techniques to watch the mechanical response of the cochlea's basilar membrane, to the point that he started to be able to see the response to even fairly low-level sounds. The normal nonlinear compressive behavior of healthy cochlear mechanics was first seen in these experiments—earlier mechanical experiments were seeing the passive behavior of dead or dying cochleas, or were at such a high intensity that the response was essentially passive. Yet for a long time after Rhode's observations across a wide dynamic range in healthy cochleas, there remained a disconnect between the sharpness of mechanical tuning, which seemed broad, and neural tuning, which seemed sharper. In succeeding decades, mechanical experiments were refined and replicated with newer techniques, until the disconnect was resolved: both kinds of measurements are equally sharp when plotted the same way, as iso-response or frequency–threshold curves with comparable response criteria and equivalently healthy preparations. Both the mechanical and neural responses measured this way are quite a bit sharper than the rather unsharp plots of response versus frequency at a fixed intensity. In linear systems, such measurements would be equivalent, and their curves equally sharp. Therefore, the sharpness of neural frequency–threshold curves eventually came to be understood largely as a by-product of the cochlea's nonlinearity and the way tuning curves are measured, and not in conflict with the less sharp plots measured differently—you just can't compare one kind of plot to the other, as we discuss in Section 10.5, when the system is healthy and behaving nonlinearly.

Studies of single-unit and evoked-potential recordings in numerous different structures of the auditory nervous systems of animals have added a treasure trove of data, but not always a clear picture. Some of the clearest results come from specialized animals, not necessarily even mammals. For example, the barn owl's ability to swoop down and catch running mice in total darkness, just by listening, has been explained with the aid of registered auditory/visual space maps in the optic tectum, to which interaural time and intensity difference signals decoded by lower structures are projected (Knudsen and Knudsen, 1983; Konishi et al., 1985; Sullivan and Konishi, 1986).

In recent decades, another valuable tool has been the study of *oto-acoustic emissions*—actual sound produced by the cochlea and coming out of the ear—for assessing both the function of the human cochlea and the plausibility of models (Whitehead et al., 1996; Epp et al., 2010).

4.3 Key Problems in Hearing

Many introductions to sound and hearing will tell you that a sound, or a musical tone at least, has a pitch and a loudness, and maybe a timbre or other properties. Too often, the notion of a sound will even be reduced to a sine wave, in which case it's almost completely safe to say that the perceived pitch is determined by the frequency, and that for any fixed frequency the loudness is determined by the amplitude or power (for example, as measured by *sound pressure level*, SPL, of the sound). These relations seem simple enough, but for sounds that are not sine waves—which means for essentially all sounds—the picture is considerably more complicated.

Consider loudness. Why do sounds of different frequencies have different curves relating loudness to sound power? Why is the loudness of a band of noise so dependent on the bandwidth, for large bandwidths, but relatively independent of bandwidth when the bandwidth is small? The psychophysical experiments that provide data to answer these questions are diverse and complicated, and devising an explanation for how perceived loudness relates to physical stimulus parameters has been a long-standing important problem in

hearing science.

Consider pitch. For periodic vocal and musical sounds, the perceived pitch usually corresponds to the repetition rate of the pressure waveform; the pitch of a 100 Hz sine wave is the same as the pitch of almost any wave shape repeated 100 times per second. For some wave shapes, however, the pitch will be near 200 Hz, even if there is no 200 Hz periodicity. For example, alternating positive and negative pulses, corresponding to a sum of only odd harmonics of 100 Hz, will typically be matched by a subject to a pulse train near 200 Hz, especially if the lowest frequencies are filtered out or masked by noise. The subtleties of trying to describe the connection between physical acoustic attributes and perceptual pitch for odd stimuli like these have driven much of the progress in hearing in the last few centuries.

Timbre is typically defined as whatever differentiating percept is left in the sounds of tones of equal loudness and pitch and duration; the old dimension of *volume* has not generally survived except as a synonym for loudness. Thus many perceptual differences are lumped together as timbre (from the French), or *tone color* (from the German *Klangfarbe*). The notes of instruments were sometimes called their "clangs" in English. Tyndall (1867) suggested we call their differences "clang-tint," but that didn't catch on the way the French *timbre* did.

The study of musical consonance and dissonance is another key problem that has motivated research and progress in hearing. Even the ancient Greeks knew that small integer ratios of string lengths or pipe lengths would lead to consonant sounds, as illustrated in Figure 4.1. It was much later that these lengths were associated with frequencies or pitch periods. The observations and interpretations were elaborated over many years in the development of music theory, and hearing science has generally scrambled to catch up with explanations. With a modern hearing model that generates auditory images, it is now possible to see how consonance and dissonance may be represented in the brain, potentially taking the problem out of the domain of numerology and into the domain of pattern recognition.

Speech communication is perhaps the most important type of sound that humans rely on in normal life. Characterizing and understanding speech perception has been a key problem in industry and academia for more than a hundred years, initially driven by the needs of the telephone business. It's complicated enough that we'll barely be able to touch on it here.

The other most important use to which humans put their hearing is to be aware of where things are going on around them. "The function of the ears is to point the eyes," as Wenzel (1992) says. This notion relates to the discovery of registered auditory and visual maps of space in the tectum of barn owls (Knudsen, 1982), and is supported by multispecies correlations between auditory localization acuity and the width of the region of best vision (Heffner and Heffner, 1992). The study of binaural hearing is therefore a big part of the attempt to understand how hearing works in general.

The natural tendency of humans to parse sounds into *auditory streams* is a higher-level function that underlies much of music perception and the ability to follow speech in interfering sounds. The *cocktail party problem* refers to a human's ability to extract useful meaning from mixed speech and music sources, sometimes even attending to more than one stream at a time. This ability is also known as *auditory scene analysis*, by analogy with the ability to analyze complex visual scenes (Bregman, 1990).

Besides these psychoacoustic aspects of human hearing, research and progress in hearing has also been driven by findings from auditory neurophysiology, especially in mammals, but also in birds, reptiles, fish, and insects. Recordings of action potentials from single neurons in the auditory nerve of cats, for example, provided early key insights into the function of the ear, and data to constrain theorizing about how to explain many properties of hearing.

The brain regions that process sounds are thought to be organized, like those for vision, into a *what* pathway and a *where* pathway (Rauschecker and Scott, 2009); the *what* pathway deals with classification of sounds, for example of different vowels or musical instruments, while the *where* pathway deals with location and orientation in space. The interactions of these pathways are complicated, as they don't remain entirely

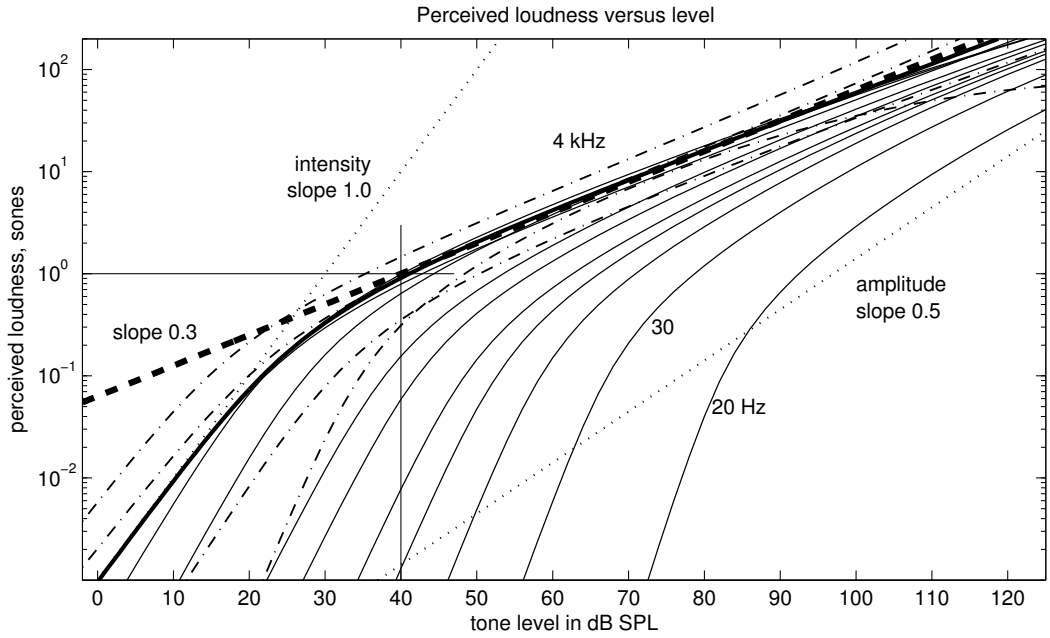


Figure 4.2: Perceived loudness in *sones*, as functions of sound intensity, for selected tone frequencies. The heavy diagonal dashed line (slope 0.3) is based on the conventional approximate definition that loudness in sones is proportional to the 0.3 power of intensity, with 1 sone at 40 dB SPL, at 1 kHz. As the upper right portion of the figure shows, this is a fair approximation at high enough intensities and high enough frequencies. The other curves are based on a *stabilized power law*, using a stabilizing offset corresponding at each frequency to the power level at the 20 phon curve of Figure 4.3; a power-law exponent of 0.28 is used. The solid curves are for frequencies up to 1 kHz (20, 30, 40, 50, 60, 80, 100, 200, 400, and 1000 Hz), with the 1000 Hz curve heavier; the dash-dot curves are for 2, 4, 8, and 15 kHz. The dotted line at the far left (labeled "intensity") shows the slope that would correspond to a linear relationship between intensity and loudness; this slope is approached at low intensities. The dotted line at the far right (labeled "amplitude") shows the slope that would correspond to a linear relationship between pressure amplitude and loudness (intensity power-law exponent 0.5); very low frequencies come close to this slope at high levels.

separate at higher cortical levels. Understanding psychophysical effects in terms of such physiological organization is another key problem in hearing research.

The concepts of linearity and nonlinearity come up in many of these areas. The action of neurons is easy to accept as nonlinear. The idea that nonlinearity is important even in the earliest mechanical parts of the hearing process is more surprising, though it has been a point of discussion for well over a century.

All of these topics require a certain level of familiarity and understanding to see how they motivate and constrain the design of machine hearing systems. The holy grail in this field is to come up with simple machine models, from which emerge behaviors that agree with all the complicated experimental details of human and animal hearing.

4.4 Loudness

Neither the Weber–Fechner law, that perceived loudness varies as the log of the sound intensity (Garrison, 1914; Howell, 1915), nor the Stevens law, that perceived loudness varies as a power-law function of intensity

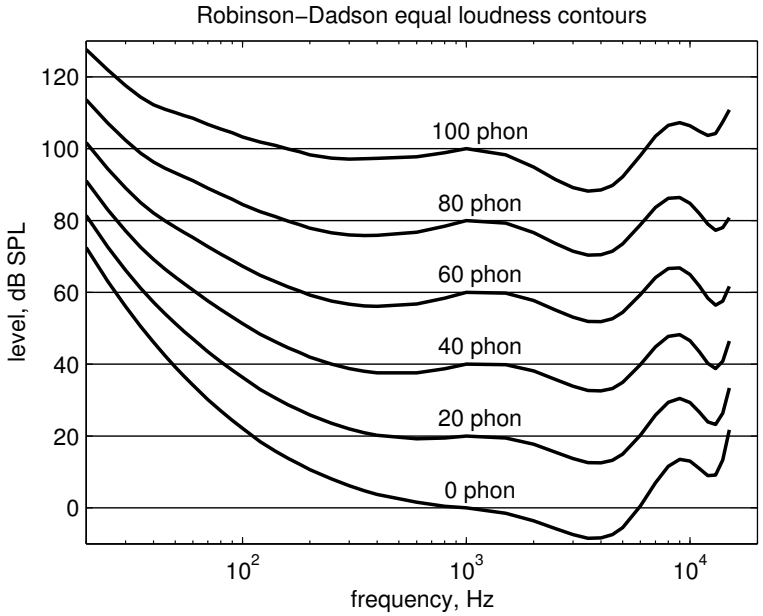


Figure 4.3: The equal-loudness contours known as Robinson–Dadson curves, plotted from their original model parameters (Robinson and Dadson, 1956), map intensity in dB SPL to a loudness-related log-like measure, *phons*. The dB SPL scale is an objective intensity scale, relative to the nominal 1 kHz threshold sound pressure of 20 μpascal RMS, and the phon scale is defined to be equal to dB SPL for a 1 kHz sine wave, but to connect other frequencies and intensities of the same perceptual loudness. The equal loudness contours do not show how the percept of loudness grows with level—that is the function of the *sone* scale.

(Stevens, 1961), is particularly accurate over the whole range of audible frequencies and levels. And loudness behaves somewhat differently for sine waves than it does for noise (Allen and Neely, 1997).

The psychophysical sensation of loudness is quantified in *sones*, which should not be confused with *phons*, which is basically just a dB scale of intensity that is warped to correlate better with loudness for frequencies away from 1 kHz.

By convention, 1 sone is the loudness of a 1 kHz sine wave at 40 dB SPL, and, as a first approximation, the loudness (in sones) is proportional to the 0.3 power of intensity, at least for frequencies not too far above or below 1 kHz, and for signals sufficiently far above the absolute detection threshold and sufficiently far below the threshold of pain. With the 0.3 exponent, each 10 dB intensity difference gives a doubling or halving of loudness, so a 50 dB increase to 90 dB SPL is five doublings, or 32 sones. A more detailed model of loudness in sones as a function of frequency and level is illustrated in Figure 4.2.

The Fletcher–Munson *equal-loudness contours* (or the later Churcher–King or Robinson–Dadson curves) show another way to map loudness, in *phons* (Eargle, 1994). See Figure 4.3. The phon scale is not a loudness scale the way the sone scale is, but rather a tool for tying together intensities of sine waves at different frequencies with the same loudness. The curve for 40 phon, for example, connects the points (frequency–intensity pairs) that have the same loudness as a 40 dB SPL tone at 1 kHz. These curves can be used to map intensities of different frequencies to loudness-equivalent intensities of 1 kHz, such that the Stevens power law with exponent 0.3 can be adapted to apply with more accuracy to a wider range of frequencies. A very detailed loudness model that includes a stabilized power law mapping with exponent slightly below 0.3 is given by Moore et al. (1997).

Weber's and Fechner's and Stevens's Laws and Loudness JND

The threshold for detection of loudness differences, the *just-noticeable difference* (jnd) of loudness, is often expressed as the *Weber fraction* of intensity: $\Delta I/I$. Weber's law says that this fraction is constant, across a wide range of intensity. This law is related to Fechner's law, that sensation scales as the log of intensity: they become equivalent on the assumption that the jnd is an equal increment of sensation level at all intensities.

Experiments find a "near miss to Weber's law," in which the Weber fraction decreases somewhat with increasing level, but not as quickly as the Stevens power law would suggest. For sine waves of mid frequencies, the Weber fraction decreases from about 30% near 20 dB SPL to about 10% near 90 dB SPL. For broadband sounds such as white noise, the fraction is even closer to being level independent. At very low levels, near the threshold of hearing, of course it must be much larger. If we were to predict a Weber fraction on the assumption that sensation scales with a Stevens power law, with exponent 0.3, and assuming a jnd is a given sensation increment, we would have a much further miss from Weber's law. Taken together, these observations suggest that the idea that a jnd is a constant increment of sensation is not plausible. It would be more accurate to say that the jnd corresponds to a constant Weber fraction of sensation: $\Delta S/S$. At least, it would be a near miss, with jnd always around 3–10% loudness change, across a wide range of loudness.

Krueger (1989) proposes this compromise:

> Fechner and Stevens erred equally about the true psychophysical power function, whose exponent lies halfway between that of Fechner (an exponent approaching zero) and that of Stevens. To be reconciled, Fechnerians must give up the assumptions that Weber's law is valid and that the jnd has the same subjective magnitude across modalities and conditions; Stevensians must give up the assumption that the unadjusted (for the use of number) magnitude scale is a direct measure of subjective magnitude.

Loudness versus intensity, and the jnd of intensity, are complex aspects of loudness perception. More interesting, perhaps, is the question of how loudnesses combine when multiple sound signals are added. For signals that are similar enough, adding them is equivalent to a change of intensity, and the loudness follows its usual intensity pattern. But if the signals are different, in that they have different frequency content, such as different sine-wave frequencies or different noise frequency bands, then the loudness increases faster than would be predicted by just considering the intensity, or acoustic power, of the combination—which brings us to critical bands ...

4.5 Critical Bands, Masking, and Suppression

The notion of a *critical band* or *critical bandwidth* can be invoked to explain several psychophysical observations, including the effect of signal spectrum on perceived loudness. The critical bandwidth represents the bandwidth of cochlear filtering, defined as roughly the bandwidth within which frequency components interact strongly, as opposed to being treated relatively independently. The definition or estimation of the critical bandwidth can depend on the problem being considered, such as loudness combination, masking, roughness, etc. The notion of critical band should not be interpreted as implying a bank of filters that divide the spectrum into a discrete set of bands.

As an example application of the critical band concept, for added signals with similar frequency content, as measured at the resolution of critical bands, their *intensities* add, and the resulting loudness is roughly a power-law function of the total intensity; whereas for signals with very different frequency content, that is, where two signals have little energy in common within any critical band, their loudness *sensations* add instead. Since intensity is strongly compressed to give sensation, the sensation of loudness will increase only slowly for the case of sounds of similar spectra. On the other hand, if the sounds being combined are separated by more than about a critical bandwidth, their sensations will be (more or less) independently computed as compressive functions of their intensities, and those loudness sensations will add. The distinction between these cases is never so clean or simple, but plots of perceived loudness versus noise bandwidth or tone separation will show a soft *corner* near the critical bandwidth. Other perceptual measures as well show changes when tone separation or bandwidth crosses roughly a critical band, supporting the idea that the band corresponds roughly to the analysis bandwidth of the cochlea. The critical bandwidth is somewhat less than one-third octave at frequencies above a few hundred hertz, and considerably more than one-third octave at lower frequencies. Exact estimates vary, depending on the task, and to some extent on the intensity level of the experiment. The critical band is a bridge between linear and nonlinear aspects of hearing; the dependence of the critical bandwidth on intensity is a reflection of nonlinearity even in the analysis filter stage of the hearing process.

For a sound with a steady spectrum (or steady power spectral density in the case of a noise-like sound), the loudness can be fairly well modeled in terms of the spectrum as analyzed by a bank of filters with widths of about a critical band (or one-third octave in the typical engineering approximation). The power in each band is mapped through a compressive power-law nonlinearity, according to the Stevens power law model of sensation as a function of stimulus intensity, and then added up to give a number that correlates well with perceived loudness. This model is known as a Zwicker-type model, as it is based on the research of Eberhard Zwicker and his colleagues, who actually codified the method as an international standard loudness computation algorithm (Zwicker and Scharf, 1965; Zwicker et al., 1984).

Improvements to these models account for mutual *suppression* and *masking* between the signals in adjacent or nearby filter channels. Masking refers to the observation that one sound (the masker) can reduce the loudness of another (the probe), or can even make another sound completely inaudible (often described as raising the threshold for detection of the probe sound). Suppression, on the other hand, is a physiological, rather than psychological, effect: one sound can suppress the response to another simultaneously presented sound, and even suppress the total response, measured at various places in the auditory nervous system or even in the mechanics of the cochlea. A reduction in response can also be caused by a sound presented earlier than the probe; the physiological mechanisms include what is called *adaptation*, to distinguish it from the suppression caused by a simultaneous sound.

Two main types of masking are important. *Simultaneous masking* is an interaction of concurrently presented sounds (as in a Zwicker-type model for the loudness of steady sounds). *Forward masking* is the effect whereby after a louder sound stops, its lingering effect can still mask a weak sound that comes later, by up to a few tens of milliseconds (longer-time effects are usually called fatigue or adaptation). The extent to which

suppression and adaptation can explain the difference between simultaeous and forward masking is still being explored (Rodríguez et al., 2010).

There is also some *backward masking*, but it is a much weaker effect, since much of the work to detect a weak sound is done before the stronger sound arrives. We will ignore it.

Simultaneous masking is highly asymmetric with respect to the frequencies of the two sounds, with low-frequency sounds masking high-frequency sounds more than the other way around. When Alfred Mayer (1876) first reported the effect, he stated that it was *completely* asymmetric, under these topics:

> 1. On the Obliteration of the Sensation of one Sound by the simultaneous action on the ear of another more intense and lower sound.
> 2. On the Discovery of the Fact that a Sound, even when intense, cannot obliterate the Sensation of another Sound lower than it in pitch.

His initial higher- and lower-pitched experimental sounds were the ticks of his watch and his clock, respectively, neither of which has a real pitch; but they do have rather different spectra, with the tick of the small watch having higher frequencies than the tick of the large clock. By setting the two time pieces to run at slightly different rates, he could let the ticks move slowly on top of each other and away from each other; by moving the watch and clock to different distances, on a quiet night, he could map out the conditions under which the clock tick could obliterate the watch tick:

> The general result of the numerous experiments thus made shows that the sensation of the watch-tick is obliterated by a coincident tick of the clock when the intensity of the clock-tick is three times that of the watch-tick. This result, however, must be regarded as merely approximative, not only from the manner in which it was obtained, but from the complexity of the sounds on which the experiments were made. It is interesting, however, both as being, I believe, the first determination of this kind that has ever been made, and as having opened out a new and important field of research in physiological acoustics.

Experiments with tuning forks and organ pipes further convinced him that a high tone could not mask a low tone. With more modern instrumentation, Wegel and Lane (1924) showed that this asymmetry is not absolute, and that the patterns of masking are complicated. Many others have worked to characterize masking since then. The critical band concept was actually proposed as part of an explanation of masking, not of loudness summation (Fletcher and Munson, 1937; Fletcher, 1940). Sounds within a critical bandwidth can mask each other strongly; beyond a critical band, masking rapidly diminishes.

For nonsteady sounds, such as speech and music, the estimation of perceived loudness is more complicated, particularly as it needs to include the effects of forward masking. The datasets from many experiments on such sounds have served to motivate and evaluate loudness models of increasing sophistication, known as dynamic loudness models (Chalupper and Fastl, 2002).

Increasingly, models of loudness come to resemble full-blown models of the functions of the auditory system, with asymmetric masking between filter channels and with forward masking in the dynamics of the sensation estimation. Loudness estimators become auditory filterbank models; the experimental data serve to calibrate and evaluate these models.

It is tempting to think that perceived loudness correlates with the total firing rate of afferent auditory-nerve neurons—this is known as the *Fletcher–Munson hypothesis*, after their 1933 conjecture (Fletcher and Munson, 1933). Modern physiological data and analysis show, however, that this simplified view does not work very well (Allen and Neely, 1997)—especially at high loudness levels where the firing rates are all essentially saturated at their maximum, yet listeners still distinguish loudness changes with a Weber fraction

$\Delta I/I$ not too far from what it is at more moderate levels (Delgutte, 1996; Heinz et al., 2005). In a high-loudness region, increased loudness of narrowband sounds leads to a spread of activation to fibers that are not very sensitive to the frequencies in the sound, so the total rate may still increase. But for wide-spectrum sounds, when most fibers are firing near their saturation rates, loudness must be coded on the auditory nerve in some way other than total firing rate: either in the temporal patterns of firing, or in the firing rates of the smaller number of high-threshold, low-spontaneous-rate neurons, or in the firings of neurons involved in loudness level adaptation, or some combination of these, to supplement the information in the saturated high-spontaneous-rate neurons.

4.6 Pitch Perception

Pitch is a percept closely associated with frequency and measured in the same units: hertz, or cycles per second (sometimes just called cycles in older works). For sound signals that are periodic and contain low harmonics, the perceived pitch is equal to the fundamental frequency, the reciprocal of the period. The idea of a fundamental and harmonics is based on a decomposition into sine waves; Helmholtz thought that the fundamental was needed to evoke a pitch, and that pitch would be perceived through a sympathetic resonance in the cochlea, each distinguishable pitch having a resonator that would be excited by the fundamental frequency associated with that pitch. Various observations and demonstrations of good pitch perception from signals with a weak or missing fundamental led to many arguments against the Helmholtz theory, and many alternative theories to supplement or replace it.

For sounds that are not periodic, or for sounds that are periodic but are missing not just the fundamental but also other low harmonics (second through tenth harmonics, say), pitch is more complicated. There is a huge body of experimental data on pitch perception with different types of signals, and a vibrant history of attempts to explain perceived pitches that are not simply the reciprocal of the sound signal's period. Generally, subjects are very good at making high-precision pitch matches, by adjusting the pitch of one source to match that of another. The sources generally do not sound alike even when they have the same pitch; besides pitch and loudness, sounds are distinguished by their other characteristics, usually lumped under the term *timbre*, which for steady tones can be associated with spectrum or waveform shape.

To first order, pitch is easily explained by autocorrelation or by power spectrum: find the delay that makes the signal waveform most nearly resemble a delayed copy of itself, and the reciprocal of that delay will be the pitch frequency; since the autocorrelation function is the inverse Fourier transform of the power spectrum, this is equivalent to finding a pattern of equally spaced peaks (a "ripple") in the power spectrum that looks most like a sequence of harmonically related frequencies. These temporal and spectral techniques will handle the missing fundamental, addition of noise, a little bit of jitter relative to perfect periodicity, and other pitch phenomena. But the psychoacoustic experimental data are good enough to easily point out signals for which this first-order model is inadequate: complexes of nonharmonically related sinusoids, filtered noises, inharmonic musical instrument notes, long-period waveforms where the pitch does not match the repetition rate, etc. Explaining pitch in these marginal cases requires consideration of the functional properties of the ear: the cochlea's frequency analysis, the hair cells' detection nonlinearity, and the auditory nerve's limited firing synchrony.

A long-investigated case is the pitch of the *strike note* of a chime (a tubular bell) or a carillon bell (a classic church bell). The strike note is the sound dominated by prominent partials (single-frequency components) in the early part of the bell's ringing after it is struck, and is associated with the musical pitch that the chime would be used for in a melody, even though the bell does not produce a periodic or harmonic sound. After studying the sounds of church bells for years, Jones (1928) focused on the frequency relationships between the fifth, seventh, and tenth partials of the bell sound, and concluded "that the relative frequencies of these partials are not far from 2:3:4, thus perhaps giving rise in the ear of the observer to the pitch of the strike

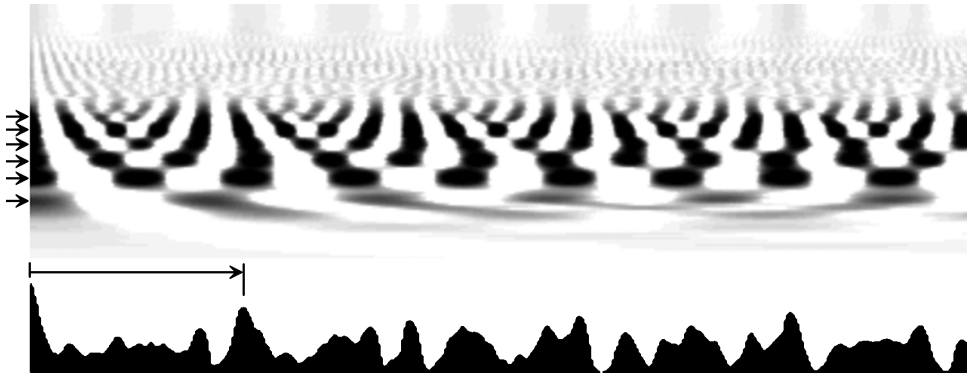


Figure 4.4: One frame of a Licklider-style auditory image of the orchestral chime sound from the Acoustical Society of America's *Auditory Demonstrations* CD (Houtsma et al., 1987), showing several strong partials (nearly-sinusoidal components, from the resonant "modes" of the bell) that contribute to a strong peak at the delay corresponding to the perceived strike note pitch period. The upper part of the figure is the auditory image, or correlogram, with vertical axis corresponding to cochlear place, or frequency, and horizontal axis representing the delay or lag parameter, measured from the left edge; see Chapter 21 for details on such images. The graph at the bottom shows the sum across frequency channels; peaks in this graph are likely pitch periods; the period corresponding to the strike note pitch is indicated by an arrow. The arrows on the left indicate the positions of strong partials along the frequency axis. Several strong partials are close to a 2:3:4:5 relationship, but not quite harmonic.

note." In spite of the intervening other partials, this alignment of frequencies gives the listener the percept of a pitch of the approximate "missing fundamental." This result foreshadows many years of work on the pitch of sounds with three successive nearly harmonic components. As shown in Figure 4.4, the stabilized auditory image, or auditory correlogram, as proposed by Licklider, gives us a good way to visualize and assess this relationship of nearly-aligned partials.

The pitch of periodic signals with only higher harmonics has been called *residue pitch*, or the *pitch of the residue* (Schouten, 1970). Somewhat more generally, the pitch of a signal missing a fundamental has been called *periodicity pitch*, or *virtual pitch* (Terhardt, 1974), to distinguish it from pure-tone pitch. Residue pitch can be heard in the 40 to 800 Hz range (findings on these numbers vary, depending on conditions such as how many low harmonics are missing) even when the signal components are all above 4 kHz, too high for the auditory nerve to synchronize to the spectral components, even with the volley principle (as described in Section 2.2). Whether the residue pitch is heard or not, or whether it is strong or weak, depends on the relative phases of the harmonics, not just on their amplitudes, when the component frequencies are high. These phase differences have no effect on the power spectrum or the autocorrelation function, so the perceived differences can't be explained by that model. Rather, it seems to be more the time-domain *envelope* of the waveform that matters in such cases. If the phases are such that the envelope has strong periodic peaks and valleys, the residue pitch will be heard, whereas if the envelope does not, then a low pitch will often not be heard (Licklider, 1956). For lower component frequencies (say around 2 kHz), to which the auditory nerve can synchronize, the envelope still matters, but the detailed waveform synchrony is more important in determining the pitch percept. A two-tone complex has a modulated envelope that evokes a weak pitch percept; more components typically give a more pronounced envelope modulation and a more salient pitch, as in the three-tone complex illustrated in Figure 4.5.

Repetition pitch, or reflection pitch, is a clear pitch percept associated with a noise added to a delayed copy of the same noise, as in hearing a noise source and its reflection off a wall behind it. Depending on the noise spectrum, the delay, and the attenuation of the echo, there is a complicated space of possible results on

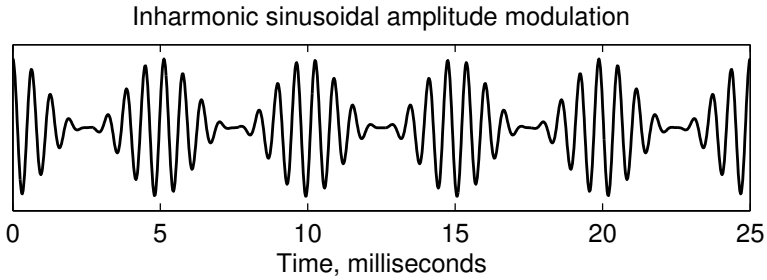


Figure 4.5: A three-component tone complex can be made by modulating a carrier with a slower envelope (here a 1560 Hz carrier and 200 Hz modulator). When the carrier and modulator frequencies are not harmonically related, the resulting signal is not quite periodic. The perceived pitch is usually close to the modulator frequency, but is pulled toward the time interval between a pair of carrier waveform peaks; which interpeak interval is used can sometimes be ambiguous, and pitch matches are sometimes ambiguous. Here the carrier is close to 8 times the modulator, so the interval between corresponding peaks is about 8 cycles of 1560 Hz: $8/1560$ s. According to the *first effect of pitch shift*, the perceived pitch is close to the reciprocal of this interval: $1560/8 = 195$ Hz; but according to the *second effect of pitch shift*, a subject will match it closer to a pitch that is the 7th subharmonic of the lower sideband tone frequency: $(1560 - 200)/7 = 194.3$ Hz. This fractional-hertz difference is enough to show up as a significant effect in pitch-matching experiments.

pitch and on the strength of the pitch percept. Further confusion is created if the echo is added with a negative polarity; this puts a dip in the autocorrelation function at the point corresponding to the delay, rather than a peak. By analyzing such signals, Bilsen and Ritsma (1970) found evidence for an explanation that merges the Helmholtz resonance–place theory of the cochlea with a correlation-like temporal analysis approach to pitch perception:

> ... evidence is presented that underlines the concept that both repetition pitch and residue pitch are the result of a combined frequency and time analysis in the hearing organ; viz., the perceived pitch appears to correspond to the reciprocal value of the time interval between two prominent positive peaks in the temporal fine structure of the displacement wave form evoked by the signal at a dominant frequency region on the basilar membrane.

Licklider (1951, 1956) had previously published related ideas in the *duplex theory* and *triplex theory* of pitch perception. Licklider's duplex theory should not be confused with Rayleigh's duplex theory of binaural localization, which we discuss in Section 22.1. In these theories, Licklider sought to bring together the pitch of sinusoids or fundamental components (Helmholtz's place pitch) with residue or repetition pitch, and in the triplex theory also the *Huggins pitch* or *dichotic pitch*, a binaural effect. Huggins had presented a structureless white noise to both ears, with one being a copy of the other except for a phase reversal above a selected frequency in the range of 200 to 1600 Hz; the result was a weak pitch percept near the phase reversal frequency, that could only have been due to interactions between the two signals, presumably in the auditory brainstem (Cramer and Huggins, 1958).

In both the Bilsen and Ritsma scheme and the Licklider scheme, the period estimation, or autocorrelation, focuses on movement of the cochlea's basilar membrane in one direction, and ignores the other direction. That is, the motion is *half-wave rectified* (keep the positive parts, and set the negative parts to 0): the "time interval between two prominent positive peaks in the temporal fine structure of the displacement wave form evoked by the signal" is a concept that works either on the original or the rectified signal. But implementing it via autocorrelation of cochlear responses works better if the cochlea's response is rectified and smoothed

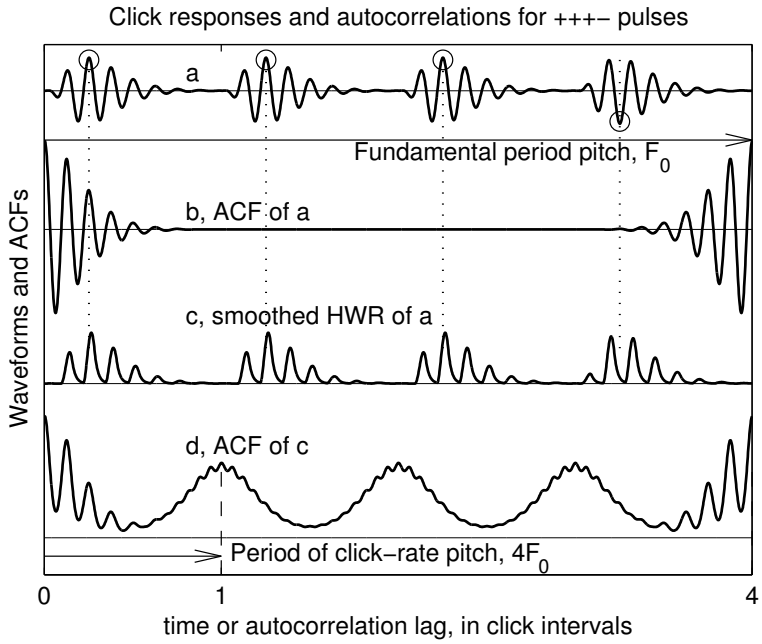


Figure 4.6: The effect of half-wave rectification (HWR) on the autocorrelation function (ACF) of the response to a click train with polarity sequence "+++–": (a) the bandpass-filtered click train; the clicks may be delivered as a sound that looks like this waveform, or as a broadband click train to which a point in the cochlea responds this way; the polarities can be seen by comparing the waveform at the times of the equally-spaced dotted lines. (b) the ACF of (a), showing maxima at zero and at the period, 4 interclick intervals. (c) the result of HWR and smoothing of (a). (d) the ACF of (c). One period of four clicks is shown, since the signals and ACFs repeat with this period. Depending on the parameters, such as the bandpass center frequency and click rate, the perceived pitch may correspond to either the interclick interval (dashed line at lag 1), or the period (at lag 4). The HWR is needed to get the ACF to explain the pitch perception at the interclick interval, since the ACF (b) has no peak at that period.

a bit first. Consider high-frequency responses to trains of clicks: there will be prominent peaks in clusters, separated by the interclick interval. When some of the clicks are inverted, as illustrated in Figure 4.6, the exact timings of the peaks will shift a bit, but the intervals between the clusters won't; the smoothed rectified signal's autocorrelation function will have peaks near the interclick interval even if some of the clicks are negative.

If the rectifier were missing, then for some stimuli like clicks in a repeating "+ + + –" polarity sequence, autocorrelation would merge opposite-polarity correlations, completely canceling each other at and near the interclick interval, so there would be no peak left near the interval that corresponds to the perceived pitch. Flanagan and Guttman (1960), Guttman and Flanagan (1964), Rosenberg (1965), and Pierce (1991) studied human pitch perception with such signals, pointing out that the ACF is identical for sounds with just a single positive click per period (with twice the amplitude, so same energy as four of the other clicks). The "+1 +1 +1 –1" pattern of the same period therefore differs only in phase from the signal that sometimes has a two-octaves-lower pitch. It is the rectification nonlinearity, matching the nonlinearity of the inner hair cells that detect basilar membrane motion, that breaks the equivalence between these post-cochlea autocorrelation techniques and spectral techniques.

In this sense, the explanations of psychoacoustic pitch phenomena were steadily converging on modern physiological concepts of cochlear function. Licklider in particular made this connection explicit, hypothesizing brain structures that would carry out the operation—*neuronal autocorrelators*—and showing their output as what we now call *auditory images*; he explained (Licklider, 1951):

> The essence of the duplex theory of pitch perception is that the auditory system employs both frequency analysis and autocorrelational analysis. The frequency analysis is performed by the cochlea, the autocorrelational analysis by the neural part of the system. The latter is therefore an analysis not of the acoustic stimulus itself but of the trains of nerve impulses into which the action of the cochlea transforms the stimulus. This point is important because the highly nonlinear process of neural excitation intervenes between the two analyses.

Licklider's approach was largely ignored for more than thirty years, probably due to the difficulty of finding enough computational power to implement it. Eventually, Langner (1981) found evidence for a periodicity detection mechanism in the auditory system of guinea fowl, and I implemented computer models of Licklider's proposal (Lyon, 1984), displaying the first real auditory images.

A further wrinkle in pitch perception is the *second effect of pitch shift*, discovered by de Boer (1956); see the example in Figure 4.5. Schouten (1970) describes de Boer's reported discovery, relative to a more easily understood *first effect*:

> If a residue is obtained by modulating a carrier frequency f of say 2000 Hz with a modulating frequency g of say 200 Hz, the pitch p of the residue corresponds to that g. The first effect consists of a pitch shift Δp proportional to the shift Δf in the carrier frequency [in the same proportion as p to f]. The second effect consists in a slight but systematic deviation of the constant of proportionality. It also consists in a *downward* shift in pitch when the modulating frequency g is raised. The latter effect was met with doubt up to incredulity.
>
> On the basis of modelmaking, de Boer ascribed the first effect to the shift in the fine structure of the modulated sound from one period of the modulating frequency g to the next. As to the second effect, he suggested that somehow the lower Fourier components might carry more weight in the formation of the residue than the higher ones.

Terhardt (1970) mentioned a rule for estimating periodicity pitch (PP) from the frequencies of the partials:

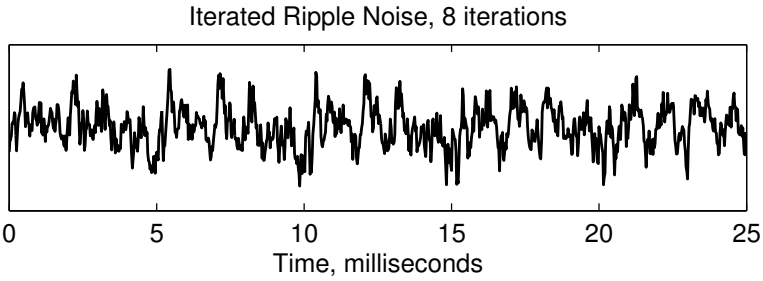


Figure 4.7: Iterated ripple noise (IRN) is another special stimulus signal that has been extensively studied. A noise signal is added to a delayed version of itself, giving a "ripple noise" with a pitch sensation determined by the delay. The signal illustrated here went through 8 iterations of delay and add, with a 5 ms delay; waveform features can be seen repeating with 5 ms separation even though the signal is still a random noise.

> The simple rule for the determination of the PP of any sound with equidistant partials where the lower harmonics are removed is: the frequency f_{PP} corresponding to the perceived PP is that subharmonic of the lowest present partial f_u which lies nearest to the beat frequency f_b between two neighbouring partials.

This rule deviates from the first effect, in the direction that makes it closer to the second effect—but it's still not quite right. As de Boer pointed out, the first effect is consistent with time intervals between peaks in the waveform, and explains the data if the harmonic numbers are low (for example, 400 Hz modulating 1600 Hz, so the central or "carrier" frequency is the fourth harmonic of the pitch) and the second effect is more or less consistent with the same idea if the signal is filtered to emphasize lower frequencies before measuring the time interval between peaks. For harmonic numbers around 7 to 9, waveform peaks of the lower sideband frequency $f - g$ align pretty well with the pitch period, as Terhardt's rule would suggest.

For higher harmonic numbers, such as 12, cochlear nonlinearities seem to come into play: the pitch period aligns more nearly to peaks of $f - 2g$, as if a new lower sideband frequency was added by a cubic intermodulation distortion (Ritsma, 1970). Intermodulation between $f_1 = f - g$ and $f_2 = f$ gives a cubic distortion tone frequency $2f_1 - f_2 = f - 2g$. The amount of second-effect shift corresponding to this newly introduced "lowest present partial" is somewhat level-dependent: at 15 dB SPL, the second effect is negligible. Additionally, different subjects show different amounts of second effect in this region, probably due to differences in the health of their outer hair cells, the presumed site of the distortion product generation.

Smoorenburg (1970) showed a considerably larger second effect with two-tone inharmonic complexes, as opposed to Ritsma's three-tone complexes. He points out that the shifts found accord with the strong third-order $2f_1 - f_2$ and fifth-order $3f_1 - 2f_2$ distortion products, or combination tones, that are detectable with such signals.

Goldstein and Kiang (1968) showed that auditory nerve fibers with *characteristic frequencies* (CFs, the frequencies to which they are most sensitive) slightly lower than the tone frequencies f_1 and f_2 (the *primary frequencies*) do in fact respond with strong synchrony to $2f_1 - f_2$, irrespective of whether the primary frequencies were harmonically related. They implied that time intervals between neuron firings, on the low-CF side of the cochlea's region of response to the tones, would be a plausible cue for the extraction of the perceived pitch, explaining the second effect. They did not check for synchrony to the fifth-order component. More recently, strong fifth-order distortion responses, or synchrony to $3f_1 - 2f_2$, have been observed in cochlear mechanics (Robles and Ruggero, 2001a).

These pitch-shift effects, and the idea of a *dominance region* preferring components closer to about 4 times the pitch frequency, do a fair job of connecting nonlinear cochlear physiology with the psychophysics

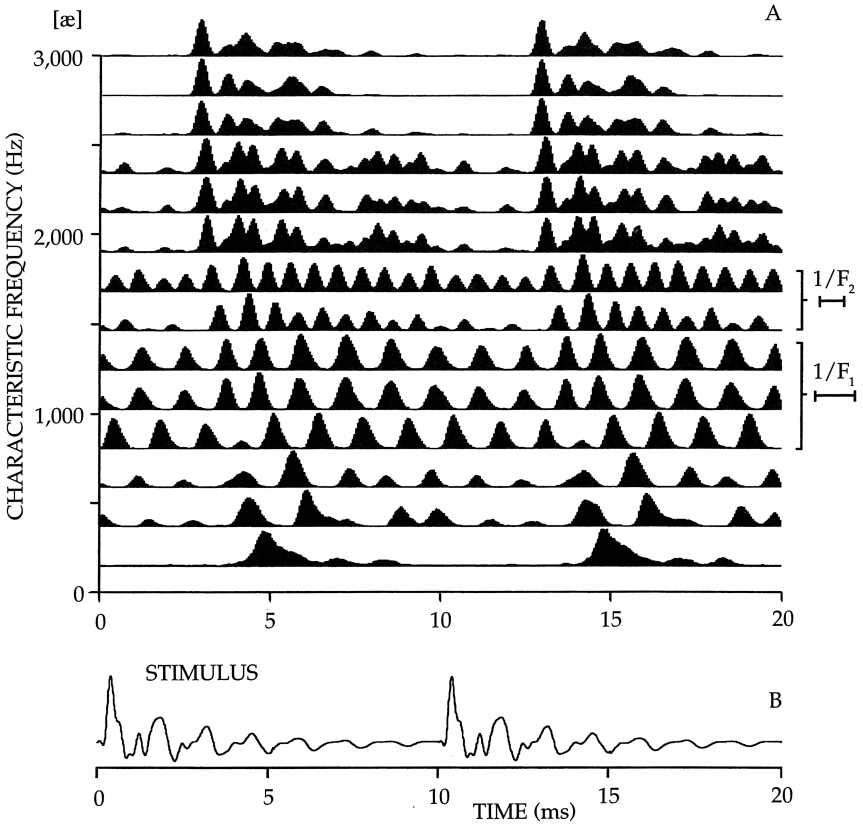


Figure 4.8: Period histograms of auditory nerve fiber firings in response to a periodic vowel sound show pitch-synchronized activity, for fibers of all CFs (Delgutte, 1997). Even fibers that primarily synchronize to the formant (vocal tract resonance) frequencies (here F_1, 8 cycles per pitch period, and F_2, 14 cycles per pitch period) show a pattern that repeats at the pitch rate. Synchrony to the formant frequencies spreads to fibers of higher CF. Fibers with CF above 2 kHz show synchrony to a wide range of lower frequencies, in a pattern prominently synchronized to the pitch rate. The pitch here, 100 Hz, is quite low relative to the cat's auditory-system tuning, so we do not see the resolved low harmonics (2 through 5 cycles per pitch period) that would likely be apparent in human auditory nerve data. [Figure 3 of Chapter 16 (Delgutte, 1997) reproduced with permission of John Wiley & Sons.]

of pitch perception, at least for the class of three-tone inharmonic stimuli. For speech vowel stimuli, several investigators have shown that most primary auditory neurons fire in a pattern locked to the pitch period, both in regions where harmonics are resolved (more than about a critical bandwidth apart) and where they are not resolved (within a critical band), and where the response is dominated by a formant (vocal tract resonance) or not (Young and Sachs, 1979; Delgutte, 1997). That is, the auditory nerve carries the information needed for a temporal explanation of pitch, along with harmonics and timbre. Auditory nerve responses for a variety of other stimuli and conditions, in addition to speech sounds, are shown to support a temporal-code approach as well (Delgutte and Cariani, 1992; Cariani, 1999).

Other special signal types, such as iterated rippled noise (IRN), have been used to investigate the extent to which pitch perception depends on temporal versus spectral cues (Patterson et al., 1996). Using a delay-and-add operation (a *comb filter*) on a noise gives its spectrum harmonically related peaks: a "ripple." Repeating the delay-and-add operation on the ripple noise, with the same delay, gives an IRN with an increased pitch salience. The relative pitch salience for different numbers of iterations and different delays has been studied as a way to compare the explanatory abilities of temporal and spectral pitch perception models (Patterson et al., 1996). An example IRN waveform is shown in Figure 4.7. Studies of auditory nerve responses show correlates of both spectral and temporal cues, but the temporal cues seem to be most prominent in many cases, as illustrated in Figure 4.8 for a speech signal.

Evidence for coding of pitch by time intervals in the fine time structure on the auditory nerve and in brainstem structures has been collected and analyzed by Cariani and Delgutte (1996a,b), and has been found compelling (Cariani and Delgutte, 1996b):

> Taken as a whole, the physiological data presented here provide strong evidence that interspike interval information plays an important role in the perception of the low pitch of complex tones. The predominant interval hypothesis for pitch yields surprisingly robust, comprehensive, and unified explanations for a very wide range of pitch phenomena: the missing fundamental, pitch invariance with respect to level, pitch equivalence of spectrally diverse stimuli, the pitch of unresolved harmonics, the pitch of AM noise, pitch salience, pitch shift of inharmonic AM tones, pitch ambiguity, phase insensitivity of pitch, and the dominance region for pitch. Its main weaknesses are its failure to account for the rate pitches of alternating click trains and its underestimation of the salience of low-frequency tones.

Some surveys of pitch perception experiments and models have come down in support of spectral template matching approaches, as opposed to approaches based on repetitions in temporal fine structure as just described (de Boer, 1976b; Hartmann, 1996); but other more recent studies conclude that only temporal approaches are consistent with experiment (Patterson et al., 1996; Yost, 2009). It is generally acknowledged that explanations of pitch perception need to be in terms of the spatial–temporal coding of sound on the auditory nerve, but approaches unconstrained by that requirement are still sometimes encountered. A careful study of the signals on the auditory nerve (of cats) supports the idea that temporal cues are dominant for low pitches, and neural firing rate versus place is the dominant cue for high pitches, above about 1300 Hz where synchrony begins to fall (Cedolin and Delgutte, 2005); this two-pitch-mechanisms view is essentially Licklider's duplex theory.

The synchrony of auditory nerve firings to a periodic input is sometimes presented as a period histogram, an estimate of instantaneous firing rate or probability constructed from the nerve firings in response to presentation of many cycles of the stimulus, as illustrated in Figure 4.8. But the brain does not have access to these histograms synchronized to the sound period; for general nonperiodic sounds, period histograms don't even make sense. The brain gets the periodic or aperiodic patterns, but not the sync signals that we use to summarize them this way. One way to stabilize the patterns of time is to make an interval histogram: count the numbers of intervals between firings at different delays. A first-order interval histogram (considering intervals

only between immediately neighboring firings) will show a pattern that depends very much on the average firing rate, so will not be stable as loudness changes. On the other hand, an all-order interval histogram—counting all intervals, not just from immediately time-adjacent nerve firings—is essentially equivalent to an autocorrelation function of the firing patterns, as Licklider suggested, and is a robust way to pull out periodicities in the nerve firing pattern. This is the sort of auditory image that many models of pitch perception are based on (Lyon, 1984; Weintraub, 1987; Meddis and Hewitt, 1991; Patterson et al., 1992; Slaney and Lyon, 1993; Cariani, 1999).

While "special" stimuli have been very useful in investigating the limits of human pitch perception, and in evaluating the abilities of models to explain such effects, the sound signals of more general interest that involve a strong pitch sensation are speech and music signals. In speech, and in the sounds of many instruments such as woodwinds, horns, and bowed strings, the pitch is the result of repeated excitation events—glottal pulses for speech, reed or lip pulsations for wind instruments, and stick–slip events for bowed strings (Hartmann, 1998). The events usually come at fairly regular intervals, and produce fairly similar waves each time, so the detection of approximately repeated patterns works very well for such signals, even though they are not exactly periodic.

4.7 Timbre

Timbre, or *tone color*, is what we call differences between tones that have equal pitch and loudness—it's the *everything else* character of sound, which makes it hard to pin down.

Plomp (1970) wrote a good survey of ideas about timbre, in which he particularly emphasized the depth of the study that Helmholtz had done. Like Helmholtz, he simplified the problem by restricting attention to the study of steady periodic tones; others often include things like attack characteristics of musical instrument notes—the "clang"—but this broad definition seldom leads to any tractable theory. Helmholtz had concluded that the timbre is primarily determined by the relative amplitudes of harmonic partials, with relative phases being usually unimportant. We now know a lot more about the conditions under which phase differences can matter, but he was not far off.

The 1973 ANSI definitions give us this rather ambiguous guidance about timbre:

Loudness: "... that intensive attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud."

Pitch: "... that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low."

Timbre: "... that attribute of auditory sensation in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different."

In the pitch section, we pretended that pitch is measured along a low-to-high scale as this definition suggests. However, pitch is really more complicated than that, and is often split up into *pitch height* (sort of a spectral centroid) and *pitch chroma* (the *pitch class*, or position within an octave, sometimes called *tone chroma* or *tonality*). Pitch chroma is what needs to be manipulated to make musical melodies or chords. Both the height and chroma dimensions generally allow a higher–lower comparison, but not always unambiguously. In terms of connecting pitch to the well-ordered keys on a piano, for example, sounds with a rich timbre will sometimes have octave ambiguity; a tone might be a good "C" in terms of its pitch *chroma*, while being ambiguous about its pitch *height*, or which octave it fits with. Nontonal sounds such as bandpass-filtered noise can have a clear height, corresponding to the filter center frequency, but without any match to a note on a musical scale. Chroma and height can be independently manipulated, even in opposite directions. These extra complications of pitch sometimes make it difficult or impossible to order pitches on a scale from low to high. For example, a musical scale of steps of increasing pitch (in terms of musical chroma) can make a sound that ends up lower or the same (in terms of height), using Roger Shepard's sound synthesis tricks (Shepard,

1964). With such tricks, it is possible to construct three tones that subjects will rank $A > B$, $B > C$, and $C > A$; that is, tone pitches are not even partially ordered, in general.

Perhaps, then, pitch height is not really separable from timbre—and chroma, the musically most important part of pitch, does not actually let us order sounds from low to high. There is no simple untangling of these concepts. Speech perception can also be considered an application of timbre perception; each vowel is just a steady tone whose spectrum determines its timbre, and from that its vowel category, almost independent of its pitch.

Timbre depends somewhat on phase, as when the relative phases of a group of harmonics within a critical band are changed in a way that changes the resulting envelope modulation. Changes of amplitudes of harmonics have a much more direct effect on timbre, such that perceptual distances between tones of different timbres correlate well with spectral differences, for example as measured in terms of the log power outputs of a one-third-octave filterbank (Plomp, 1976). This simple log-spectrum approach to timbre is not quite accurate or complete, but captures a good part of the space of variations that we call timbre, and is the basis for sound representations typically used in automatic speech recognition.

Studies of timbre are as diverse as timbre is. Humans are good at recognizing timbre categories, for example to say what instrument is playing a note, but also are very good within categories, as in comparing one violin to another (Dinther and Patterson, 2006). Some aspects of timbre recognition, such as the categorization of speech sounds, clearly involve higher-level parts of the auditory brain, while in other aspects, timbre may simply mean spectrum. We'll leave timbre to be defined differently within different applications that need to distinguish sounds.

4.8 Consonance and Dissonance

The fact that pleasant-sounding *harmonious* or *consonant* musical intervals are based on small integer ratios of pitches has long been known, as Figure 4.1 and Figure 4.2 illustrate. In western musical tradition, relationships between notes are named according to their positions in an seven-tone scale—a scale that evolved to support the production of music with consonant intervals. The first or unison represents notes of equal pitch: 1:1. The eighth or octave represents a doubling of pitch: 2:1. These are the most consonant intervals, but the fifth and fourth are also very consonant: 3:2 and 4:3 respectively. Other intervals come in major and minor versions, depending on the scale, and are generally less consonant as the integers needed to express their ratios increase. The major sixth and major third, 5:3 and 5:4 respectively, are more consonant than the minor third, 6:5.

Integer ratios requiring an integer greater than 6 are not very consonant, according to Holder, who explains that consonance comes from frequent alignments in the waveforms (the "Courses and Recourses of the Motion") (Holder, 1731); see Figure 4.9. The minor sixth (the integer ratio 8:5 in typical *just* tunings), is sometimes considered consonant. The minor second (semitone, about 16:15) and major second (whole-tone, 10:9 or 9:8) are definitely not consonant. Neither is the tritone, the note between the fourth and the fifth on a chromatic scale, with an interval of a half octave, near the square root of two, sometimes approximated as 7:5, 10:7, 17:12, 45:32, or 64:45.

Explaining why small integer pitch ratios lead to consonant versus dissonant note combinations has been a longstanding problem in music and auditory science. John Hershel explains, "The sense of harmony, too, depends on the periodical recurrence of coincident impulses on the ear, and affords, perhaps, the only instance of a sensation for whose pleasing impression a distinct and intelligible reason can be assigned" (Herschel, 1930). The reason may be distinct and intelligible, yet it remains unsatisfactory.

In 1973, Ernst Terhardt introduced the idea of *psychoacoustic consonance*, as distinct from the traditional musical consonance, explainable on the basis of *roughness*, the medium-speed beating of low harmonic components of two sounds that fall within a critical bandwidth of each other (Terhardt, 1974). When two harmonic sounds have pitches in a small integer ratio, their low harmonics either coincide or fall far enough from each

I ſaid, that all Concords are in Rations within the Number Six; and I may add, that all Rations within the Number Six are Concords: Of which take the following Scheme.

6 to 5 3d *Minor.*	4 to 3 4th	6 to 5 3d *Minor.*
to 4 5th	to 2 8th	5 to 4 3 d *Major.*
to 3 8th	to 1 15th	4 to 3 Fourth
to 2 12th		3 to 2 Fifth
to 1 19th	3 to 2 5th	2 to 1 Eighth
	to 1 12th	
5 to 4 3d *Major.*	2 to 1 8th	
to 3 6th *Major.*		
to 2 10th *Major.*		
to 1 17th *Major.*		

Figure 4.9: William Holder explained that "From the Premises, it will be easie to comprehend the natural Reason, why the Ear is delighted with those forenamed Concords ; and that is, because they all unite in the Motions often, and at the least at every sixth Course of Vibration, which appears from the Rations by which they are constituted, which are all contained with that Number, and all Rations contained within that Space of Six, make Concords, because the Mixture of their Motions is answerable to the Ration of them, and are made at or before every Sixth Course. First, how and why the Unisons agree so perfectly ; and then finding the Reason of an Octave, and fixing that, all the rest will follow." (Holder, 1731).

other to not cause roughness. For example, a fourth, in terms of harmonics of the *root* pitch (the greatest common divisor of the two pitches), has harmonic numbers 4, 8, 12, 16, 20 and 3, 6, 9, 12, 15, 18. Of the ones that do not coincide, the closest ratios, 9:8 and 16:15, do fall within a critical bandwidth, so some regions of the cochlea's basilar membrane will exhibit a beating interaction between those. But the beats of the 9:8 are probably too fast to contribute much to roughness. Compare that to an 8:5 minor sixth; in terms of harmonics of the root, 8, 16, 24, 32 versus 5, 10, 15, 20, 25, 30. Here the 16:15 and especially the 25:24 contribute to roughness. To be more definite, consider the fourth 400:300 Hz and the minor sixth 400:250 Hz. Their roots, and the differences between the adjacent beating harmonic numbers, are 100 and 50 Hz respectively. The fourth therefore has some 100 Hz beating relationships (for example 800:900, 1600:1500), while the minor sixth has 50 Hz beating relationships (800:750, 1200:1250). That lower beat frequency makes the minor sixth more rough than the fourth. If the minor sixth is tuned closer to an equal-tempered intonation (a power of the twelfth root of two), 1.587:1, its dissonance gets worse. Consider 400:252—instead of second harmonic 800 being near third harmonic 750, the third harmonic moves up to 756, reducing the beat frequency to 44 Hz and increasing the roughness.

According to Terhardt, roughness is the percept that separates consonant from dissonant. He says this simple approach supports Helmholtz's explanation of consonance, but doesn't totally explain the basis for small-integer ratios in music. As Terhardt points out, according to Rameau (Rameau, 1722; Rameau and Gossett, 1971) the existence of the root pitch is what makes the sounds consonant and harmonious in the musical sense: "The harmony of these consonances can only be perfect if the first sound is found below them, serving as their base and fundamental ... Thus the first sound remains the source of these consonances and

of the harmony they form." But this is not really much of an explanation, just a restatement of the condition of small integer ratios.

A literal interpretation of the idea that consonance is just lack of roughness has been successfully carried to some interesting extremes in modern reengineering of musical scales and musical tones by computer. Instead of using periodic tones, nonperiodic tones are built with inharmonic partials (*partials* or *part tones* is what we call harmonics when they're not harmonically related). In this scheme, consonant intervals are those relationships that put the inharmonic partials at frequencies where they either align closely or stay away from each other. This theory connects intervals, note spectra and timbres, and scale tunings, into systems that work for making music, without integer ratios or conventional western scales (Sethares, 2005). The idea of a *dissonance meter* for evaluating the roughness between pairs of tones was put forward both by Terhardt and by Sethares. The meter basically measures the power fluctuations in the outputs of a one-third-octave filterbank analysis of the sum of the tones, to say how dissonant they are when sounded together.

Music made with non-integer-ratio intervals, using notes with inharmonic partials, is a bit odd to the western sensibility, but is not dissonant when done right. The pitches of the notes are a bit ambiguous, since the tones are not periodic. Familiar instruments that make such tones are mostly percussion instruments. Drums, bars, bells, etc., typically have inharmonic partials, and the notion of the pitch of the *strike note* has been a much-discussed topic. When the pitch is ambiguous, one still needs to assign a value to fit these instruments into a suitable place with other more harmonic instruments.

The difference in sound between typical western harmonic music and the music of inharmonic instruments is probably attributable more to presence or absence of the root tone, as Rameau supposed, corresponding to periodic versus nonperiodic sounds, than to consonance versus dissonance, or roughness. Human mechanisms of pitch perception may be well suited for the extraction or recognition of consistent repetition intervals across separated frequency bands, as Licklider's duplex theory suggests, so there is at least a plausible theory, based on auditory images, for this difference between inharmonic and harmonic consonances.

As Mathews and Pierce (1980) concluded:

> Our experiments do not decide finally among three views of harmony: that harmony depends on a fundamental bass or periodicity pitch (Rameau), that harmony depends on the spacing of partials (Helmholtz and Plomp) or that harmony is a matter of brainwashing.

With auditory images, we can study consonant and dissonant sounds visually, and perhaps find a better characterization of consonance and dissonance in terms of auditory representations, rather than in the more abstracted terms of frequency components, periodicity, and roughness. Figure 4.10 shows how different degrees of consonance and dissonance can be visualized as degrees of regularity in auditory images.

4.9 Speech Perception

Speech is a relative latecomer, evolutionarily. The mammalian auditory system, up through cortical areas, was well developed, and widely used for intraspecies communication sounds, long before humans evolved speech and language as we know it (Kojima, 2003). Nevertheless, we do find areas in man's highly evolved neocortex that seem to be specialized for speech and language (and for music, too). The field of speech perception, speech communication, language, etc., is too big to give a credible summary of it here. Machine hearing will eventually serve as a basis for improved spoken communication with machines. But tackling speech understanding in general is a task that will take a lot of effort at higher levels, involving language modeling, perhaps via brain modeling, that is as applicable to written language as to spoken language. In this sense, speech is largely above the level that we intend to address in the present work.

The part of the study of speech perception that might be within our scope of machine hearing is known as *acoustic phonetics*: the study of the *units* of speech in terms of their physical realization as sound, how

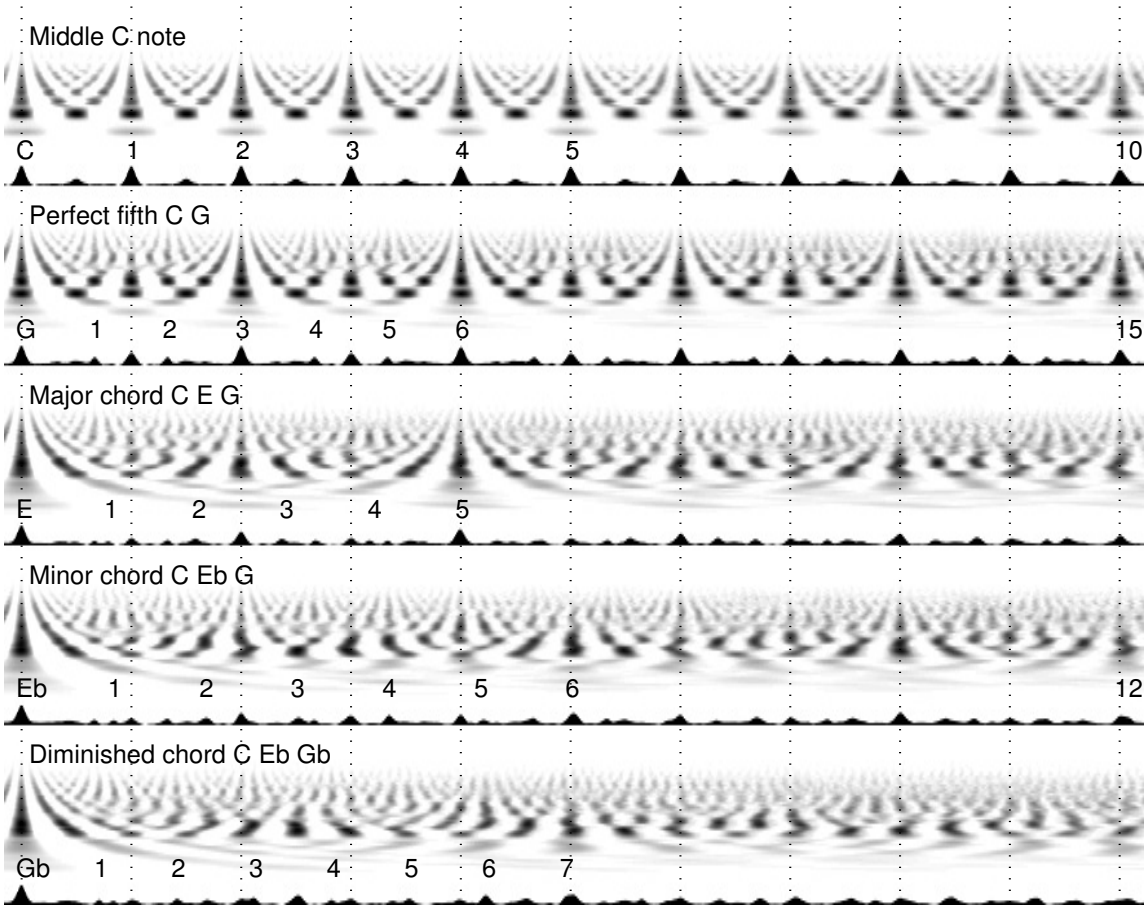


Figure 4.10: The stabilized auditory image (SAI) displays a steady musical tone or chord as a steady image. Each stripe of this figure is an SAI of a musical tone or chord, starting with the middle-C (260 Hz) note of a bassoon played alone (top stripe), followed in the subsequent stripes by four combinations of that C tone with bassoon notes of other pitches, to visualize how harmonicity, consonance, and dissonance appear in this representation. Below each stripe of SAI is a *summary SAI*, a graph of the sum of the SAI over frequency channels, peaks of which correspond to likely perceived pitch periods, including induced root pitches. Between the SAI and the summary SAI, the most recently introduced note is named and its periods along the time lag axis are labeled. The patterns get increasingly complex, from the simple structure of the C note, to the pattern with twice the period when the G is added (3 periods of G aligning with two periods of C, a perfect fifth, making a root pitch of C, down an octave), to four times the period for the major chord CEG (pitches near 4:5:6, root pitch down another octave, at four cycles of C), to a more complex pattern of partial alignments for the minor chord (pitches near 10:12:15), and finally the relative disorder of the more dissonant diminished chord (pitches near 20:24:29 or somewhat near 5:6:7). According to Holder, it is only the alignments out to about 6 cycles that make "concords."

these sounds are represented in the auditory nervous system, and how they are perceived (this latter being dangerously open-ended, of course).

Phoneticians break words down into syllables, and syllables into phonemes. Listeners can relate their perception to these units, and can be trained to transcribe phonemes according to a standard alphabet. But the details of how we hear phonemes, and of how phonemes are represented in the acoustic signal, are quite complicated.

Phonemes are typically described in terms of *articulatory features*, and how those features affect the acoustics can be very complicated and context dependent. Vowels are described by their *formant frequencies*, the center frequencies of the first several resonances of the vocal tract; but what values correspond to what vowel is complicated by language, dialect, and the speaker's vocal tract length. Consonants are often most apparent in how they modify adjacent vowels. For example, a consonant with a labial *place of articulation* (that is, associated with a closure of the lips) will typically be evident in the acoustics by a dip in the formant frequencies of an adjacent vowel; for example, the "b" sounds in "Bob" will cause the formants of the vowel to come up from a dip by the initial labial stop, and dip down again into the final labial stop. For a velar place of articulation (at the velum, or back roof of the mouth, as "g" in "get," say), the directions of the formant transitions typically push the second and third formant resonance frequencies toward each other, in what is seen on a spectrogram as a "velar pinch" (spectrograms are discussed in Chapter 5). The effect of an alveolar place of articulation (a "d" sound, say, articulated at the top front of the mouth, the alveolar ridge behind the top teeth) are not so simple to describe, being in different directions for different vowels. For consonants, the main features other than place features are the *manner of articulation* features (stop, nasal, fricative) and the *phonation* features (voiced, unvoiced). A labial (at the lips) consonant is "b" if it's a voiced stop, "p" if unvoiced stop, "m" if nasal; in English we don't have a pure labial fricative, but we do have the voiced and voiceless labio-dental (lips and teeth) fricatives "v" and "f." These features are all about what the mouth is doing to make the speech sound, but even there the correspondence is only approximate, due to *coarticulation* effects and other shortcuts in typical fluent speech.

It is not possible, in general, to prevent the influence of higher-level (top-down) information from our linguistic expectations, or even from simultaneous visual input. Two examples will serve to illustrate. First, if a portion of a speech signal is replaced by a noise burst, the speech will often still be heard correctly, and when this happens, the listener is typically not able to say which phoneme was replaced—suggesting that the word decision can determine the phoneme perception, as opposed to the other way around (Samuel, 1981). Second, if the sound conflicts with a video of a person speaking, say by having a different initial consonant on a word, the visual input can often cause a listener to "hear" a different phoneme from what the sound alone would suggest. This *McGurk effect* can combine audio and visual cues, especially about place of articulation, which is something that most people naturally pick up partly by watching the lips, to cause the listener to perceive a phoneme that was not spoken either by the person on the audio track or the one on the video track. For example, audio "ba" and video "ga" typically combine to generate the perception of "da"; the subject will believe the "da" was in the sound, though the sound would clearly be "ba" without the video (Massaro, 1998). The video is telling the subject that this can't be a labial (that is, it is not articulated by the lips), and the consonant most nearly consistent with the formant transitions, but not labial, is the "d," so that's what the subject perceives.

The perception and understanding of speech is even possible when the sound is very un-speech-like. For example, it is possible to make intelligible speech by adding three sinusoids whose frequencies follow the formants (Remez et al., 1981). It sounds like modulated sinusoids, yet the speech part of the brain is still able to understand words from this nonspeech signal. An unvoiced whisper is also very unlike normal speech, yet we communicate effectively with this noise-like inharmonic sound, presumably because we still get enough information to decode enough of the articulation features.

Normal speech is incredibly robust, though, compared to whispers and sine-wave speech. We can under-

stand it after all sorts of severe modifications, such as filtering, addition of noise, various kinds of clipping and other waveform distortions, and even adding other speech to it. The nearly periodic glottal excitation of the formant resonances gives the speech waveform a strong temporal structure, lots of redundancy and predictability, locally strong signal peaks in restricted time–frequency regions, etc. The lower parts of the auditory system help by preserving these cues in a form that the higher levels can exploit—not because the ear evolved to suit speech, but because speech evolved to suit the ear. For example, the formant bandwidths are about the same as the analysis bandwidths of the cochlea, resulting in locally high signal-to-noise ratios at cochlear places matched to the formants.

The vowel part of the speech signal is usually regarded as having a fairly steady spectrum, with harmonics of the glottal rate forming the fine spectral structure and the formants imposing a coarse, or large-scale, spectral structure. Whispered speech and sine-wave speech somewhat support the idea that a frequency-domain view is a good description. But on the auditory nerve, the temporal structure of the glottal excitation is robustly present. Consonants also make a variety of acoustic events that are most apparent in the time domain. It is hard to explain the robustness of speech in interfering speech without using the fine time structure of the signal. Whispered speech is not robust to interference from other whispered speech, and sine-wave speech is not robust to interference from other sine-wave speech, because their combined spectra don't include useful clues for how to separately interpret them (Scheffers, 1983; Assmann and Summerfield, 2004). The useful clues in normal speech are the regularly repeating glottal pulses; in the frequency domain, a narrowband analysis can resolve the harmonics of each speaker if the pitches are constant enough, but in typical speech the pitch will change too fast for this frequency-domain approach to be effective. It is more likely that we key in on repeating glottal pulses, which can be reliably localized, even in the case of multiple simultaneous talkers (Irino et al., 2006).

4.10 Binaural Hearing

Our system of two ears is descended from the bilaterally symmetric *lateral line system* of the early fishes. We and the fishes both use our hair cells to detect fluid motions, to let us know about things going on around us, by comparing signals received on opposite sides of the body. As Weiss and Buchanan (2004) explain, "The distinction between 'hearing' and the detection of other forms of environmental vibration is after all a human invention."

In terrestrial vertebrates, the mechanisms to support binaural spatialization of sound, by comparing signals from the two ears, are hardwired in the brainstem, the most primitive part of our brains; in particular, the *olivary complex* is the part of the brainstem that takes input from both ears and extracts binaural cues. We can detect the direction of signals from slightly to left or right of center based on time-of-arrival differences of a few tens of microseconds, by comparing neuron firing times in the medial superior olive (MSO). The lateral superior olive (LSO) compares the level of sounds at the two ears. Both the time differences and level differences are functions of frequency, and these brain structures maintain a frequency axis to map out the patterns. The interaction of sound waves with the head, ears, shoulders, and torso give every direction a fingerprint pattern of interaural time difference (ITD) and interaural level difference (ILD), which can be combined and interpreted in higher brain levels, based on experience.

Jeffress (1948) proposed a cross-correlation model of binaural ITD sensitivity, using neural delay lines to test a range of different alignments of signals received at the two ears, preserving a tonotopic axis as well. This early binaural auditory image model is now seen as a good model of what has been found in the neurophysiology of the MSO. Comparison of relative intensities in the LSO is not quite so simple, but seems to involve excitatory–inhibitory interactions of inputs from the two ears (Brownell et al., 1979).

In indoor situations, we encounter multiple sound reflections and reverberation that can confuse our sense of direction. Fortunately, we pay a lot of attention to the onsets of sounds, or transients in sounds, which

arrive directly from the direction of the sound source, and discount directional cues from subsequent echos from other directions. This trick of human perception is known as the *precedence effect* or the *law of the first wavefront*—or in the fields of architectural acoustics and sound reinforcement as the *Haas effect* (Muncey and Nickson, 1964). It is a very compelling effect perceptually, but has not been easy to explain or model effectively. It is sometimes described as a reduction in ITD sensitivity for a period of 0.5 to 10 ms after an onset (Zurek, 1980). Several attempts have been made to incorporate a precedence effect into a Jeffress-style cross-correlation model of the MSO (Lyon, 1983; Lindemann, 1986; Tollin, 1998), and it has been shown that the effect occurs neurally either in the MSO, or in the next station, the inferior colliculus (Yin, 1994).

In ongoing sound streams, local events such as stop consonant releases and vowel onsets in speech can act as first wavefronts, providing good directional cues and suppressing attention to their echos. As Wallach et al. (1949) observed, "the precedence effect requires sounds in which there occur some sharp discontinuities or transients. Clicks, interrupted tones, and piano music all contain the requisite transients and show the precedence effect. Orchestral music and continuous tones show the effect less well."

Besides the localization effects, there are other important and often-investigated binaural effects, such as *binaural release from masking*. A mixture of a signal (speech) with interference (noise) presented to one ear may have low intelligibility, partly attributable to masking. Simultaneous presentation of the noise alone, even if filtered or with inverted polarity or delayed, to the other ear can very significantly *unmask* the signal. Any clue that helps identify what part of the mixture is the signal and what is the interference can help the auditory system hear and understand the signal. In another example, signal plus noise are presented to both ears, but the polarity of either the signal or the noise is inverted at one ear; the amount of interference that it takes to reduce speech to 50% intelligibility is about 6 dB higher (4 times the power) when the signal is inverted at one ear, compared to the case of identical signal plus noise in both ears (Levitt and Rabiner, 1967). This is an impressive amount of release from masking and suggests that the binaural auditory system is able to use various sorts of interaural differences to allow the separation of, and focus of attention on, a chosen part of the sound mixture, even if that chosen part has been oddly and unnaturally modified.

In a more natural scenario, more interference is tolerated from a source at a different location in space than from an interference at the same location as the signal source. Even a "better ear" strategy, simply listening with whichever ear happens to have the higher signal-to-interference ratio, can provide a significant advantage.. In solving the cocktail party problem, we manage to attend and understand one speaker even in interference from many others, and even in a reverberant environment, leveraging the fact that most of the interference comes from different directions than the attended signal does (Hawley et al., 2004). Modeling and replicating this ability in machines is a big open problem in hearing.

4.11 Auditory Streaming

Al Bregman's *Auditory Scene Analysis* (ASA) (Bregman, 1990) has been an inspiration for a number of advances in *computational auditory scene analysis* (CASA) (Wang and Brown, 2006). The basic idea is that in natural acoustic scenes, the auditory system follows many kinds of cues from the sound input, and uses these cues to both separate out simultaneous sounds and connect together sequential sounds from the same source. Under the general term *grouping*, Bregman proposes separation and streaming processes that use all the available cues to try to decide which sound fragments to put together, and which to take apart, to come up with an interpretation of sound sources and auditory *streams*. The idea of *scene analysis* is borrowed from the vision community, where it has been in use since at least 1963 (Marrill et al., 1963).

Sound fragments that have similar spectra, or similar pitches, or corresponding amplitude or frequency modulations, or common onsets, for example, tend to group together, as they're likely to have been generated by a common process or source. Sound fragments that don't have much in common, such as simultaneous spoken vowels from speakers of different pitches, or sounds from different directions, will tend to be sepa-

rated, and interpreted as different sources. When cues conflict, listeners will sometimes flip between different stream interpretations; studying the relative strength of different competing cues in such situations can tell us something about what the auditory system is doing.

As a simple example, the alternation of two sounds, A and B, might be heard as one stream or source, alternating between two states (as in a two-tone siren), or as two independent streams, a stream of A sounds and a stream of B sounds, depending on how compatible the sounds are, and on how their onset and offset timings are aligned.

Bregman talks about the "components" of sound as if they are inherently distinct, and says that the default treatment of components is that they are fused, grouped, or integrated, into one sound stream unless there are good reasons to segregate them. Each resulting sound stream can be said to have a timbre associated with its components; Bregman says "timbre is a perceptual result, not a physical cause, and it may be simply a parallel result of the physical causes and not, in itself, a cause of grouping."

Plomp (2002) takes a related approach, assuming that the ear first separates, then groups, frequency components: "The ear distinguishes between frequency components originating from different sound sources ... The analyzing process 'overshoots' its task of separating the individual sounds, and then a subsequent synthesizing process is employed to 'repair' the defects, resulting in an astonishingly reliable picture of the world of sounds reaching the ear."

From a machine hearing point of view, finding discrete components without contributions from multiple sources would not be an easy place to start; and the idea that the components to be dealt with are frequencies is quite at odds with the importance of transient and temporal cues in sounds. Rather, we need to extract cues from the sound mixture, and work from the whole back toward the streams and components. The auditory image has been a popular representation for doing that, since its multidimensionality (for example, frequency, pitch, and time axes) provides a basis for supporting multiple types of cues (Weintraub, 1987; Duda et al., 1990; Cooke, 1993; Brown and Cooke, 1994; Ellis and Rosenthal, 1998; Slaney, 2005). Binaural auditory images are also used, to support separation based on spatial coherence (Lyon, 1983; Shackleton et al., 1992; Hartung and Trahiotis, 2001; Roman et al., 2003).

4.12 Nonlinearity

Many aspects of human hearing exhibit nonlinearity, in physiological mechanisms and psychophysical effects. We've discussed nonlinear loudness perception and nonlinear pitch perception, which are the most useful and positive aspects of nonlinearity. Masking and suppression are nonlinear interactions between different sounds, such that the addition or increased level of one sound can make another sound less loud, or even inaudible. Effects such as two-tone suppression and synchrony suppression are also seen on the auditory nerve. The detection of vibration by the inner hair cells is approximately a half-wave rectification nonlinearity, which manifests itself in the fact that pure tones and envelope-modulated high-frequency signals have corresponding periodicities on the auditory nerve, and hence corresponding pitches. Categorical perception of speech sounds, and the overriding of some speech sound cues by visual cues suggest strongly nonlinear processing in the brain. All of these effects involve nerves, and nobody is surprised that nerves behave nonlinearly, just as nobody is surprised that a computer made of logic gates would exhibit very nonlinear behaviors.

What is more surprising is the extent to which important nonlinear behaviors are manifest very early in hearing, in the wave mechanics of the cochlea. Also surprising is the extent to which some of these mechanical nonlinearities directly correlate with or explain psychophysical and neural nonlinearities.

We see two main types of nonlinearity in the cochlea: first, an instantaneous interaction of waveforms, a distortion, such that a pair of interacting sine tones will generate intermodulation products (also known as combination tones, distortion tones, Tartini's tones, or distortion products) with frequencies equal to the difference between integer multiples of the original frequencies; second, level-dependent and suppression

effects that correlate with the compressive loudness sensation, two-tone suppression, variation of critical bandwidth with level, etc. For brevity, we refer to these two types of nonlinearity as distortion and level dependence. In the ear, they mostly occur together, and to a large extent come from the same nonlinear mechanisms.

Distortion, particularly the $2f_1 - f_2$ cubic distortion tone (CDT), plays a role in the study of the subtleties of pitch perception, as noted above, and can produce audible musical notes below the pitches of the notes played, as Tartini observed. Moreover, in recent years the detection of this CDT propagating back out of the ear canal has been developed into a test of hearing function. A normally functioning cochlea generates a robustly detectable CDT oto-acoustic emission when the ear is stimulated by a pair of sine tones separated by about a critical bandwidth. As we age and lose the effectiveness of our cochleas' outer hair cells as active amplifiers, the CDT emission diminishes or disappears, especially in response to tones at higher frequencies; this loss of distortion is well correlated with elevated thresholds in audiograms. The CDT evoked-emission test is robust enough that it is becoming a standard way to assess infant hearing (Roush, 2001).

Level-dependent effects in cochlear mechanics were first measured by Rhode (1971) when he developed a sensitive technique for observing the small vibrations of the basilar membrane of the cochlea at low sound levels. Over time it became clear that the nonlinear effects are enormous, changing the gain of the mechanics by more than 50 dB as the sound level changes over a 100 dB range, and accounting for most of the compression in the perception of loudness as a function of sound pressure level. These gain changes cause big two-tone suppression and masking effects, and moderate bandwidth effects.

The fact that these two types of nonlinearity generally act together, and have a common origin in the outer hair cells and the system that controls them, has made them confusing and difficult to analyze. The distortion products, for example, depend on level in a way that is very puzzling in the context of what we know about electronic circuits with distortion (Goldstein, 1967), though inclusion of nonlinear damping in a traveling-wave model did close the gap on understanding it (Schroeder, 1975). Few models of the cochlea include both types of mechanisms; many models use instantaneous nonlinearities, and some use level-dependent gains. Authors who compare them will usually point out that the instantaneous nonlinearity is not enough to be accurate, but then opt to pick the other, rather than both (van der Heijden, 2005).

After years of studying the problem, we now think that including both types of nonlinearity in the sound analysis is not just a way to model the cochlea better, but a practical idea for efficient sound analysis, extracting robust representations that better correspond to human perception of a wide variety of sounds—an idea that we develop in later chapters.

4.13 A Way Forward

In this book, we focus on the duplex theory and auditory images as an intermediate representation for the perception of pitch, loudness, consonance, timbre, and binaural interactions, and for supporting higher levels of analysis.

If your interest in this book is to understand human hearing better, then the machine hearing parts of the book should be a viewed as a contending explanatory model. Indeed, it's a falsifiable model, which makes it better than many models of hearing—you can run it and try to show where it fails to account for observed facts about human hearing, as a way to refine our understanding.

On the other hand, if your interest is primarily in machine hearing, then I hope you'll be able to see how the machine models that we develop do a good job of reproducing, at least qualitatively, most of the phenomena of human hearing introduced in this chapter—and where we've come up short, maybe you'll see a way to do better.

After we can build systems that analyze sound as humans do, we can try optimizing away some of the biomimicry to see how much it matters in machine hearing applications—but we caution against premature

optimization, since we don't yet have a full appreciation of what aspects of human hearing are going to matter most to machines.

Chapter 5

Acoustic Approaches and Auditory Influence

> Machines which automatically recognize patterns from a stream of acoustic events, for example a spoken command, would have great utility in both communications and data processing. This paper reviews two applications of an elementary recognizer to the problem of actuating certain logical functions, and indicates how more ambitious recognizers might be utilized. In this regard, the automatic measurement of a talker's voice pitch and voicing dynamics appears fundamental to speech analysis, and hence to many recognition schemes. Visual inspection of spectral data taken from different speakers supports this contention.
>
> — "Artificial auditory recognition in telephony," E. E. David (1958)

5.1 Sound, Speech, and Music Modeling

The sound analysis approach most commonly used in speech and music processing systems is to extract features of sound sources based on models of their acoustics, in a "front-end" process that is somewhat tuned with respect to properties of the auditory system. The presumption is that a description of the vocal tract or musical instrument that produces a sound can be reliably extracted from the sound waveform, and that such a description is very informative about the underlying speech or music communication intent.

The extraction of representations of sound sources works better when the analysis respects some of the basic properties of human hearing, such as the varying widths of *critical bands* with respect to frequency. All of the techniques commonly used to compress, code, or recognize speech or music have evolved to incorporate at least some influence from models of hearing. This approach—acoustic analysis influenced by hearing—stands in contrast with our more explicit hearing approach: the hearing approach is to model the kind of signal representation that the ear sends to the brain, for arbitrary sound types and sound mixtures, and to defer the interpretation into sound-source features to later higher-level processes.

Although these approaches are differently motivated, they do share some conceptual basis. Understanding the nuances of the hearing approach will be easier once the acoustic approach is understood as a baseline.

The discussion in this chapter is primarily for readers who come from a speech processing background, or who already know enough about systems theory. For others, it might be best to skip it, and perhaps come back later after absorbing Part II of the book. We will not go very deeply into the mathematics of the methods discussed. There are excellent references that cover them in more detail (Gold and Morgan, 2000; Deng and O'Shaughnessy, 2003; Schroeder, 2004; Rabiner and Schafer, 2007, 2010).

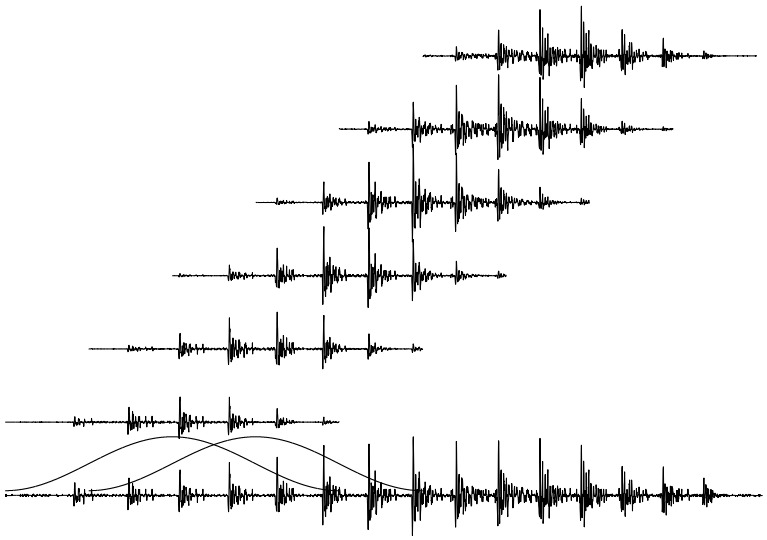


Figure 5.1: Short-time analysis is typically done on windowed segments of a sound waveform. This diagram illustrates a bell-shaped window function, placed in two different positions separated by a "hop," on a waveform of speech (of the word "I" by a low-pitch male voice). Multiplying the window, point-by-point, by the original waveform generates a *windowed segment* at each position; six of them are shown above the positions that they came from. In this example, the window is a *Hamming window* (a *raised cosine*) of 80 ms duration, and the hop size is 20 ms.

5.2 Short-Time Spectral Analysis

Most acoustic methods extract a representation of the short-time power spectrum of a sound signal, essentially in accordance with the Ohm–Helmholtz phase-blind resonance theory of hearing. A bank of bandpass filters, or resonators, or a Fourier transform of a short segment of sound, generates an output whose squared magnitude represents the power in each of many frequency bands. Such a *power spectrum* estimate changes over time, as the segment of sound analyzed is advanced in time, or as the filters respond at their own time scale to ongoing sound. The *short-time Fourier transform* (STFT) approach, popular for its ease of computation with *fast Fourier transform* (FFT) algorithms, is essentially a constrained *filterbank* approach, with all the filter bandwidths being equal and their center frequencies being equally spaced.

As Rabiner and Schafer (2007) explain:

> It can be argued that the short-time analysis principle, and particularly the short-time Fourier representation of speech, is fundamental to our thinking about the speech signal and it leads us to a wide variety of techniques for achieving our goal of moving from the sampled time waveform back along the speech chain toward the implicit message. The fact that almost perfect reconstruction can be achieved from the filter bank channel signals gives the short-time Fourier representation major credibility in the digital speech processing tool kit. This importance is strengthened by the fact that ... models for auditory processing are based on a filter bank as the first stage of processing. Much of our knowledge of perceptual effects is framed in terms of frequency analysis, and thus, the STFT representation provides a natural framework within which this knowledge can be represented and exploited to obtain efficient representations of speech and more general audio signals.

This is an elegant description of the usual acoustic approach and its motivation in speech processing. But

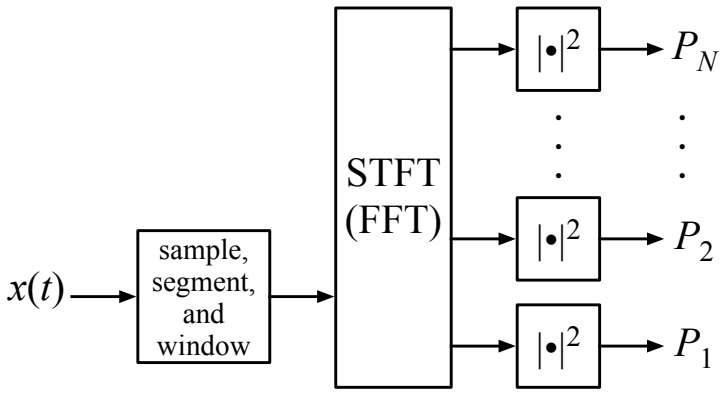


Figure 5.2: The short-time spectrum of a signal $x(t)$ can be estimated by a short-time Fourier transform (STFT, typically using the fast Fourier transform algorithm, FFT). The center frequencies are in arithmetic progression, and the bandwidths all equal, unless a further stage of channel combining is added. A segment length and window function establish the time scale of the STFT. The power outputs from such an analysis are typically next compressed through a logarithmic or power-law function.

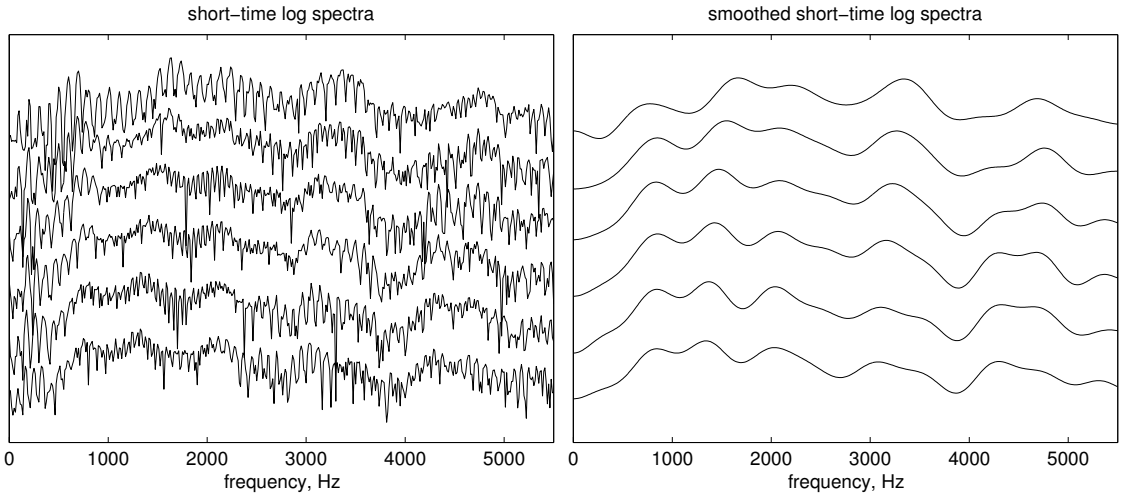


Figure 5.3: Short-time log spectra (starting from zero frequency on the left) from an FFT analysis of the 80 ms Hamming windowed segments reveal too much irrelevant detail (left). The logarithm compresses the dynamic range, but puts too much emphasis on low values (making the down-going spikes) and flattens high values. These spectra correspond, from bottom to top, with the segments shown in Figure 5.1. On the right, each spectrum is smoothed to remove "ripples," or details not very relevant to the acoustic source. Such operations in the log-spectrum domain are explained in Section 5.4.

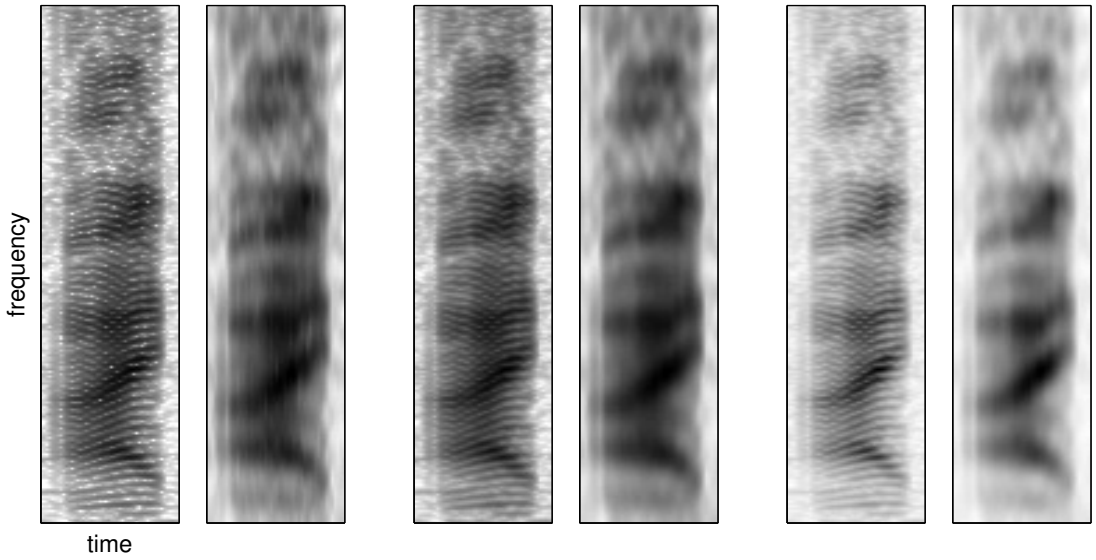


Figure 5.4: Three pairs of spectrograms and frequency-smoothed spectrograms of the word "I" are shown, using a conventional light-to-dark scale (white for silence, dark where there is energy). Each spectrogram is about 0.25 s along the time axis, and the frequency range axis is linear from 0 to 5000 Hz. In the left pair, a log spectrogram with dynamic range from white to black of 60 dB is shown. White specks where estimated power is near zero (like the downward spikes in Figure 5.3) persist as light streaks and raggedness after smoothing across the frequency dimension. In the middle pair, a very slight smoothing across frequencies is applied in the power spectrum domain before the logarithm, which eliminates most of that anomaly by keeping values away from zero. This pre-logarithm smoothing is done with a [0.1, 0.8, 0.1] filter (each spectral energy value is replaced by 80% of itself plus 10% of each of its lower- and higher-frequency neighbors). In the right pair, a power law with exponent 0.15 is used instead of the logarithm, but with the same pre-compression smoothing as in the center, resulting in more contrast in the high-power areas, and less in the low-power areas. The results with the slight pre-compression smoothing (center and right pairs) are much cleaner, with the few very-near-zero values having been fixed before allowing the nonlinearity to spike downward there. The white specks are gone, and the noise or raggedness that these specks add in the smoothed spectrum is removed. Spectrograms and smoothed spectrograms like those on the left are commonly seen in publications, and are a clear indication that the effect of the logarithm was not carefully considered or dealt with. The rightmost spectrograms are most “meaningful” in the sense that they give a better visualization of the aspects of the sound spectrum that encode the spoken word, and as a result may also be better as input to a machine learning system.

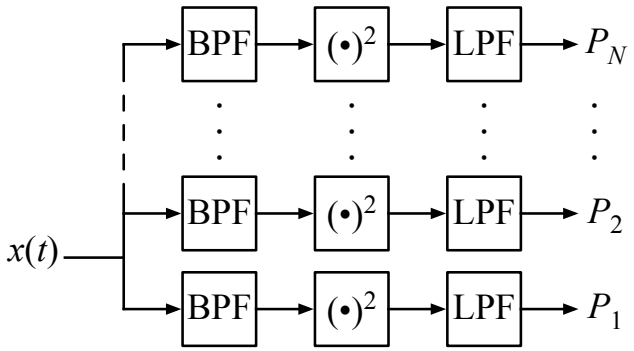


Figure 5.5: The short-time spectrum of a signal $x(t)$ can be computed by a bank of bandpass filters (BPF), followed by squaring to detect instantaneous power, and finally lowpass filters (LPF) that establish the smoothing time scale. An N-point spectrum is made via N different BPFs with different center frequencies, and often different bandwidths, and N usually identical LPFs.

if "the STFT representation" is interpreted, as it usually is, to mean *short-term power spectrum*, neglecting phase, then this explanation is less than convincing in justifying the approach by its connection to models of auditory processing. The auditory system uses not just the power, but also the fine temporal structure, in the filterbank outputs, making a much richer representation of "more general audio signals" than can be obtained via short-time power spectral analysis. The "almost perfect reconstruction" of which they speak is true only if the phase information is kept at the filterbank output—which it usually is not. Good reconstruction from short-time power spectral information, without phase, may also be possible in the case of a single clean voice signal. However, an attempt to reconstruct a mixture of two voices, or other sound mixture, with this approach will demonstrate that the reconstruction is far from adequate unless the time–frequency plane is oversampled and represented with high accuracy—essentially capturing enough fine time structure to allow the reconstruction of the sound in detail.

The short-time spectral analysis methods ignore phase information, or fine time structure, by taking the squared magnitude of Fourier coefficients, as shown in Figure 5.2. Almost equivalently, a *running* or short-time spectral representation can be computed by taking a running time average of the instantaneous square of filter outputs, as shown in Figure 5.5; here the smoothing done by the lowpass filters is more explicit; the impulse responses of these filters essentially define the analysis window.

Another common approach to computing a phase-blind representation is autocorrelation. A *short-time autocorrelation function* (STACF) can be computed as running time averages of products of a sound signal times a range of delayed versions of the sound signal; or it can be computed as the Fourier transform of a short-time power spectrum estimate from an STFT or filterbank. The relationship of the autocorrelation function to the power spectrum allows going in both directions, from STACF to short-time power spectrum estimate or the reverse, as Schroeder and Atal (1962) detailed.

These short-time analysis techniques often benefit from being tuned to properties of human hearing, such as nonlinear frequency and amplitude scales. While the Fourier transform approach naturally gives equally spaced frequency bins and equal bandwidths, the filterbank approach can be easily tailored to give a nonlinear mapping of frequencies to channels, and a different bandwidth for each channel. A popular alternative filterbank technique for sound analysis is the *constant-Q filterbank*, that is, with each channel having a bandwidth that is a constant fraction of its center frequency. Spacing the center frequencies by the bandwidth, or a fixed fraction of the bandwidth, leads to a geometric sequence of center frequencies, or a logarithmic mapping of frequency to channel number, which is a better approximation for what the ear does than a Fourier trans-

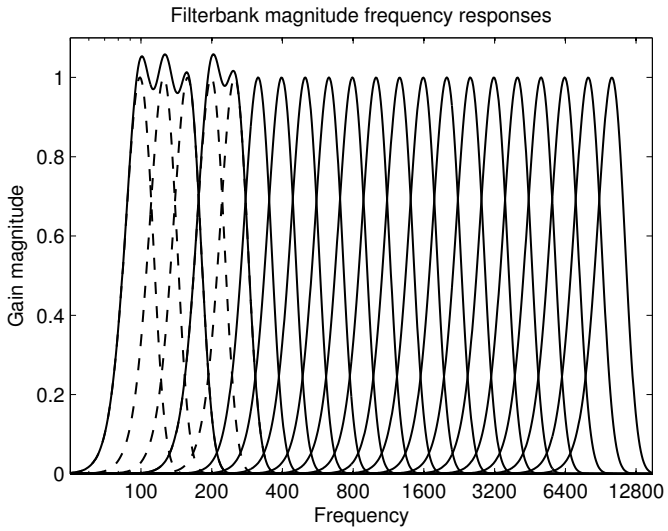


Figure 5.6: A constant-Q filterbank, here represented by 21 similar Gaussian-shaped frequency response gain curves, can give rise to an approximately auditory-scale filterbank analysis by combining some of the low channels to make an 18-dimensional spectrum vector, as Plomp et al. (1967) did; channels that are combined are shown dashed, and their sums solid. Conversely, an FFT-based analyzer, with channel responses of equal width on a linear frequency scale, can give rise to an auditory spectrum by combining increasing numbers of channels at the high-frequency end.

form is. On a log-frequency scale, the filters in a constant-Q filterbank are all alike, and equally spaced, as illustrated in Figure 5.6. Since the Q of the ear (ratio of center frequency to bandwidth) in reality decreases a lot at low frequencies, sometimes such filterbanks that have been designed for constant Q are used with modifications such as combining several channels together at the low end, as indicated in the figure.

5.3 Smoothing and Transformation of Spectra

When a short-time spectrum of a sound is computed, it may be represented by dozens to hundreds of numbers representing different frequency channels. Sometimes this amount of detail can be useful, as in narrowband analysis of music to find the partials of tones that are present. For other applications, especially in speech processing, this is more spectral detail than is useful; it is desirable to project this high-dimensional vector to a vector of lower dimension that captures most of the relevant information and omits the irrelevant detail, and in a way that eliminates or at least reduces the correlations among the dimensions.

A popular technique is *principal components analysis* (PCA), also known as the *Karhunen–Loève transform* (KLT) and by various other names (Vaseghi, 2007). It works by projecting the spectral data onto a set of basis functions that capture as much as possible of the variance of the data. The first dimension in the transformed vector is formed by projecting along the first principal component, the direction of greatest variance in a statistical model or training set that represents the distribution of spectra. The second dimension, orthogonal to the first, captures as much as possible of the remaining variance, etc.

An interesting observation about speech spectra is that a cosine transform comes fairly close to this ideal variance-capturing transform (Mermelstein, 1976) (a cosine transform is like a Fourier transform, but uses an orthogonal basis set of sinusoids in cosine phase, each having an integer number of half cycles across the domain of the spectrum). PCA basis functions tend to be in increasing order of frequency-domain detail, even if

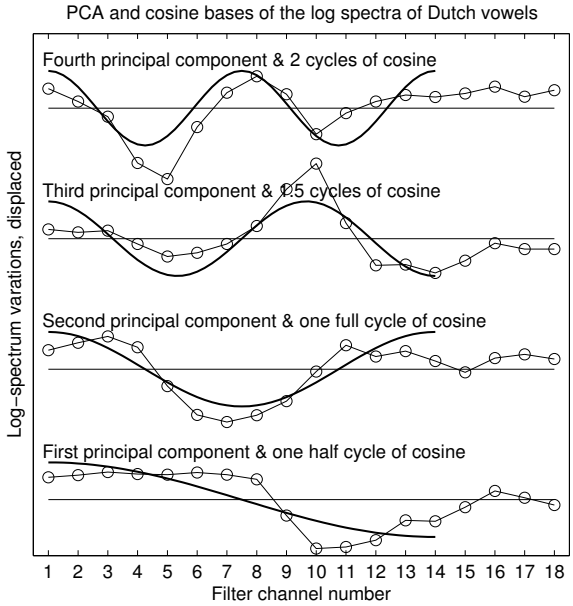


Figure 5.7: The first four principal components of speech log spectra found by Plomp, Pols, and van de Geer (1967), light curves with circles, and the cosine-transform basis functions that approximate them, heavy curves. The filter channels are approximately as shown in Figure 5.6. To get a fair fit to the cosine basis, we have ignored the last four filter channels, with center frequencies of 5 kHz and above. The analyzed speech signals were just vowels at a constant level, so there is no zero-order or constant function in the basis set.

they're not exactly cosine waves (Plomp et al., 1967), as illustrated in Figure 5.7. These PCA or cosine shapes can be used as an *orthonormal basis*: a dot product of a log spectrum with one of these components gives a coefficient that indicates how much of that component is in the spectrum shape. There is good empirical evidence that a cosine transform works well for many kinds of data (Ahmed et al., 1974); theoretical reasons to expect the cosine transform to approach the performance of PCA have also been described (Shanmugam, 1975). The cosine transform is used, for example, in the popular JPEG image compression transform, to capture most of the variance of typical image patches, using just a few numbers.

The first cosine transform basis function is the constant. For spectra, that means a constant function of frequency—a flat spectral shape. A variable amount of this function represents an overall raising or lowering of the spectrum—an intensity or loudness variation. Much, perhaps most, of the variance of typical speech log spectra is captured by this loudness dimension (unless the dataset has been pre-normalized for loudness). The second principal component resembles the next basis function of the cosine transform, a half cycle that is high at low frequencies and low at high frequencies, or vice-versa. Projecting onto this direction gives a signed quantity that measures the *spectral tilt*; in speech, this dimension captures the important difference between vowels, which tend to be heavy in low frequencies, and fricative consonants, which tend to be heavy in high frequencies. This dimension captures nearly half of the remaining variance of typical speech spectra. Higher-order coefficients fill in increasing levels of spectral detail, each capturing less variance than the ones before.

The fact that the cosine transform uses a *fixed* basis set, as opposed to PCA that derives a data-dependent basis set, makes it easier to use, to describe, and to communicate.

Based on these observations and on theoretical considerations concerning spectral smoothing, the cosine transform has long been popular in speech processing, with log spectra and root-compressed spectra, on linear frequency scales and warped frequency scales. The *cepstrum*, introduced in the next section, is an

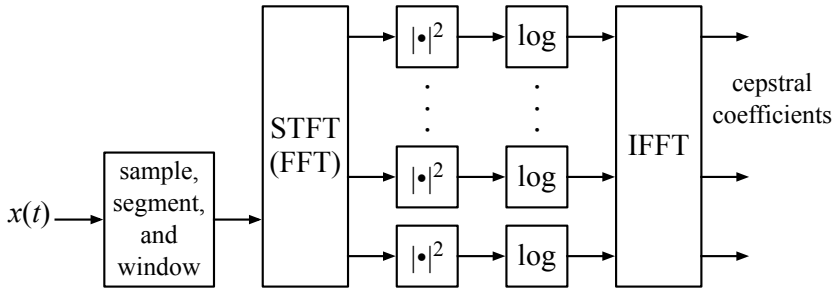


Figure 5.8: A cepstrum analyzer based on fast Fourier transform (FFT) operations. The first FFT produces complex coefficients, so a squaring or absolute value is needed. The second FFT can be a cosine transform if the spectrum is interpreted as symmetric in frequency, since a real symmetric function has no sine-phase components (no imaginary parts in its Fourier coefficients). The outputs are arranged from low to high *quefrency*, representing the smooth part and detailed part, respectively, of the log spectrum.

example of the use of cosine transforms of log spectra. Other PCA-like and related transformations are popular dimensionality reduction techniques used in machine learning systems of all sorts. However, the benefits of capturing most of the variance in a few dimensions can sometimes be outweighed by the drawbacks of too much smoothing and simplification of the signal representation.

5.4 The Source–Filter Model and Homomorphic Signal Processing

The speech signal is often described in terms of a *source–filter model*. The vocal cords (the *glottis*) act as a source of roughly periodic bursts, or impulses, which excite the *vocal tract*, in which the position of the tongue and lips act to make a filter to shape the spectrum of the sound. Since in many languages the phonetic structure of speech is created mainly through variations of vocal tract shape, a description of the filter is a useful description of the phonetic content of the speech signal. The voice pitch, the rate of impulses delivered by the glottis, carries mainly information about the speaker and about the prosodic content of the speech. Separating these two pieces of information is one goal of the source–filter model and of analysis techniques such as *homomorphic signal processing*.

The key to the model and its connection to analysis methods is that the speech signal's spectrum, or Fourier transform, can be modeled as the product of an excitation spectrum and a filter frequency response. The excitation spectrum is rippled, with peaks at every multiple of the excitation rate or pitch f_0, accounting for what's called the *spectral fine structure* of the speech signal. The filtering due to the vocal tract, on the other hand, has a frequency response, or spectral modification factor, that imposes a smoothly varying *coarse structure* on the speech spectrum. The fine structure and the coarse structure are multiplied together, point-by-point for every frequency point, to give the speech spectrum.

Converting the spectrum to a log spectrum converts multiplicative effects (such as source and filter spectral combination) to additive effects, which can make it easier to separate the fine spectrum of the source from the smooth spectrum of the filter by linear operations on the log spectrum. A mapping of functions between domains in a way that also maps an operator (such multiplication or convolution) to another operator (such as addition) is known as a *homomorphism*. The idea of using logarithms and other mappings is powerful and general enough that a whole subfield called *homomorphic signal processing* developed around the idea in the 1960s, when Oppenheim, Schafer, and Stockham (1968) showed how to formalize these ideas and apply them to image and sound processing.

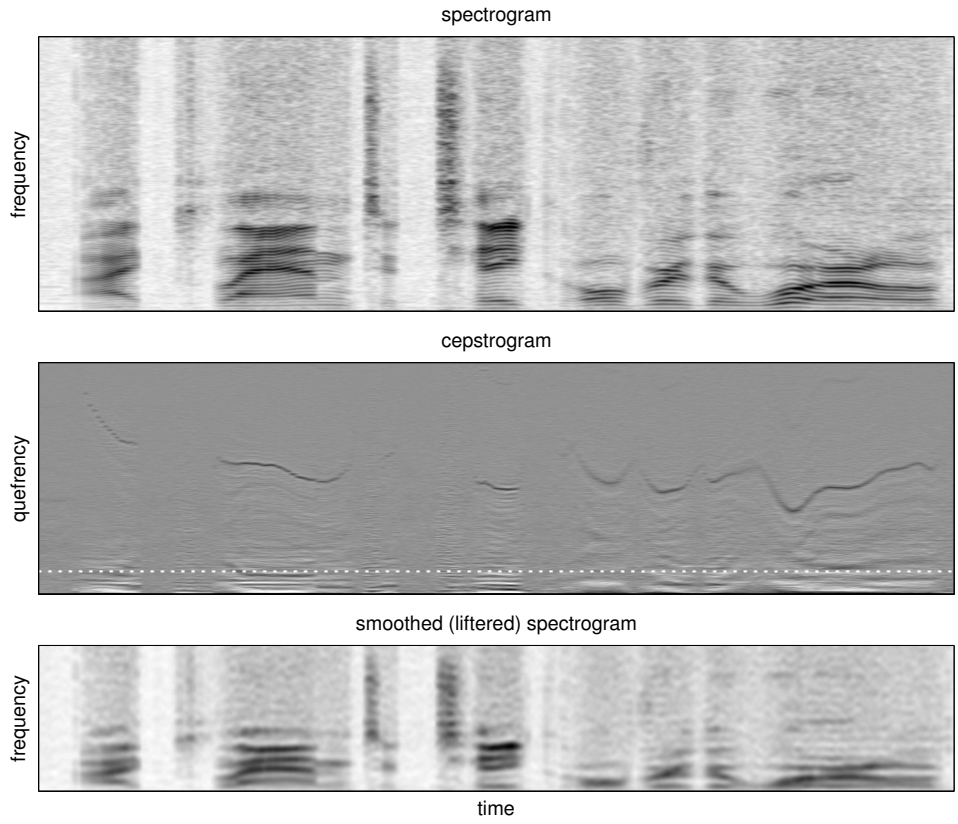


Figure 5.9: Power-law-compressed (exponent 0.15) spectrogram (top), cepstrogram (middle), and liftered spectrogram (bottom) reconstructed from the low-quefrency cepstral coefficients (from below the dotted white line). The first word in this utterance is the "I" illustrated in Figure 5.4. The full 11 kHz frequency range of the voice recording is shown in both spectrograms, at different scales, though this is a larger frequency range than is typically shown on a spectrogram with linear frequency scale. The dark curve visible in portions of the cepstrogram represents the pitch period, mirroring the movement of the pitch harmonics that are resolved in the narrowband spectrogram at top. The cepstrogram vertical axis (quefrency) can be thought of as the lag parameter of an autocorrelation, which is what it would be if the log or power-law nonlinearity were omitted and the second Fourier transform done on the power spectrum; here the lag runs from 0 to about 15 ms; cepstral coefficients are signed, with positive values darker, and negative values lighter, than the gray background. Cepstra are not usually displayed as cepstrograms, since there's not much interesting structure to look at, other than the pitch track, but they are used as representations for further sound processing.

A widely used homomorphic technique for sound processing, known as *cepstral analysis* (pronounced with a hard *c*, like the *c* in *spectral* that the name derives from), was actually developed and named a bit earlier, independent of Oppenheim's elegant mathematics, by Bogert, Healy, and Tukey (1963). They showed how convolution (filtering) in the time domain (which as we will see is equivalent to multiplication in the spectrum domain) could be converted to addition in the log-spectrum domain. They then transformed from there to the *cepstrum* domain, where a cepstrum is the Fourier transform of the logarithm of the power spectrum of the signal, to separate short-time (coarse spectral structure) and long-time (fine spectral structure) effects into *low-quefrency* and *high-quefrency* parts of the cepstrum function. Smoothing in the log-spectrum domain, preserving only low-quefrency information, is known as *liftering*, or cepstral smoothing; the concepts of *lifter*, *cepstrum*, and *quefrency* are twisted from *filter*, *spectrum*, and *frequency*.

See Figure 5.8 for a block diagram of a typical cepstral analyzer, and Figure 5.9 for an analyzed signal, where the long-time pitch-related information is particularly clear in the cepstrum. Bogert et al. (1963) called Fourier analysis of a log spectrum *quefrency alanysis* to keep it from being confused with *frequency analysis* and its inverse, which are the more common uses of a Fourier transform.

In the cepstrum, therefore, high-quefrency coefficients capture most of the information about a slow (glottal rate) repetitive excitation source that corresponds to a fine spectral comb, and low-quefrency coefficients capture most of the information about a fast filter (for example, a resonant vocal tract) that modifies those excitation pulses to shape a signal from a voice or a musical instrument. In speech processing, it is typical to want to extract a description of the filter, determined by the vocal tract shape, since that's what mostly encodes phonemic information, and to want to ignore the source, the periodic glottal excitation whose rate carries pitch and thereby signals inflection and something about the speaker but not much about the words; that is why cepstral smoothing, or truncation of the sequence of cepstral coefficients, is used.

5.5 Backing Away from Logarithms

The system theory needed to start to understand the source–filter model and the cepstrum is reviewed in part II of this book, but we will not go further into the details of cepstral analysis, nor of the other acoustic methods we survey here. The point is this: cepstral analysis aims to go directly from sound to a description of vocal tract shape, via logarithms of short-time spectra, bypassing any explicit notion of what the sound "sounds like" to the human ear.

Due to the way it separates pitch from spectral envelope, the cepstrum is also sometimes used in pitch tracking algorithms. The cepstrum can be viewed as a modified autocorrelation calculation, as it differs from a power-spectrum-based autocorrelation calculation only in the logarithmic nonlinearity. With a generalization to intermediate compressive nonlinearities—power-law curves between linear and logarithmic—the result of this process has been dubbed the *generalized autocorrelation* by Tolonen and Karjalainen (2000). For pitch extraction, power-law exponents of 0.25 to 0.67 have proved to be more robust than either linear or log (the low exponent of 0.15 in Figure 5.9 probably is too compressive, too close to the log, putting too much emphasis on the low-level noisy parts of the power spectrum). This generalized autocorrelation approach has also been adopted for robustly analyzing musical beats (Percival and Tzanetakis, 2013). The mid-range exponents make this processing more like that of the auditory system, as opposed to the processing of the more mathematically motivated linear and log extremes, as discussed in Chapter 3.

5.6 Auditory Frequency Scales

The scaling of pitch as the logarithm of frequency is a useful idea, especially for music, but is not an accurate perceptual scaling at low frequencies. Experiments on pitch comparison formed the basis for a better scaling,

the *mel scale* of pitch, named for *melody* (Stevens et al., 1937; Stevens and Volkmann, 1940) even though it is not based on musical aspects of pitch. Various tables, plots, and formulas of a modified log-like scale of perceptually scaled pitch in mels, versus frequency in hertz, evolved over the years, usually following the convention that 1000 Hz would be called 1000 mels. The most commonly used formulas are based on a logarithm with an offset that linearizes the function near zero frequency.

The offset that marks the division between a low-frequency nearly linear region and a high-frequency nearly logarithmic region is usually 700 Hz (Makhoul and Cosell, 1976; O'Shaughnessy, 1987), though other values are sometimes used. Typically, we scale the frequency by the break frequency (e.g. divide by 700), then stabilize by adding 1, so that 0 Hz maps to 0 mel:

$$\text{mels} = 1127 \log(f/700 + 1)$$

where the log base used is the natural log; a different base would work the same, but a different scale factor would be needed (instead of 1127) to map 1000 hertz to 1000 mels.

Given its circuitous history, nobody really believes the mel scale is an accurate reflection of anything about the ear or about pitch perception. The relatively high break frequency of 700 Hz is more an engineering convenience: it limits the number of low-frequency filter channels needed, and broadens their bandwidths such that they tend to not resolve pitch harmonics, making extraction of vocal-tract features more stable. It is an *acoustic frequency scale*, in the sense that it suits the acoustic model of speech production, smoothing away spectral fine structure due to pitch.

Stevens et al. (1947) characterized the mel scale as a place map: "it has been found most convenient to analyze speech by dividing it up into bands that stimulate equally wide regions on the basilar membrane. This is accomplished by choosing filter cut-offs at equal intervals along the mel scale of subjective pitch." But this is not accurate. Much more realistic *auditory frequency scale*s of the same form as the mel scale, but with lower break frequencies separating their linear and log regions, have been proposed since then. Greenwood (1961) found the mel scale lacking as a cochlear place map, and proposed a better function with a break at about 165 Hz instead of 700 Hz (though this was long before either function was written in the modern form).

Consider the general form:

$$f_{\text{scaled}} = A \log(f/f_{\text{break}} + 1)$$

where A is an arbitrary scaling constant. For $f_{\text{break}} = 228.8$ Hz, we get the ERB-rate scale of Glasberg and Moore (1990), a psychophysical scale based on *critical bands* via the *equivalent rectangular bandwidth* (ERB) of auditory filters, as estimated by tone-in-noise masking experiments. For $f_{\text{break}} = 165.4$ Hz we get Greenwood's cochlear place map (Greenwood, 1961, 1990). Either of these makes a better *auditory* frequency scale, and we'll use something in that range for our model of the auditory periphery (the cochlea). Nevertheless, a mel scale with 700 Hz break frequency is still conventionally used in *acoustic* analysis for speech and music.

5.7 Mel-Frequency Cepstrum

The ideas of cepstrum and mel frequency introduced in the previous sections are often combined into what is call a *mel-frequency cepstrum*. When this method was introduced, its results were called *mel-based cepstral parameters* (Mermelstein, 1976), but they are now more commonly called *mel-frequency cepstral coefficients*, or MFCCs. In this technique, the short-time power spectrum is *warped* to a mel frequency scale (either by merging bins of Fourier spectrum, or via a bank of filters), these power values are compressed through a logarithmic nonlinearity, and then this mel-scale log spectrum is cosine transformed to cepstral coefficients. Paul Mermelstein attributed the mel-based cepstral analysis to Bridle and Brown (1974), who described their analysis this way:

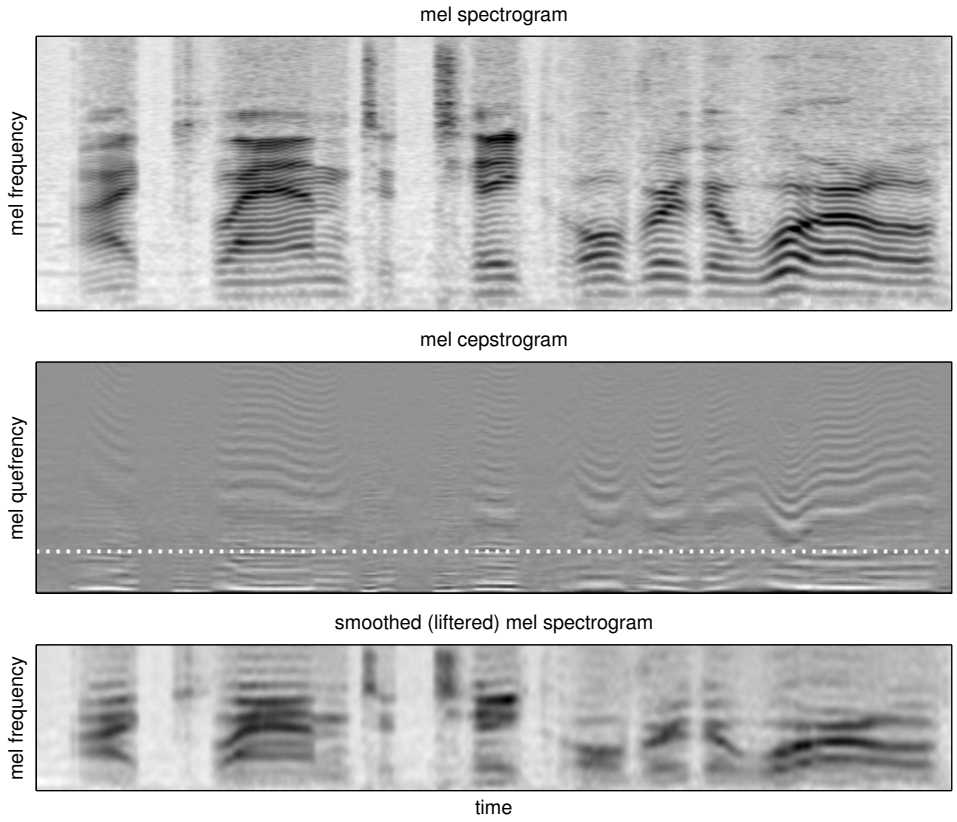


Figure 5.10: Mel-scale power-law spectrogram (top), mel-frequency cepstrogram or MFCCs (middle), and liftered mel spectrogram (bottom) reconstructed from the low-quefrency cepstral coefficients (from below the dotted white line); though this low-quefrency region looks bigger than in Figure 5.9, it is only 22 coefficients instead of the 33 used in the linear-frequency case. The mel frequency warping interferes with the usual cepstrum's simple representation of periodicity in the high-quefrency region, smearing out the pitch curve into ripples, but gives a much better distribution of spectral information.

> The initial analysis is performed by an experimental 19-channel vocoder, which has been designed for low bit-rate voice communications. The analysis can therefore be assumed to give a compact description of the speech signal, while preserving at least enough information to allow speech communication.
>
> The vocoder analyser describes the speech in terms of 'frames', at the rate of 50 frames per second. Each frame is coded in 48 bits. The result of unpacking and decoding each frame is a 19-point logarithmic, short-term, power spectrum with a nonlinear frequency scale and an amplitude of 0 to 15, representing 48 dB. Each analysis frame also contains a decision as to whether or not the speech at that time is voiced, and, if voiced, an estimate of the fundamental voicing frequency.
>
> ... There were several reasons for using this method of analysis: the vocoder was available and easily interfaced to the computer, the encoded speech data could be stored efficiently, and the associated vocoder synthesiser could be used to 'play-back' any selected portion of the data. The last feature has proved to be extremely useful.
>
> ... The 19-channel log spectrum is transformed, using a cosine transformation, into 19 'spectrum-shape' coefficients, which are similar to cepstrum coeffcients.

To a large extent, this long-evolving approach and rationale are still representative of the thinking in the field of automatic speech recognition (ASR). In particular, the ability to optimize representations based on listening to resynthesized speech led quickly to meaningful and efficient ways to analyze and represent speech spectra. Representations that are good for voice coding and resynthesis (vocoder systems) have sometimes proven to also be good in recognition systems.

What is often left unstated, however, is that representations that are good for low-bit-rate communication of speech do rather poorly for communicating speech mixed with interference, such as background noise including other speech—if you use a mobile phone, you've probably experienced the problem. Such techniques are also of only limited use in representing music. For music, speech mixtures, and sound in general, the fine structure due to multiple speakers, multiple instruments, general mixtures from different sound sources, and even reverberation needs to be dealt with if a tolerable resynthesis is desired. Music coder/decoder systems use a much higher bit rate to get a more acceptable quality, compared to the coding rates and techniques typically used in the voice coders in mobile phones or speech recognizers. Instead of just coding power spectra, they capture temporal fine structure, or phase, of waveforms. For example, the MP3 music coding standard starts with a bandpass filterbank, but it encodes the waveform in each band—not the power in each band. No system is able to encode and reproduce tolerable music via short-time power spectra.

5.8 Linear Predictive Coding

Concurrently with the popular use of MFCCs, *linear predictive coding* (LPC) has been a widely used method for speech analysis, representation, compression, and synthesis (Atal and Hanauer, 1971; Itakura, 1975; Markel and Gray, 1982). The technique has two equivalent descriptions: first, the signal is represented by the coefficients that make the best predictor for each sample of the sound waveform in terms of a weighted sum of a small number of previous samples; second, the signal is represented in terms of a frequency-domain spectral shape described by a small number of resonances (of the vocal tract, presumably, in the case of speech signals) (Johnson, 2003). The first description is what makes the method easy to realize computationally, and the second is what makes it give sensible parameterizations of speech spectra.

Like MFCCs, LPC provides a pretty good model of the speech spectrum, for both recognition and voice coding applications. Both do a good job of separating the overall spectral shape due to the vocal tract from the spectral fine structure due to voice pitch, at least when the voice pitch is not too high. Sometimes the higher

pitches of women and children cause the resonance parameters to lock onto harmonics, rather than formants (vocal tract resonances), which distorts and confuses the LPC representation of phonemic spaces.

The LPC prediction coefficients (typically 6 to 22 numbers per short-time analysis frame) are found from an equal number of points of the STACF of each frame of audio. The STACF can be found directly, or as a cosine transform of the power spectrum; hence, LPC can be viewed as another way to smooth and parameterize a power spectrum. The LPC predictor coefficients define what is known as an *autoregressive* (AR) model, or *all-pole* model, of the sound source. Such models are sensible for speech, at least for vowels, as an AR model is a plausible model of the acoustics of the vocal tract. For speech applications, a rule of thumb is to use a model order of about $4 + f_s/1000$; for example, 12 coefficients for 8 kHz sample rate, 20 for 16 kHz (Rabiner and Schafer, 2007).

5.9 PLP and RASTA

Putting the frequency domain on a nonlinear mel scale is one way that acoustic analysis and representation techniques have benefited from ties to the auditory system. There are several other ways that auditory techniques have been been applied to improve acoustic techniques. One of the strong proponents of such techniques is Hynek Hermansky, who originated both PLP (*perceptual linear prediction*) (Hermansky, 1990) and RASTA (a strained acronym for *RelAtive SpecTrAl* processing) (Hermansky and Morgan, 1994).

The idea of PLP is to modify LPC to be more "perceptual," by distorting the spectrum according to a mel scale, and cube-root compressing the power to be more like perceptual loudness, before transforming to an STACF-like representation and then to prediction coefficients. The coefficients are no longer usable directly as a prediction filter, but they still make a good smooth parameterization of the spectral shape, and admit the same kinds of comparisons (distance functions) as real LPC coefficients do. Hermansky found that for speech recognition, PLP model orders of only 5 to 8 were optimal, suggesting that the PLP method captures the relevant spectral shape information more easily than an LPC model does, as LPC models typically need to be somewhat higher order.

Whether the compressed spectrum on the auditory frequency scale is modeled via LPC-like analysis or via a cepstrum-like cosine basis decomposition, further modifications of the spectrum can make the model more powerful (at least for speech recognition). A popular class of modifications is represented by the RASTA technique, which is essentially a bank of linear filters operating on the frame-rate spectral values. The term *relative spectral* refers to the idea of making the spectral values relative to a recent history, rather than using absolute measurements. This relative measurement can be thought of as the difference between the current spectral value and a running average of recent values in the same frequency channel. We haven't covered the background needed to describe these filters mathematically, but basically the result is to suppress very slow, say below about one cycle per second, components of the spectral variations that are more likely to represent characteristics of the environment or the communication channel than of the speech signal.

If the compressed spectrum being filtered is a log spectrum, then a difference relative to a running average of recent values is equivalent to applying a time-varying gain before the log operation. This interpretation as an *automatic gain control* operation connects RASTA processing to models of cochlear function, and to psychophysical forward masking (Hermansky and Pavel, 1995). These interpretations are approximately valid even if the spectral power compression is not logarithmic.

5.10 Auditory Techniques in Automatic Speech Recognition

The idea of basing automatic speech recognition on models of the ear and auditory brain is very old. Ed David (1958) explicitly shows an "auditory nerve" in his efficient speech coder proposed for a recognition system;

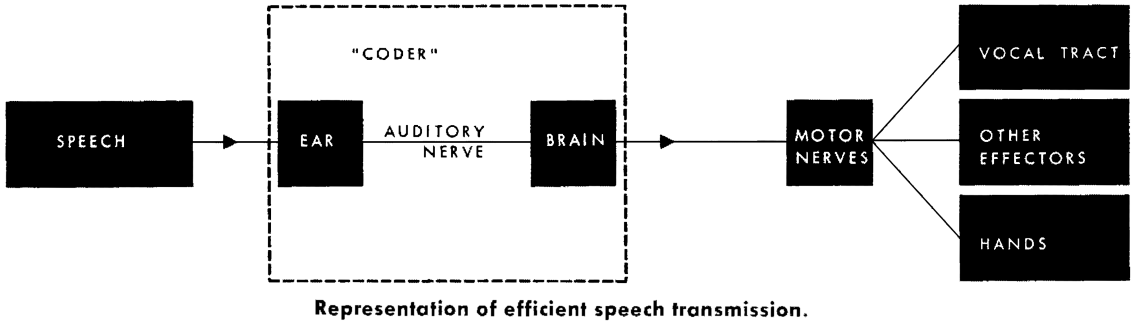


Figure 5.11: An early proposal for a speech analysis and recognition system based on mimicry of human processing (David, 1958). In agreement with our current recommendations, it respects the auditory nerve as a key representational funnel. [Figure 5 of (David, 1958) reproduced with permission of IBM.]

see Figure 5.11. This idea has recurred many times, with varying degrees of success.

Auditory influences on short-time spectral representations for speech recognition have often been reported as successful, for example in this abstract by Jordan Cohen (1989):

> Some aspects of auditory processing are incorporated in a front end for the IBM speech-recognition system. This new process includes adaptation, loudness scaling, and mel warping. Tests show that the design is an improvement over previous algorithms.

Morgan, Bourlard, and Hermansky (2004) explore the potential advantages and pitfalls of auditory approaches in automatic speech recognition systems, but in the end come back around to acknowledging the point of view that fundamentally distinguishes a speech-acoustics approach from a hearing approach:

> Clearly, listeners do not act on all of the acoustic information available. Human hearing has its limits, and due to such limits, certain sounds are perceptually less prominent than others. What might be more important for ASR is not so much what human hearing can detect, but rather what it does (and does not) focus on in the acoustic speech signal. Thus, if the goal of speech analysis in ASR is to filter out certain details from the signal, a reasonable constraint would be to either eliminate what human listeners do not hear, or at least reduce the importance of those signal properties of limited utility for speech recognition. This objective may be of greater importance in the long run (for ASR) than improving the fidelity of the auditory models.

In ASR systems, auditory techniques have been gradually incorporated as they fulfill this goal of creating better representations of speech signals. In the machine hearing approach, on the other hand, we advocate starting with a rich sound representation, and narrowing down to speech-relevant features at a later stage, in order to give the system a better chance of dealing well with nonspeech, and with speech mixed with interfering sounds.

5.11 Improvements Needed

Auditory considerations have beneficially affected many aspects of acoustic representations used in modern speech and music processing systems, including representations and processes such as mel-frequency cepstral analysis, perceptual linear prediction, and relative spectral filtering. Yet these representations are still basically optimized for how well they extract vocal tract shape, more than for how well they represent what arbitrary

signals "sound like." They are still basically spectral, with filterbank tuning, nonlinearities, smoothing, and postfiltering to tailor them to better represent those aspects of the short-time spectrum that speech recognizers most need.

Among acoustic or spectral representations, the log mel spectrum and mel cepstrum remain widely used. Though their amplitude scale is too logarithmic, and their frequency scale not logarithmic enough, they have been the standard in ASR and music work. Some groups have reported improvements by stabilizing the low end of the log energy scale (Wu and Cao, 2005), or switching to power-law energy compression (Zhao and Wang, 2013), or by switching to a more auditory-like frequency scale (Atame and Therese, 2015), or by both frequency scale and energy compression changes (Tchorz and Kollmeier, 1999).

What these auditory-influenced acoustic analysis techniques usually still omit is the pitch information, or fine temporal structure, in the sound. Such temporal information is largely independent of the smoothed short-time spectrum that describes the vocal tract, and is another cue that humans pay attention to and leverage for interpreting natural sound mixtures.

Systems using conventional acoustic representations are typically not very robust to noise or interference, nor able to represent information about sound mixtures, since the spectrum is not a rich enough starting point. The ear does more: it sends to the brain all the information needed to describe sounds. This information represents not just the energy in each spectral band, but all the fine time structure from which pitch and sound-mixture cues can be extracted. To give our hearing *machine* a chance to do what human *brains* do, we must move on to a model of the auditory periphery that is more realistic in what it represents from sound, and leave behind the idea that sound can be adequately represented by a phase-free spectral description. To get there, we first need a firm grasp of the relevant systems theory, which is what Part II of the book is for.

Part II

Systems Theory for Hearing

Part II Dedication: Charlie Molnar

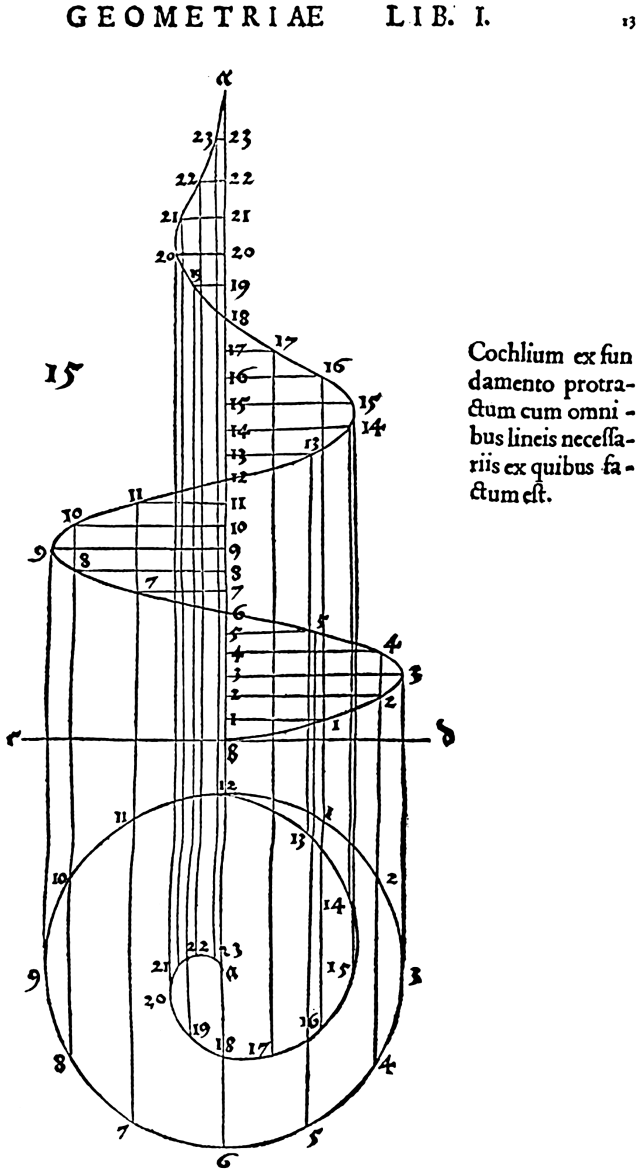
This part is dedicated to the memory of Charles E. Molnar (1935–1996). Charlie is mostly known outside the hearing field for his invention, with Wesley A. Clark, of the LINC—the "laboratory instrument computer"—which was according to many the first personal computer; and for his work in asynchronous and self-timed computer circuits, which is what he was working on when he died in 1996. He was a super generous guy, always willing to discuss and advise, and our talks about hearing and circuit design were very important to me. His "system of nonlinear differential equations modeling basilar-membrane motion" (Kim, Molnar, and Pfeiffer, 1973) was probably the first example of a great way to integrate nonlinearity into a filter-cascade model of the cochlea.

In this part, we develop the mathematical and engineering basics needed to model the ear.

We start with a review of linear systems theory, the body of knowledge that allows the design and construction of efficient and flexible filters. Even for readers very familiar with linear systems theory, a reading of this chapter should be a useful refresher and an introduction to our terminology and approach.

After a chapter on the discrete-time version of linear system theory, we apply the theory to resonant filters and elaborated resonant filters such as the gammatone family. Then we extend into nonlinear systems, with a whole chapter on automatic gain control.

Finally we discuss wave propagation in distributed systems, and how we model it with linear systems of the sort that will lead to good machine models.



Albertus Durer's 1532 *cochlium*, a "spiral extended from the base with all the necessary lines from which it is created," is unrelated to hearing, but generates a waveform that resembles the impulse response of a cochlear filter.

Chapter 6

Introduction to Linear Systems

> The fact that representation of waveforms as a sum of sine waves is useful in the elucidation of human hearing indicates that something involved in hearing is linear or nearly linear.
>
> — "The nature of musical sound," John R. Pierce (1999)

A basic understanding of linear systems is fundamental to understanding the nature of sound and hearing, including sound sources such as speech and music, sound propagation and mixing, and sound analysis in the inner ear. An understanding of linear systems is also a base on which to build an understanding of important *nonlinear* aspects of hearing.

In this chapter, we attempt to bring the novice and the expert to a common level of shared understanding about linear systems. The terminology and understanding developed here will support the material in subsequent chapters.

The concepts that we discuss include filters, circuits, impulse responses, frequency responses, convolution, transfer functions (including magnitude, phase, and delay), poles and zeros, transforms, time and frequency domains, eigenfunctions, sinusoids, and complex exponentials. The typical electrical engineering undergraduate education covers all these concepts, but too often some of the important connections between them are neglected. The typical bioengineering or hearing science education may only touch on half of the concepts, sometimes omitting Laplace and Z transforms and the important concept of poles and zeros that we need for our cochlea models. Depending on the level at which one wants to work in human or machine hearing, a deeper understanding of this area can be very helpful.

Linear systems are often encountered in the study of electrical, mechanical, acoustic, and even quantum systems, as well as in the processing of nonphysical data sequences such as stock prices. In a modern engineering or computer science curriculum, linear systems will likely be taught using discrete-time data sequences, processed by digital filters in computer programs. A more traditional approach, at least in electrical engineering (EE), is to teach linear systems using electrical circuits. This approach connects more directly with the continuous-time nature of sound waves and sound processing in the mechanical structures of the ear, so that's where we start. But we only use the simplest possible circuits, and quickly abstract them to differential equations—so if circuits are unfamiliar to you, just ignore them and go with the equations that relate inputs to outputs. In Chapter 7, we'll come back to how to map this approach into digital computers.

6.1 Smoothing: A Good Place to Start

A noisy signal can often be made cleaner, or more useful, by *smoothing*. Consider the daily closing price of a stock—a discrete-time signal. It is typical to plot a *moving average* to show price trends with less noise.

E.E. Connection: Direct and Alternating Current (DC and AC)

Electrical engineers often divide signals between *direct current* (DC) and *alternating current* (AC). The term *current* implied by the abbreviations is largely irrelevant, and it is not considered redundant or contradictory to speak of an AC current, a DC voltage, a DC response in a system that is not even electrical. DC just means steady, unchanging, or at zero frequency, while AC means cycling back and forth between positive and negative values, usually sinusoidally, with frequency as a parameter.

We use the term AC occasionally, in reference to frequency-dependent analysis. We use DC more frequently, to refer to steady-state conditions and low-frequency limits of signals and systems. Sound has no useful information at DC (at zero frequency), but the DC response of a sound-processing filter is often a practical characterization of its low-frequency behavior.

These terms were first popularized in the electric power transmission business in the 1880s, when Thomas Edison took the side of DC and George Westinghouse took the side of AC in their "Battle of the Currents" (Billington and Billington, 2013). Such contentiousness is not relevant now that things are better understood.

For example, given a sequence of daily values as input, compute for each day the average of the most recent 5 days' prices. This 5-day moving-average operation is an example of a discrete-time linear filter, or linear system. Linearity implies that if you want the moving average of the sum of two different stock prices, it doesn't matter if you add them first and then do a moving average on the sum sequence, or do a moving average on each and then add those—the answer will be the same either way. The moving average is also an example of a particular class of linear system known as smoothing filters, which means that over time, the average output is equal to the average input, but fluctuations are reduced, so the output sequence is more "smooth."

In this chapter, we focus on continuous-time systems, as opposed to the above moving-average filter, which is a discrete-time linear system. Continuous-time smoothing filters appear in many important physical systems, such as automobile suspensions made with springs and dampers (shock absorbers); the mass of your car body follows variations in the height of the road, on average, but smoothes out the bumps.

In electrical circuits, smoothing filters are easy to make; we'll start by analyzing the simplest possible one, one that can be described by a first-order linear differential equation. The condition of average output equal to average input is not one that we rely on heavily, but it is frequently a useful constraint. This constraint removes one degree of freedom, the *DC gain*—a smoothing filter has a DC (*direct current*) gain of 1. That means that if you apply a steady (DC) input voltage, say from a battery, to a smoothing circuit, then that same voltage will appear at the output, neither increased nor diminished; there may be a transient effect when the battery is first connected, but the output voltage will settle to the input voltage. We use this *unity gain at DC* property of smoothing filters in several examples, and in various places in our cochlea models.

6.2 Linear Time-Invariant Systems

A system is a device or a mathematical abstraction that maps an input function of time to an output function to time. In the simple case, the input and output are scalar functions of time, but we can use the same definition to handle systems with multidimensional (vector) functions of time, equivalent to multiple scalar inputs and outputs.

Linearity means that if we know a system's outputs for a certain set of inputs, then we can compute the outputs for weighted sums of those inputs *linearly*: when the system input is any sum of multiples of such inputs (that is, a linear combination of signals in the known set), the output is the corresponding sum of multiples of the corresponding outputs.

In general, a system maps an input signal or waveform $x(t)$ to an output signal $y(t)$. Suppose a system maps two inputs $x_1(t)$ and $x_2(t)$ to outputs $y_1(t)$ and $y_2(t)$, respectively; if the system is linear, then it will map the weighted combination $ax_1(t) + bx_2(t)$ to $ay_1(t) + by_2(t)$, for any weights a and b in a suitable domain, such as the real numbers.

A system is *time-invariant* if shifting the input in time has the effect of shifting the output the same amount in time, with no other changes—that is, if shifting $x(t)$ to $x(t-\tau)$ produces the shifted output $y(t-\tau)$, for every real time delay τ.

Systems can be arbitrarily complicated, but when they are linear and time-invariant they are much more tractable. We can completely characterize a linear time-invariant system in terms of its response to an impulse at a particular time, or by its response to sine waves of different frequencies.

In the realm of discrete-time systems, where inputs and outputs are functions of an integer index rather than a function of continuous time, the property analogous to time invariance is known as *shift invariance*. The shift must be by an integer number of discrete steps in that realm.

Linear time-invariant (LTI) systems are a core topic in many engineering and scientific fields. These two properties, linearity and time invariance (or shift invariance) are enough to imply a whole range of useful mathematics, with corresponding versions for continuous and discrete time, including impulse response convolutions, transforms to frequency domains, and multiplicative transfer functions. We outline this mathematics informally, from an engineering perspective, inviting the reader to study a more rigorous treatment in any of the excellent textbooks available on linear systems.

6.3 Filters and Frequencies

This chapter focuses on the class of LTI systems that can be described by linear ordinary differential equations with constant coefficients—as opposed to *distributed* LTI systems that require partial differential equations, *discrete-time* systems that require difference equations, *nonlinear* systems, *time-varying* systems, and other extensions that are covered in other chapters. An LTI system is often referred to as a *filter*—historically, because it lets some frequencies through and *filters out* other frequencies. The effect of filters on different frequencies is often the key to their functional importance.

In general, when we say "linear system" or "filter" we usually mean an LTI system, or a system with slow enough variation of parameters in time that we can treat it locally as time-invariant. For example, your stereo amplifier is an LTI system except while you're turning the volume control or a tone-adjustment knob, so we treat it as an LTI system parameterized by the knob settings.

A typical way to write ordinary differential equations of two variables is in terms of weighted sums of their different orders of derivatives. Here is a second-order example (highest derivative order being 2):

$$a_2 \frac{d^2y(t)}{dt^2} + a_1 \frac{dy(t)}{dt} + a_0 y(t) = b_2 \frac{d^2x(t)}{dt^2} + b_1 \frac{dx(t)}{dt} + b_0 x(t)$$

Mathematically, the differential equation expresses a relation between two functions of time, $x(t)$ and $y(t)$. The "system" or "filter" is the electrical, mechanical, or abstract mathematical entity that maps a function $x(t)$ (the input) to a function $y(t)$ (the output) in agreement with such an equation. It is common to describe what such a system does in terms of its effects on different frequencies. For example, a *lowpass* filters *passes* low frequencies from the input to the output, but filters out, or *attenuates* high frequencies.

The effect of the RC filter circuit of Figure 6.1 is like what you might hear by turning down the treble on your music player—turning the treble knob lower will do more smoothing, reducing high-frequency ("treble") content and making the sound more "muffled." As the quote at the head of the chapter suggests, the relevance of frequency—that is, the frequency of sinusoidal components—follows mathematically from the definition

E.E. Connection: Linear Electrical Circuits and Filters

Electrical circuits made of linear components, such as resistors, capacitors, inductors, transmission lines, and ideal amplifiers, are what we normally think of when we speak of electrical or electronic filters. Consider the first-order lowpass filter of Figure 6.1: a circuit with a resistor of resistance R connects an input terminal to an output terminal, and a capacitor of capacitance C connects from the output to ground (that is, to a common circuit node, with respect to which the input and output voltages are measured). Typographically, we often use roman letters such as R and C as reference designators for components, and corresponding italic variable names such as R and C for their resistance and capacitance values.

Unlike a resistor, a capacitor is a component that has *state* or *memory*; it holds electrical charge, an imbalance in the number of electrons on its two plates, and the voltage across it is proportional to the charge that it is holding. The charge stored in the capacitor is the time integral of the current into it (the current I in Figure 6.1), and the voltage is proportional to that charge, with a ratio called the *capacitance*. This time integration is what gives it state: the present voltage is a function of the history that determines the stored charge. The other kind of electrical component with simple state is the *inductor*; the current through an inductor tends to persist, proportional to the time integral of the voltage across the inductor.

The resistor–capacitor or "RC" filter of Figure 6.1 is called *first-order* because it has only one *state variable*: the voltage across the capacitor. In general, the order of a system is the number of state variables needed to specify its instantaneous state. In our RC circuit, the voltage across the capacitor, call it V_C, is the filter output signal $y(t)$:

$$y(t) = V_C = \frac{1}{C}\int I dt \quad \text{or} \quad I = C\frac{dV_C}{dt}$$

The resistor, on the other hand, is a much simpler *stateless* element, characterized by *Ohm's law*, which says that current is proportional to voltage (the voltage V_R across the resistor corresponds to $x - y$ in our circuit), at every instant of time:

$$x(t) - y(t) = V_R = IR \quad \text{or} \quad I = \frac{V_R}{R}$$

The effect of connecting these two elements as drawn is that their currents I are equated, causing the output voltage to change more slowly than the input voltage, because it takes time for the integration of the current to change the voltage across the capacitor. The mathematics describing it is worked out in the main text.

The filter passes slow (low-frequency) fluctuations pretty well, but fast (high-frequency) fluctuations less well, so it is called a *lowpass filter*. The output is *smoother* than the input, but follows very low frequencies with a gain of 1, so it is called a *smoothing filter* (lowpass filters in general might have other gains). It may be referred to by such terms as *RC filter*, *RC lowpass*, or *RC smoothing circuit*. The latter names uniquely specify the circuit: the only way to connect one R and one C to make a filter that preferentially responds to low frequencies.

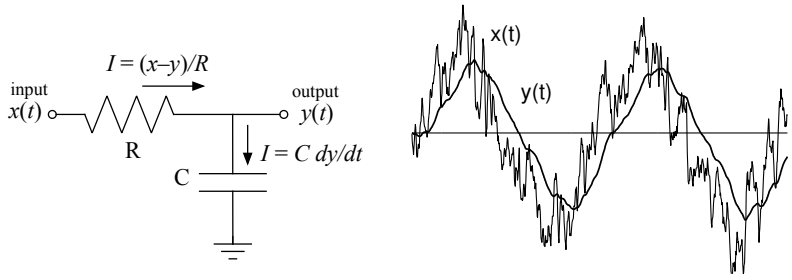


Figure 6.1: An example filter schematic diagram, and example input and output waveforms. The RC filter, with a resistor R in series with the input and capacitor C shunting the output, is the simplest smoothing filter. The differential equation that relates the output voltage $y(t)$ to the input voltage $x(t)$ is found via Kirchhoff's current law, which equates the currents $I(t)$. The graph on the right shows how a noisy input voltage signal $x(t)$ results in a relatively smooth output signal $y(t)$.

M.E. Connection: Mechanical State

Mechanical engineers will note that we have started with electrical examples. There are corresponding linear mechanical systems that have similar equations. Electrical circuits are often used as models of mechanical systems, and we use them that way when we connect them to mechanical filtering in the ear.

In mechanical systems, masses have state, because their velocity tends to persist, with momentum being proportional to the time integral of applied forces that accelerate the mass. And springs have state, since they push back with a force proportional to displacement, where displacement is the time integral of the velocity of whatever is displacing. Like inductors and capacitors, these elements store energy: masses, like inductors, store kinetic energy, while springs, like capacitors, store potential energy. Mechanical systems can be modeled by electrical ones, and vice versa, by making the appropriate analogs, such as current for velocity and voltage for force.

If you have come to hearing via mechanical engineering, acoustics, or applied physics, I trust you will be able to make the appropriate connections.

E.E. Connection: Lumped and Distributed Circuits

Both resistors and capacitors, as well as (idealized) inductors and transformers and amplifiers, are what are known as *lumped* elements, since they lump the effects of physical structures into simple device models that are described by just the voltages and currents at their terminals.

Another important class of linear element is the *distributed* element, most notably the transmission line. Any piece of wire long enough to introduce an appreciable delay, such that points along the wire cannot be treated as all having the same voltage, must generally be treated as a distributed element. The analysis of distributed systems, or transmission lines, was developed to characterize and improve telegraph lines in the nineteenth century (Heaviside, 1892). The treatment of distributed elements has a lot in common with the treatment of lumped elements, but the math is a bit different. In particular, distributed systems involve partial differential equations, to describe functions of both time and space, while systems of lumped elements get by with ordinary differential equations, describing functions of time only. We'll treat distributed systems in Chapter 12 and later chapters, as we get closer to the mathematics of waves in the cochlea.

A continuous-time moving-average filter is perhaps the simplest example of an LTI system that cannot be represented as a lumped system. An exponentially-weighted moving average, on the other hand, is what the RC lowpass filter example of this chapter computes, using the state of a single lumped element, a capacitor.

of an LTI system. But we before we get to that, it is best to understand more about how the filter works in the time domain.

6.4 Differential Equations and Homogeneous Solutions

For the RC circuit of Figure 6.1, the differential equation that describes the filter is found by equating the current that charges the capacitor to the current that flows through the resistor, driven by the difference between input and output:

$$\frac{x-y}{R} = C\frac{dy}{dt}$$

At this point, we can forget about resistors, capacitors, currents, voltages, and circuits, and just work with the simple abstracted equations of linear systems. To help us forget, we introduce the parameter τ to stand in for the product RC; since R and C are necessarily positive, so is τ. It has units of *seconds* (in the case of the RC circuit, it is a product of volts per ampere and ampere-seconds per volt) and is known as a *time constant*, or the *RC time constant*. The differential equation relating the input and output is then:

$$x - y = \tau\frac{dy}{dt}$$

The homogeneous solution, or what the output $y(t)$ can do while the input $x(t)$ is just sitting at zero, is a decaying exponential function of time, as is easy to check by differentiating:

$$y = A\exp\left(\frac{-t}{\tau}\right)$$

$$\frac{dy}{dt} = \frac{-1}{\tau}A\exp\left(\frac{-t}{\tau}\right) = \frac{-y}{\tau}$$

where the factor A is any real or complex constant. Of course, if we're talking about a real circuit, we need real-valued solutions, like those plotted in Figure 6.2, but the differential equation is more flexible than that. Its complex-valued homogeneous solutions will turn out to be useful.

The same differential equation with $\tau < 0$ would have the same homogeneous solutions, but in that case the solutions would be exponentially growing in time, representing an unstable system, which is not a case that we are usually interested in.

6.5 Impulse Responses

The homogeneous solutions help us find the system's *impulse response*: what the output does if the circuit starts in a zero state and we send a *unit impulse* into the input at time 0. A unit impulse is a very high value for a very short time, with an integral of one (one volt-second in the case of a voltage impulse, or whatever units are appropriate to the system being modeled).

Think of a unit impulse as a square pulse of duration D and height $1/D$, starting at $t = 0$, for some D much shorter than any time resolution that we care about. The limit for small D does not exist as a function, but mathematicians have notions of generalized functions, distributions, and measures, that can handle that problem, such that we can think of the unit impulse (delta function or Dirac distribution as they sometimes call it) as a well-defined input to a system, without needing a value of D (Redheffer, 1991).

The impulse response of a continuous-time LTI system is then defined as the system's output when the input is the unit impulse. For some systems, the output will include an impulse resembling the input (as the

system $y(t) = x(t)$ does). For others, like our RC smoothing filter example, the impulse response will be bounded, because it integrates the impulse.

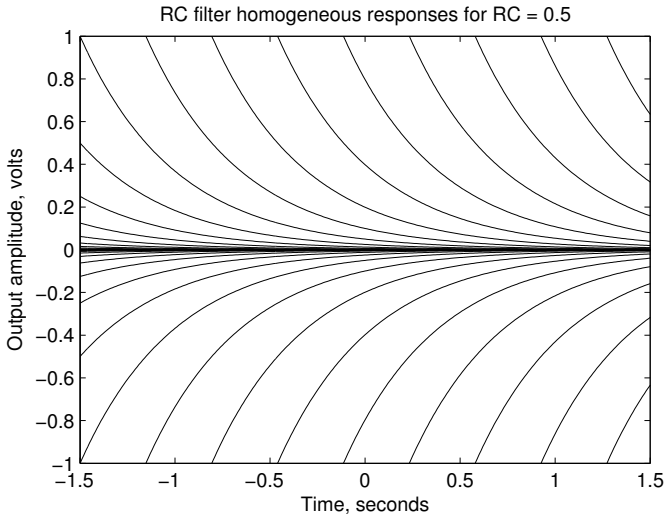


Figure 6.2: The homogeneous responses of the RC circuit are the signals that can appear at the output when the input is zero. The first-order system response is an exponential decay toward zero, from any starting value specified at any time value. For this example, the decay time constant is $\tau = 0.5$ s.

At $t < 0$, before the impulse, the input and output can be assumed to have been resting at zero forever. After the impulse, at $t > 0$, the input is again at zero, so the response will be a homogeneous solution during that time. The general homogeneous solution may have one or more free coefficients (depending on the order of the equation; only A in the first-order case); yet the impulse response is uniquely determined as the response to the unit impulse at time $t = 0$, so these free coefficients need to be tied down, either mathematically from the equations, or by reasoning about the physics of the system being modeled.

In the case of the RC lowpass filter, the coefficient A can be determined such that the integral of the output is equal to the integral of the input (one volt-second, as we defined for the unit impulse)—a special condition for smoothing filters, for which the average output is equal to the average input. The impulse response that fulfills this condition is:

$$\begin{aligned} h(t) &= \frac{1}{\tau}\exp\left(\frac{-t}{\tau}\right) \quad \text{for} \quad t > 0 \\ &= 0 \qquad\qquad\qquad\quad \text{for} \quad t < 0 \end{aligned}$$

where we introduce the new notation $h(t)$, the conventional term for an impulse response. The impulse response of the RC lowpass smoothing filter is shown in Figure 6.3.

The impulse response turns out to be a complete characterization of an LTI system, as it will allow the computation of the output given any arbitrary input, by linearity. Using a succession of pulses of duration D, back-to-back, we can approximate as closely as required any input waveform that doesn't have important structure at arbitrarily short time scales, by choosing D small enough. This decomposition into finite pulses is one way to motivate impulses as a basis set for waveforms.

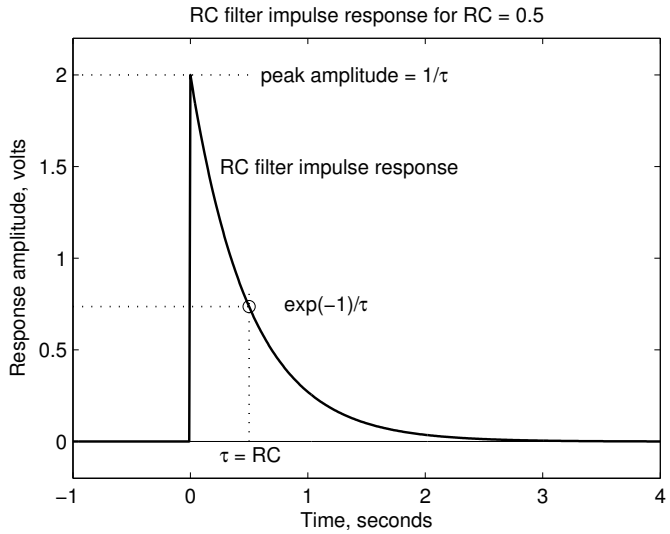


Figure 6.3: The impulse response of the example first-order RC smoothing filter of Figure 6.1 (solid curve), with time constant $\tau = 0.5$ s.

6.6 Causality and Stability

Real-world systems generally have the property that their outputs don't respond until an input arrives. This property is called *causality*, and such systems are *causal* systems: the output has to be *caused* by an input. The alternative is a *noncausal* system, which responds before its input arrives.

Causality is reflected in the impulse response $h(t)$. For a causal system, the output will be identically zero for all times before the impulse arrives at $t = 0$. Therefore, $h(t) = 0$ for all $t < 0$. We often leave this case implicit when giving a formula for an impulse response; the formula should be replaced by zero for negative t.

We generally assume that any system we discuss, whether real or simulated, is causal. Then our machine implementations at least have the potential of running in real time, if implemented on low-latency systems. For real-time applications, such as getting a machine to join in a musical jam session, causality is an inevitable property of the overall system: the response of the system can't depend on inputs that haven't arrived yet. For other applications, such as indexing stored sound files, causality is less relevant, since examining the "future" of a stored sound waveform is trivial.

The linear systems that we work with are usually *stable*, too, meaning that the output signal is bounded whenever the input is bounded. The alternative is *unstable*. For example, if the τ in the first-order exponential filter is negative, the output will grow exponentially without bound after an impulse gets it started.

Stability is assured for systems of passive components; for example, in our RC lowpass filter. In the analysis of that filter, we can be sure that $\tau > 0$, since both R and C are necessarily positive, so the homogeneous response decays with time, rather than growing. But when active amplifiers are involved, or when arbitrary linear differential equations are allowed, it's easy to make systems whose output amplitudes grow exponentially with time, even when the input is zero. Such behavior inevitably drives any real-world system outside the range over which it can be realistically treated as linear. For example, instability in the cochlea can make your ears ring, in the condition known as *objective tinnitus*; but nonlinearities prevent these oscillations from increasing exponentially for long.

6.7 Convolution

If the input to an LTI system consists of multiple impulses, shifted to different times and scaled to different sizes, then, by linearity and time invariance, the output is simply a sum of correspondingly shifted and scaled copies of the impulse response. If the input is not impulses, but rather an arbitrary waveform, then the corresponding limiting operation is called a *convolution integral*. This ability to decompose any input into a continuum of impulses, and to compute the output by integrating the responses to all of them, is the crucial consequence of linearity and time invariance, supporting a very simple complete system description in terms of impulse responses; such a simple description does not exist for nonlinear or time-varying systems.

The convolution integral that computes the output from the input and the impulse response is:

$$y(t) = \int_{-\infty}^{\infty} x(u)h(t-u)du$$

which we write using "*" as the convolution operator, as:

$$y(t) = x(t) * h(t)$$

We say x is *convolved with* h, or *filtered with* h, to get y. The $h(t-u)$ term represents the displaced impulse responses being summed, and the $x(u)$ represents the scale factors from the input. The limits of integration are generally from $-\infty$ to ∞, but for systems that are causal, it's just as good to put the upper limit of integration at t, since a causal $h(u-t)$ is identically zero for negative arguments:

$$y(t) = \int_{-\infty}^{t} x(u)h(t-u)du$$

The convolution can be thought of as "turn one signal around, slide it along the other, and integrate their product." Convolution is commutative, so we can swap the input signal with the impulse response:

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t-u)du$$

or for causal systems:

$$y(t) = \int_{0}^{\infty} h(u)x(t-u)du$$

The convolution integral makes it clear that the impulse response completely characterizes a linear system, in that it is all that's needed to compute the system's response to arbitrary inputs. For systems more complicated than our first-order filter with exponential impulse response, however, the expressions for the impulse responses may be hard to find, or may not provide much insight into the system behaviors. Additionally, the input signal is not generally known in closed form, or if it is, we may still not be able to do the integral. Therefore, convolution integrals are not typically tractable for the general case; other system descriptions, involving frequency, are often more useful.

6.8 Eigenfunctions and Transfer Functions

What makes *frequency* a concept relevant to linear systems? As it turns out, there are certain waveforms, namely sinusoids, that have a rather special property: if you apply one as input to a linear system, you get the

Math Connection: Complex Exponentials as Eigenfunctions of LTI Systems

Consider the generalization of real sinusoidal inputs to decaying and growing complex sinusoids of the form

$$x(t) = A_x \exp(st)$$

for some frequency-like parameter s, not necessarily pure imaginary like $i\omega$. With this complex exponential as input, manipulation of the convolution integral shows that the output will be a complex exponential with the same parameter s:

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} A_x\, h(u) \exp(s(t-u))\, du \\ &= A_x \exp(st) \int_{-\infty}^{\infty} h(u) \exp(-su)\, du \\ &= A_y \exp(st) \end{aligned}$$

That is, the output is like the input but with a different complex amplitude factor A_y, proportional to A_x in a way that depends on s (for values of s where the integral converges). Functions with this property—that the output function is like the input function times a constant factor—are known as *eigenfunctions*. The ratio of output to input, as a function of s, is called the *transfer function* $H(s)$, and the integral that determines it is known as the *Laplace transform* of the impulse response $h(t)$:

$$H(s) = \frac{A_y}{A_x} = \int_{-\infty}^{\infty} h(u) \exp(-su)\, du = \mathcal{L}\{h(t)\}$$

For the example RC filter, as we'll see in Section 6.10, the transfer function is:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{A_y}{A_x} = \frac{1}{\tau s + 1}$$

for values of s where the integral converges, which are $\mathrm{Re}[s] > -1/\tau$.

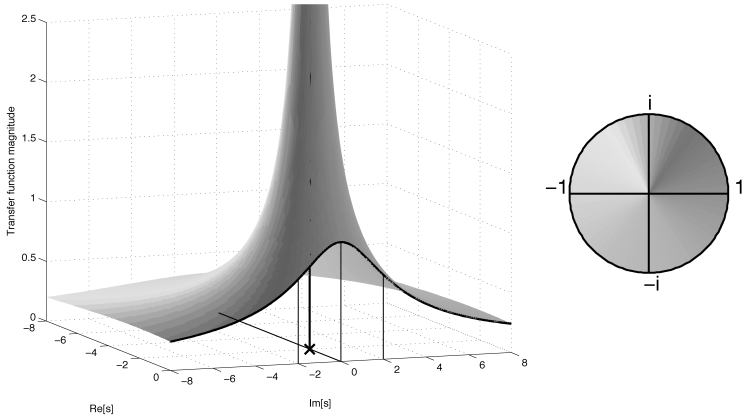


Figure 6.4: The transfer function of the example RC filter of Figure 6.1. The magnitude of $H(s)$ is plotted as a surface height above the complex s plane, while the phase of $H(s)$ determines the hue of the surface color (following the phase–hue legend on the right; see color plates). For the example filter with $\tau = 0.5$, the transfer function has a singularity at $s = -2 + i0$ (at cross and heavy vertical line). The surface is cut along the imaginary s axis to reveal the frequency response. The frequencies $\omega = 0$ (DC) and $\omega = \pm 2$ (the 3-dB points) are marked by lines.

same as output, but modified in phase and amplitude. Consider the input

$$x(t) = A_x \cos(\omega t - \phi_x)$$

where A_x is the input amplitude, ϕ_x the input phase, and ω is the frequency (in radians per second, equal to $2\pi f$ where f is frequency in cycles per second, or hertz). We use the cosine function, rather than the sine function in this example, because it's conventional to call the cosine the sinusoid at 0 phase (or because the cosine is the real part of the complex exponential of 0 phase, $\exp(i\omega t)$). We will find that the output will be of the same form:

$$y(t) = A_y \cos(\omega t - \phi_y)$$

where the ratio A_x/A_y and the difference $\phi_y - \phi_x$ (for a particular linear system) depend only on the frequency ω. This ratio, as a function of frequency, is known as the *magnitude frequency response* (or sometimes *amplitude frequency response*) of the system; the phase difference is known as the *phase frequency response*.

Uniting these response properties into a cleaner mathematical formalism, however, requires one more leap: the introduction of complex numbers. As detailed in the Math Connection box, the generalization from the real sinusoid of frequency ω to a complex exponential of complex frequency s is what makes the math work out nicely, yielding a *transfer function* $H(s)$ that multiplies the input to give the output:

$$x(t) = A_x \exp(st) \implies y(t) = A_y \exp(st) \quad \text{where} \quad A_y = A_x H(s)$$

The transfer function is a complex *gain* factor, which can be interpreted as an amplitude gain, the absolute value $|H(s)|$, and a phase shift, the angle $\angle H(s)$. For this formulation to be useful, we have to allow the gain $H(s)$, as well as s, A_x, A_y, and the system inputs and outputs themselves, to be complex-valued. This complex transfer function is a complete characterization of the linear system, equivalent to the impulse response, and computable from the impulse response. The transfer function of the example smoothing filter is illustrated in Figure 6.4; see Section 6.10 for its derivation.

Complex exponentials, that is, all functions of the form just given—and no other functions—are the

eigenfunctions of time-invariant linear systems. That means that if you put one in (for some s), you get the same one out, only multiplied by a (possibly complex) constant gain factor $H(s)$. No other signal shape will go through a general LTI filter unchanged. This fact tells us what's so special about complex exponentials and sinusoids, with respect to linear systems, and what's so special about linear systems, with respect to sinusoids or the concept of frequency.

If we really can't tolerate the idea of putting complex-valued inputs into a real-world system, that can be accommodated by always taking pairs of such signals, in a complex-conjugate relationship, such that they add up to real signals (the complex conjugate of a complex number or signal x is denoted x^*; the conjugation operation negates the imaginary part, so $x + x^*$ will always be real-valued). For example, to put the sinusoid $\cos(\omega t)$ into a system, consider two complex exponentials that add up to it, using $s = i\omega$ and $s = -i\omega$, with amplitude 0.5 for each term:

$$x(t) = 0.5\exp(i\omega t) + 0.5\exp(-i\omega t) = \cos(\omega t)$$

Real sinusoids of any phase can be made by using a pair of complex amplitudes in a complex-conjugate relationship, for example $-0.5i$ and $0.5i$ to make $\sin(\omega t)$. More generally, $A\exp(st) + A^*\exp(s^*t)$ is a real waveform made from a complex-conjugate pair of eigenfunctions.

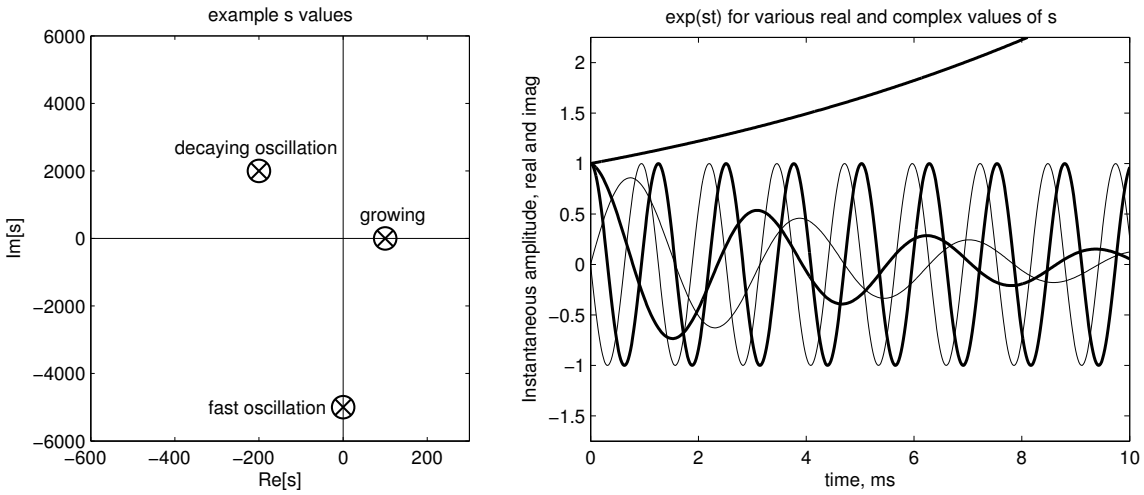


Figure 6.5: Plots of example s values (left) and the real and imaginary parts of $\exp(st)$ (right); for complex s, the imaginary part is shown by light curves, and the real parts are always heavy curves. One real s value, $s = 100$, yields the slow exponentially growing curve on top. A pure imaginary value $s = -i5000$ corresponds to the steady high-frequency sinusoids shown; notice that the imaginary part *leads* the real part, since this one is a *negative frequency* complex sinusoid. The damped sinusoid shown corresponds to $s = -200 + i2000$. In each case, the units of s are inverse seconds, or radians per second. If the input to a linear time-invariant system, or filter, is any of these signal shapes, then the output will be the same, just multiplied by the complex factor $H(s)$ for that filter at that s value. That is, these functions $\exp(st)$ are eigenfunctions of all LTI systems (subject to some region-of-convergence restrictions that we mostly ignore).

When the "frequency" s is complex, not just the pure-imaginary value $i\omega$, it is sometimes written in terms of its real and imaginary parts as $\sigma + i\omega$. This complex frequency represents an exponentially decaying or growing *complex exponential* waveform:

$$x(t) = A_x\exp(st) = A_x\exp(\sigma t + i\omega t) = A_x\exp(\sigma t)\exp(i\omega t)$$

in which the growth (for positive σ) or decay (for negative σ) with time is represented by the first exponential factor, with real argument, and the complex sinusoid $\cos(\omega t) + i\sin(\omega t)$ by the second, with imaginary argument. A few complex exponentials are illustrated in Figure 6.5.

The transfer function $H(s)$ is an *operator* that can be used like an algebraic factor, multiplying an input to give an output. Since $H(s)$ simply acts as a multiplier on frequency components, or eigenfunctions, the frequency components that come out will never include any that did not go in—but this simplicity will vanish when we consider nonlinear systems, and it is demonstrably not true in the ear.

6.9 Frequency Response

Often, we are satisfied with describing a system by its response to sinusoids or steady complex exponentials, rather than to a broader set of growing and decaying complex exponentials. Theoretically, these steady real and complex sinusoids should be an adequate set of signals to characterize a system, as they are a *complete basis* for the space of time-domain input and output signals. The general complex exponentials, by contrast, make up an *overcomplete basis*. When we limit the description to sinusoids, we call the description a *frequency response*.

We tend to use the same symbol H for the complex gain, whether it is for complex sinusoids parameterized by real frequency f or ω, or for complex exponentials parameterized by complex frequency s. To keep these related functions distinct, we sometimes use a subscript to clarify which function we mean. We thus define the frequency response in terms of the more general transfer function $H(s)$:

$$H_\omega(\omega) = H(i\omega)$$

$$H_f(f) = H(i2\pi f)$$

If we just write $H(\omega)$ or $H(f)$ we usually mean the appropriate one of the functions above (that is, $H(s)$ evaluated at the appropriate imaginary argument value, not at ω or f).

For the example RC filter, the frequency response, from the transfer function $1/(\tau s + 1)$, is:

$$H(\omega) = \frac{1}{i\tau\omega + 1}$$

We can decompose such complex gains back to amplitude gains and phase shifts, as functions of positive and negative frequency, as plotted in Figure 6.6. The magnitude frequency response is:

$$|H| = \frac{1}{\sqrt{\tau^2\omega^2 + 1}}$$

and phase response:

$$\angle H = -\arctan(\tau\omega)$$

Real systems (meaning either actual physical systems with real-valued inputs and outputs, or abstract systems for which real-valued inputs always give real-valued outputs) have a complex-conjugate symmetry in the values of the complex gain functions (as the above example has):

$$H(-\omega) = H^*(\omega)$$

Thus when the input to a filter is $\exp(i\omega t)$, the output will be $|H|\exp(i(\omega t + \angle H))$; and when the input is $\exp(-i\omega t)$, the output will be $|H|\exp(-i(\omega t + \angle H))$; so when the input is the real sinusoid $\cos(\omega t)$, the output will be the real sinusoid $|H|\cos(\omega t + \angle H)$. That is, the symmetry in H makes the imaginary output parts cancel

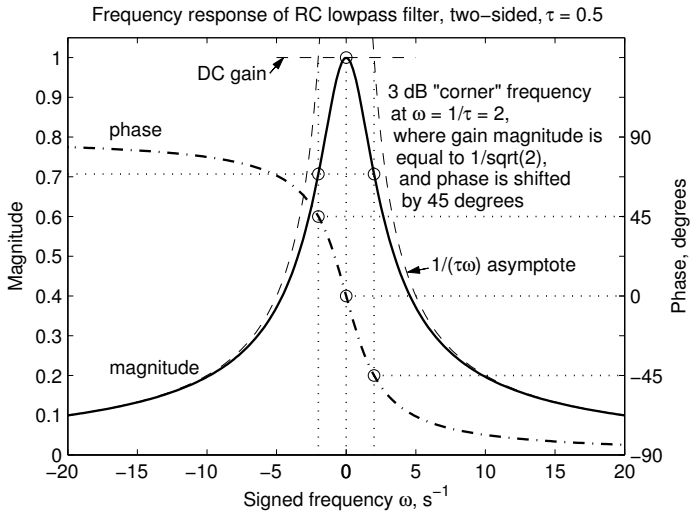


Figure 6.6: The magnitude frequency response (solid) and phase frequency response (dash-dot) of the RC lowpass filter with time constant of 0.5. The phase varies from zero at DC to a lag of a quarter cycle (90 degrees or $\pi/2$ radians) at high frequencies (and the negative of that at negative frequencies). The frequency parameter ω is in radians per second. The 3 dB *corner frequency*, $\omega_C = 1/\tau = 2$, is where the gain is 0.707 and the phase is ±45 degrees (marked with circles). Also shown (dashed) are the magnitude-gain high-frequency asymptotes, the hyperbolas $|1/(\tau\omega)|$, and the low-frequency (DC) limit of gain 1. These power-law asymptotes are shown for comparison with their straight-line versions in a log–log plot in the next figure.

when the input is real.

Frequency responses are often plotted as *Bode plots*: plots of log-magnitude gain and phase versus log frequency (for real systems and positive frequencies only). A Bode plot for the RC lowpass is shown in Figure 6.7. The complex logarithm function separates magnitude and phase out from a complex gain: the real part of the complex log is the log of the magnitude, and the imaginary part of the complex log is the phase; so both parts of the Bode plot are log–log plots. In practice, often the magnitude frequency response of a filter is plotted, and the phase response ignored. In hearing, the magnitude response plays a more important role, but the phase response also needs our attention sometimes.

6.10 Transforms and Operational Methods

Now that we have two alternative characterizations of linear systems, namely impulse responses and transfer functions, we should examine how they are related. Both are functions (or generalized functions, when impulses are included); the mapping between them is known as an *integral transform*.

The transform most familiar to scientists and engineers is the Fourier transform. In 1822, Joseph Fourier came up with the idea of solving complicated linear systems by analyzing their responses to sinusoidal components, and decomposing all inputs and outputs into those. This transform is not general enough to give us the full $H(s)$ from $h(t)$; but it will give us the frequency response, $H_\omega(\omega)$, that is, $H(s)$ evaluated only where $\sigma = 0$ in $s = \sigma + i\omega$. The Fourier transform from impulse response to frequency response is:

$$H_\omega(\omega) = H(i\omega) = \mathcal{F}\{h(t)\} = \int_{-\infty}^{\infty} \exp(-i\omega t)h(t)dt$$

The Laplace transform includes the Fourier transform as a special case where $\mathrm{Re}[s] = 0$, but is a more

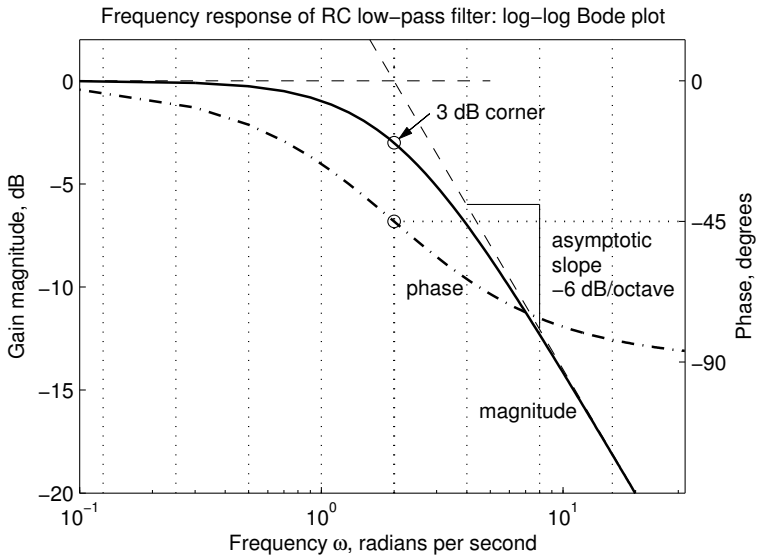


Figure 6.7: Bode plot, or log–log frequency response, of the example RC lowpass filter. Magnitude-gain asymptotes are shown dashed, and octaves are marked by dotted lines. The high-frequency asymptote has a slope of –6 dB per octave (or –20 dB per decade, that is, per factor of 10 in frequency), no matter what R and C values are used, due to the $1/f$ characteristic rolloff of a one-pole system. Note that the same data looks very different in this presentation than in the linear plot: the low-frequency asymptote and the "corner" shown in Figure 6.6 make more sense this way.

general relationship. It is often used in a one-sided form, especially when dealing with causal systems; the lower limit of integration is written as 0^- to imply that the integral includes any impulse at time zero:

$$H(s) = \mathcal{L}\{h(t)\} = \int_{0^-}^{\infty} \exp(-st)h(t)\,dt$$

But sometimes the two-sided version (or *bilateral* Laplace transform) with integration from negative infinity is used, like in the Fourier transform; if $h(t)$ is the impulse response of a causal system they are identical, but for transforming signals that might be nonzero before time 0, the two-sided version must be used.

Going the other direction, the inverse Fourier transform also serves as the inverse Laplace transform, at least in the case of stable causal systems, for which the Laplace transform integral converges wherever $\mathrm{Re}[s] = 0$:

$$h(t) = \mathcal{L}^{-1}\{H(s)\} = \frac{1}{2\pi} \lim_{\Omega\to\infty} \int_{-\Omega}^{\Omega} \exp(i\omega t)H(i\omega)\,d\omega$$

Don't worry, you'll never have to know or evaluate these integrals. There is good magic to help: *operator notation*, or *operational calculus*, which Oliver Heaviside came up with around 1880, having never heard of Laplace transforms. Promoting his methods, years later, he said "the best result of mathematics is to be able to do without it." Here's the main trick: just convert differential equations to algebraic equations on the (capitalized) signals, by putting an s wherever the original has a d/dt—supported by the observation that for all signals $A\exp(st)$, the derivative is $sA\exp(st)$. Our RC filter:

$$x - y = \tau\frac{dy}{dt}$$

becomes:

$$X - Y = \tau s Y$$

which we can easily solve for the output and transfer function as

$$Y = \frac{X}{\tau s + 1}$$

$$H(s) = \frac{Y}{X} = \frac{1}{\tau s + 1}$$

This simple algebraic formula for $H(s)$, arrived at without solving any differential equations or computing any integrals, tells us exactly how our RC filter will respond to any frequency of sinusoid, or to any exponentially growing or decaying sinusoid (subject to certain restrictions about regions of convergence of integrals, which is not an issue that we ever need to consider in practice; in this case, the region of convergence is $\mathrm{Re}[s] > -1/\tau$).

We mentioned the Laplace transform as converting between impulse responses and transfer functions. But in the equations above, the X and Y are also frequency-domain representations, related by the same transform to the corresponding time-domain signals. The algebraic expression for the ratio Y/X above can be interpreted as the ratio of the Laplace transform of the output, $Y(s) = \mathcal{L}\{y(t)\}$, to the Laplace transform of the input, $X(s) = \mathcal{L}\{x(t)\}$. These frequency-domain signals, functions of s, are usually written, as here, with the uppercase version of the time-domain signal name, just as with $h(t)$ and $H(s)$. The transforms do not necessarily exist, or converge, for general signals, but at least for absolutely integrable signals of finite duration they do. Even when they don't exist, the symbols can usually be manipulated as if they do, and ratios of them will still make sense, sometimes even for characterizing unstable and acausal systems.

There are lots of Laplace transform relationships, tricks, and tables that mean you never have to do integrals, or even understand them, to put the transforms to work. The trick we just used is that the s operator is the Laplace transform of the derivative operator; equivalently, that the derivative operation is an LTI system with transfer function s. That is the only trick needed to go from systems described as differential equations to their transfer functions, via simple algebra.

The other main trick we use about these transforms is that they (the Laplace transform or Fourier transform for continuous-time systems, or discrete-time Fourier transform or Z transform for discrete-time systems) map between domains in such a way that convolution in the time domain is equivalent to multiplication in the frequency domain, as described in Section 6.13. This observation is key to describing cascades of LTI systems, discussed in Section 6.14. We use it implicitly when we write algebraic representations in which multiplication by s represents application of a differentiation, and multiplication by s^{-1} represents an integration.

6.11 Rational Functions, and Their Poles and Zeros

A main advantage of the Laplace transform (and Z transform) is that it allows description of many linear systems in term of poles and zeros, making it easy to describe, design, analyze, and understand such systems.

The first-order RC filter, using $\tau = RC$, has the transfer function:

$$H(s) = \frac{1}{\tau s + 1}$$

In general, linear systems made from lumped components, that is, having a finite number of internal state variables, will lead by similar algebra to a transfer function in the form of a ratio of polynomials (sums of

E.E. Connection: Slightly More General Circuits

The transform variable s can be pushed all the way back into circuit elements. If Ohm's law (his electrical law $V/I = R$ relating voltage, current, and resistance of a resistive circuit element), and the idea of resistance as the ratio, are generalized to complex values, then the simple techniques of DC circuit analysis suddenly become capable of AC circuit analysis: analyzing the response for frequencies other than zero. Impedance (the complex version of resistance) can be thought of as the transfer function from current to voltage: $V(s)/I(s) = Z(s)$. The element impedances come from the usual trick of replacing differentiation by the s operator (and integration by $1/s$) in the definitions of the terminal relationships of the different component types. A capacitor of capacitance C is thus treated as an impedance equal to $1/sC$, and an inductor of inductance L as an impedance sL. This way, the electrical engineer can do a complete analysis of a circuit's frequency response algebraically, without ever looking at a differential equation or an integral. In mechanical systems, masses and springs can be similarly treated.

Systems of masses and springs, or inductors and capacitors, can *ring* in response to an impulse, rather than simply decay monotonically toward zero as our first-order example does. They can *resonate*, or respond strongly to signals of certain frequencies, as investigated in Chapter 8.

To simplify matters, we focus on examples in the form of the common *voltage divider* circuit of Figure 6.8. The impedance of block Z1 between the input and the output (for example, R in our first-order RC filter) is referred to as the *series impedance*, and that of block Z2 between the output and ground (which is $1/(sC)$ in the RC filter) as the *shunt impedance*. The currents I through the two impedance blocks Z1 and Z2 are equal, and the voltages across them (impedance times current, $Z_1 I$ and $Z_2 I$) add up to the input voltage X. The output voltage Y is $Z_2 I$, so the ratio of output to input voltage is easily found by canceling the I factors:

$$H = \frac{Y}{X} = \frac{Z_2}{Z_1 + Z_2}$$

For circuits of resistors, the impedances are just resistances, and the variables are all real scalars. The circuit is known as a *resistive voltage divider* in that case, and the output voltage is always less than the input voltage, by a ratio that does not depend on frequency. For circuits of *reactive* elements (inductors and capacitors), which store energy and induce frequency-dependent phase shifts and gains, we use the complex impedances, with the same algebra, to compute a complex frequency-dependent ratio, the transfer function. The first-order RC lowpass filter's transfer function can thus be found without explicit use of a differential equation. If the impedances Z_1 and Z_2 of the two blocks Z1 and Z2 include two reactive elements of different types (a capacitor and an inductor), the system would be second order and could resonate.

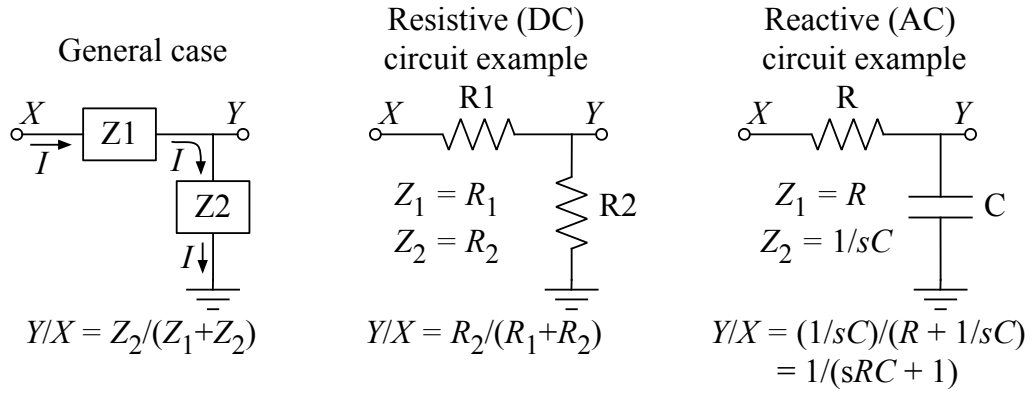


Figure 6.8: Schematic diagram of the general voltage-divider circuit (left), a simple resistive case, and a reactive case.

powers of s). This form is called a *rational function*—it is a very easy form to work with, compared to the expressions we get for systems made of distributed components, such as transmission lines.

As a complex function of a continuous complex-valued parameter, $H(s)$ looks like a formidable object. But when it is a rational function, we can represent it completely and succinctly by the *coefficients* of the numerator and denominator polynomials, or (to within a constant factor) by the *roots* of the polynomials (roots being the values of s for which the polynomial evaluates to zero), which are easier to visualize and reason about.

The roots of the numerator are the values of s for which $H(s)$ is zero—these locations in the *s plane* are called the *zeros* of the system. Similarly, the roots of the denominator, where $H(s)$ goes to infinity, are called the *poles* of the system. The poles and zeros (that is, their number and locations in the s plane), along with an overall gain factor, are a complete characterization of a linear system made of lumped components—and systems of distributed components can often be well approximated by these.

The example smoothing filter has a pole at $s = -1/\tau$, and no zeros. For convenience, we often refer to the position of such a pole in the s plane as something like p_1; for this filter, $p_1 = -1/\tau$.

Notice that the rational transfer function has real coefficients. These coefficients come from the real coefficients of the ordinary differential equations that define real systems. That means that for real systems, the roots of the numerator and denominator are either real (as in this first-order example), or in complex-conjugate pairs.

The poles correspond to frequencies s where the output is infinitely larger than the input; that means these frequencies can occur in the output when the input is zero—they are the frequencies of the homogeneous responses. The real part of the pole location, σ, controls the envelope $\exp(\sigma t)$ of the homogeneous response. It is negative if the homogeneous response decays in time, and positive if it grows in time. For stable systems, the output must decay with time, so $\sigma < 0$; that is, the poles of stable systems are in the left half of the s plane. A system may be *conditionally stable* if it has poles exactly on the imaginary axis—an *integrator*, one pole at $s = 0$, is an example.

A circuit made of passive components is always stable. When active components are included, it is possible to get poles in the right half plane, their positive real parts corresponding to growing exponentials. That is, the active component can provide the energy to make the output grow even when the input is zero. Such unstable systems are usually to be avoided.

When a pole is complex, at $p_1 = \sigma + i\omega$, the homogeneous response is an exponential envelope times the oscillating term $\exp(i\omega t)$. Such poles typically come from second-order filters, made for example by combining an inductor and a capacitor in a circuit. The result is known as resonance. Such systems tend to respond strongly to input frequencies near ω.

There is no such stability constraint on the zeros, but in many cases they too will be in the left half plane. In those cases, swapping the poles with the zeros, to make a system whose transfer function is the inverse of $H(s)$ (swapped numerator for denominator), will result in a stable inverse filter. Filters with all poles and zeros in the left half plane, or stable filters with stable causal inverses, are said to have the *minimum phase* property; they cause less phase lag, or delay, than any other stable causal system with the same magnitude transfer function. This property comes up in analysis of the cochlear traveling wave delay (de Boer, 1997; Recio-Spinoso et al., 2011), and in relating nonlinear tuning curves to filter-like behavior (Goldstein et al., 1971).

6.12 Graphical Computation of Transfer Function Gain and Phase

Besides the economy of description that poles and zeros provide—just lists of a few complex numbers to describe a linear system—they are also very useful as a way to make graphical representations of filters. A filter can be described as a simple diagram, with Xs (crosses) at the locations of poles in the complex s plane,

E.E. Connection: Second-Order Filter Circuits

Second-order filters are more interesting and relevant to hearing than our first-order example is, since they can be *resonant*, that is, particularly responsive to frequencies in a certain range. A resonant system is known as a resonator, or *single-tuned resonator* in the second-order case. Second-order resonant systems are the building blocks of almost all models of cochlear function.

In mechanics, examples of resonant systems are mass–spring systems and pendulums. Resonant systems generally have two different kinds of energy storage mechanisms, and dynamics that make the system's energy oscillate between the two types, for example, between the kinetic energy of a moving mass and the potential energy of a stretched or compressed spring.

In electrical circuits, inductors store energy as the kinetic energy of collectively moving electrons (Mead, 2002) (or in the magnetic field in the Maxwellian conception), and capacitors store potential energy by accumulating charges against a net electrostatic repulsion. The differential equations that describe the motion back and forth between kinetic and potential energy are essentially the same as for mechanical systems.

Consider the circuit of Figure 6.9, a second-order filter formed by adding an inductor to the series impedance of the RC lowpass filter. We call it "circuit A," the first of several resonant systems that we analyze in detail in Chapter 8.

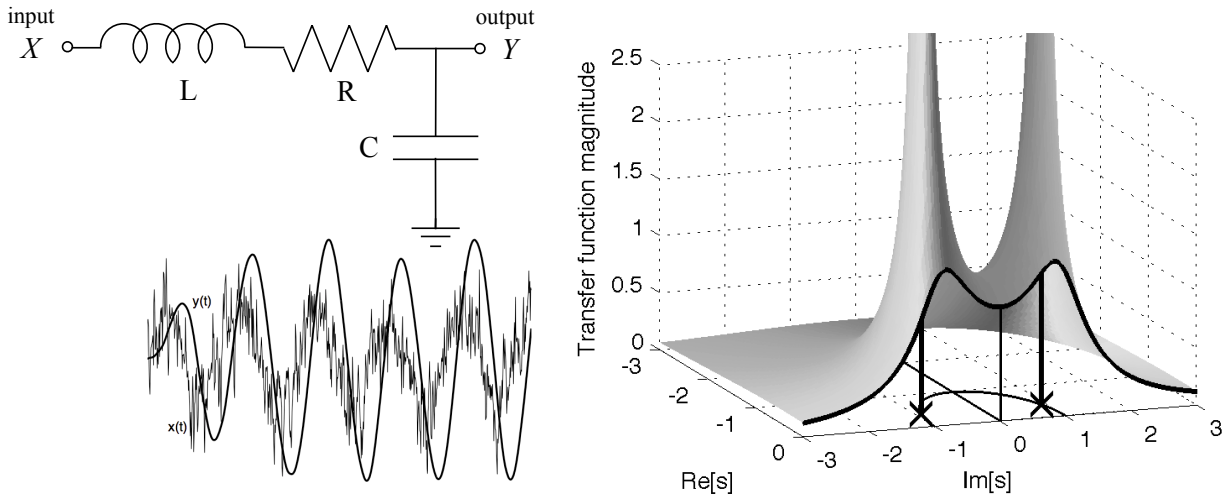


Figure 6.9: Filter A, a second-order resonant lowpass filter, is diagrammed (left) with an example of a noisy input waveform and a corresponding output waveform, which is smooth but has an increased amplitude of the input component that is close to the resonance frequency. The transfer function of filter A (right), plotted as in Figure 6.4 (see color plates), resembles a tent fabric draped over a pair of "tent poles" at the singularities, the two complex pole positions (at crosses and heavy vertical lines).

The Z1 block, or series impedance, has impedance $Z_1 = sL + R$, and the Z2 block, or shunt impedance, has impedance $Z_2 = 1/sC$, so the transfer function, from the voltage-divider approach described in the previous box, is

$$H_{\mathrm{A}}(s) = \frac{1/sC}{sL + R + 1/sC} = \frac{1}{s^2LC + sRC + 1}$$

This filter is termed second-order because it has two independent state variables: the voltage or charge stored on the capacitor, and the current stored in the inductor. In this case, the numerator is zero-order, with no roots, so the filter has two poles but no zeros. The poles can be real, for large enough R, but in the more interesting case, where the circuit is resonant, the poles are a complex conjugate pair.

and Os (circles) at the locations of zeros. This diagram is actually a computational tool to one who knows how to use it, and can lead to quick estimation (by eye) of transfer functions, or quick calculation with the aid of a ruler and protractor—measurements of distances and angles lead directly to gain magnitudes and phases, respectively. Even though we no longer need them for computation, it is still useful to understand this geometric view of rational functions, so that filters can be described by and understood through such diagrams.

We illustrate these manipulations for our one-pole–no-zero lowpass filter in Figure 6.10. The graphical computation of a resonant second-order frequency response is illustrated in Chapter 8. The two-pole filter gives us a chance to illustrate the idea of factoring, which makes it easy to see why the graphical methods work. In general, once we know the roots of the numerator and denominator, we have a factorization in terms of those values. For example, a two-pole–one-zero filter can be written in factored form as:

$$H(s) = \frac{A(s - z_1)}{(s - p_1)(s - p_2)}$$

where A is a gain factor, z_1 is the position (in the s plane) of the zero, and p_1 and p_2 are the positions of the first and second poles (typically with $p_2 = p_1^*$). The first-order pole-only filter doesn't need factoring, but can be arranged in a similar form as

$$H(s) = \frac{A}{(s - p_1)}$$

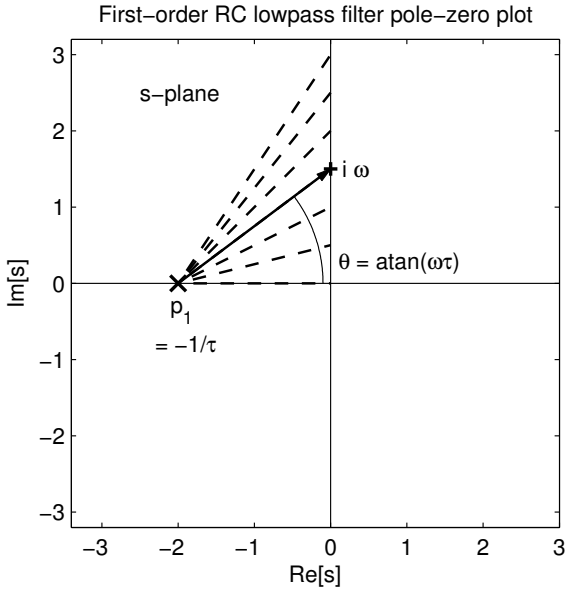


Figure 6.10: Calculating the frequency response of the example RC lowpass filter from an s-plane pole–zero diagram with one pole at $p_1 = -1/\tau$ and no zero: $1/(s - p_1)$. The magnitude response is inversely proportional to the length of $s - p_1$, the line from the pole p_1 to the frequency point at $s = i\omega$ (shown here for $\omega > 0$), and the phase lag is the angle θ of that line from the real axis. Since the angle θ appears in the denominator, it becomes a negative phase shift, which represents a lag, or delay (it would be a positive phase shift for $\omega < 0$, which is still a lag).

To compute the frequency response of such a filter, we plug in various values of $i\omega$ (or $i2\pi f$) for s, and evaluate. Each factor in the numerator or denominator, of the form $(s - x)$, becomes $(i\omega - x)$. The magnitudes of these are just distances in the s plane diagram, from the location of the pole or zero (x) to a point on

the imaginary axis ($i\omega$). The magnitude frequency response is then just the product of distances from zeros divided by the product of distances from poles, to this point $i\omega$ on the imaginary axis. Treating the differences as vectors, we can also get the phase of the frequency response as the sum of vector angles to zeros minus the sum of vector angles to poles. The operations are simple enough that engineers routinely develop an intuition for filter behavior based on looking at pole–zero plots. The extra gain factor A is typically ignored; the gain at DC (at $i\omega = 0$) is often taken to be 1, if a normalization factor is needed.

Applying this procedure to the resonant second-order system is more interesting, since moving along the imaginary axis can take one close to the position of pole, at the resonant frequency, and then further again at higher frequencies. So the gain goes up near the pole frequency, then back down again.

6.13 Convolution Theorem

When the output of one filter becomes the input to another, the filters are said to be in *cascade*. For LTI systems in cascade, the final output is obtained from the input by applying in turn the time-domain or frequency-domain operators (via convolution or multiplication, respectively) for the two filters.

This equivalence between multiplication and convolution is the basis of the algebraic operator notation, and is often expressed as the *convolution theorem*, which states that a product of Laplace transforms of two signals (for example, $h(t)$ and $x(t)$) is equal to the Laplace transform of the convolution of those signals:

$$H(s)\,X(s) = \mathcal{L}\{h(t) * x(t)\}$$

These operations are commutative and associative, so can be applied not only for getting a filter output from its input and impulse response, but also for combining the responses of cascaded filters, as:

$$H_2(s)\,H_1(s) = \mathcal{L}\{h_2(t) * h_1(t)\}$$

Similarly, the convolution theorem works for more than two factors:

$$H_2(s)\,H_1(s)\,X(s) = \mathcal{L}\{h_2(t) * h_1(t) * x(t)\}$$

Mathematically, for the theorem to apply, the relevant Laplace transforms must exist in some region of the s plane, and the resulting Laplace transform will exist in the intersection of those regions.

6.14 Interconnection of Filters in Cascade, Parallel, and Feedback

The first-order and second-order filters considered above are often useful, but have a rather simple range of behaviors. To make more interesting systems, we can combine simple systems in various ways, such as the patterns shown in Figure 6.11.

For the cascade connection in Figure 6.11-A, the convolution theorem gives a net transfer function that is the product of the individual transfer functions:

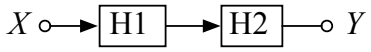
$$y(t) = h_2(t) * h_1(t) * x(t)$$

or

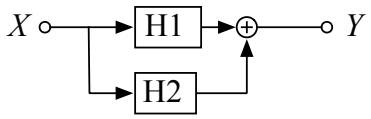
$$Y(s) = H_2(s)\,H_1(s)\,X(s)$$

where the order of application of the operator has been written to imply that H_1 is applied to the input first, then H_2. That is, the effective net filter is characterized by the product $H_1(s)\,H_2(s)$ of the transfer functions, or

A. Systems in cascade



B. Systems in parallel



C. Systems in a feedback loop

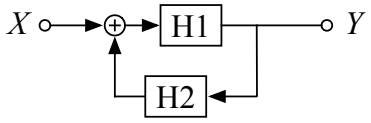


Figure 6.11: Example filter systems interconnected in cascade, parallel, and feedback configurations.

by the convolution $h_1(t) * h_2(t)$ of the impulse responses, of the individual filters. This works for any number of LTI systems in cascade, in any order.

When one input goes to two filters, and the outputs of the two filters are added together to make a single output, as in in Figure 6.11-B, this *parallel* interconnection of filters is equivalent to a single filter, described by the sum of the impulse responses or equivalently by the sum of the transfer functions:

$$y(t) = [h_1(t) + h_2(t))] * x(t)$$

or

$$Y(s) = [H_1(s) + H_2(s)]\, X(s)$$

That is, the Laplace transform leaves addition as addition when converting between domains.

If systems are connected in a feedback configuration, with a path from an input being added to a path coming back from an output or some intermediate point, the algebraic description can be solved for a net transfer function. In such cases, transfer functions combine in algebraic expressions that involve division; rational transfer functions still lead to rational-function results. Such algebraic division is easy in the frequency domain, with rational functions; but there is no equivalently simple algebra for combining impulse responses in feedback systems, as there are for combining impulse responses in parallel and cascaded systems (addition and convolution, respectively). For the example shown in Figure 6.11-C, the relationship between input and output is found by applying the additions and multiplications suggested by the internal parallel and cascade portions:

$$Y = H_1\,(X + H_2\,Y)$$

which is solved to yield the net system transfer function:

$$\frac{Y}{X} = \frac{H_1}{1 - H_1\,H_2}$$

The general rule of thumb to remember for simple one-loop feedback systems is that the net input-to-output gain is equal to the forward gain over one minus the loop gain.

For systems with rational transfer functions, we can reason about how their poles and zeros combine. The cascade is the simplest case: the sets of poles and zeros simply get combined (actually, they're not sets, but *bags*, since any given location may appear more than once, as when cascading two or more identical filters). Going the other direction, from a system to a cascade of simpler systems, is a factorization, corresponding to putting some of the poles and zeros in one system, and some in the other. For example, one factorization of a system would be as a cascade of an all-pole filter and an all-zero filter. When factoring systems into cascades, we usually avoid splitting up complex-conjugate pairs of poles or zeros, so that all the cascaded component systems will have real-valued inputs and outputs.

For systems in parallel, the net system combines the sets of poles, but not the zeros. That is, at any frequency where either system has an infinite output, so will their sum. But to get zero output, the two systems need to cancel each other, by having equal gain magnitudes and opposite phases. So some analysis is needed to find the locations of the zeros. We will use such an analysis in Chapter 9 to find the zeros of the gammatone filter that is popular in modeling hearing.

For systems in a feedback configuration, zeros are collected from any components in the forward path H_1, but new poles and zeros also come from the zeros and poles, respectively, of the denominator $1 - H_1 H_2$, one minus the loop gain. The poles of this denominator, the new zeros, are the union of the poles of H_1 and H_2; but the poles of H_1 cancel the poles in the forward path, so only the poles of H_2 end up as new zeros. The zeros of this denominator, the new poles, are more complicated, due to the one minus.

The cascade and parallel connections of stable filters are stable. But with feedback, it is quite easy to make unstable systems from stable ones, and vice versa. The roots of the denominator $1 - H_1 H_2$ need to be in the left half of the s plane for the feedback system to be stable, but these roots are not among the original poles and zeros, so some checking is required.

We will encounter examples of all of these canonical compositions—cascade, parallel, and feedback—in our machine hearing systems.

6.15 Summary and Next Steps

Linear systems are analyzed by several powerful interrelated methods. We can describe linear systems, or filters, as circuits or as differential equations. We can characterize them via impulse responses or transfer functions, and move between these characterizations using transforms. For the class of circuits made from lumped elements, we can represent transfer functions as ratios of polynomials, and factor those polynomials to get descriptions in terms of poles and zeros. With poles and zeros we have a compact description that yields a simple diagram and supports visualization and graphical calculation of frequency responses.

We've seen why sine waves are important: they are as close as we get, in real bounded signals, to the eigenfunctions of linear systems—put one in and you get the same one out, slightly modified but with the same frequency.

In later chapters, we look more at *resonant* linear systems, and build up to models of filtering in the cochlea.

While this chapter summarizes the key points needed to understand the developments in this book, many readers can benefit from either a deeper treatment or a more introductory treatment. There are numerous good books that cover linear time-invariant systems, including theory and applications. Among my favorites are Siebert's *Circuits, Signals, and Systems* (Siebert, 1986) and Oppenheim and Willsky's *Signals & Systems* (Oppenheim and Willsky, 1997). I also like Hamming's development of the frequency concept and smoothing filters in his *Digital Filters* (Hamming, 1998).

For a more elementary introduction to sound, linear systems, and hearing, Rosen and Howell's *Signals and Systems for Speech and Hearing* (Rosen and Howell, 2011) is recommended. Another excellent book

E.E. Connection: On Cascades

The notion of a cascade of filters is so central to our models of hearing that it needs a little extra attention. There are at least three relevant contexts in which cascades appear prominently in linear systems and hearing literature. These contextual meanings are closely related, and we use them all, but discussions in the literature are sometimes limited to narrower interpretations.

First, as a description of a way of interconnecting filter circuits, as traditionally used in the radio, telephone, and television fields, cascades are sometimes mentioned in the sense analyzed in this chapter. Specifically, in a cascade connection, the filter stage circuits are "buffered" (by a unity-gain follower amplifier) such that the output of each one is not disturbed when it is connected to the next; that is, such that there is no backwards influence, coupling, or "loading" from the next connected circuit. Cascades that implement what we now call gammatone filters were described in the 1940s by Eaglesfield and by Tucker. Tucker (1946) explains the difference between the generic concept of "in series" and "cascade," a difference we illustrate in Figure 6.12: "The use of a series of tuned circuits coupled together by mutual inductance, mutual capacitance, or resistance is well known, and the response of such an arrangement is analysed in various textbooks and papers. The use of a cascade of tuned circuits which have transmission coupling but no mutual coupling is not so often referred to, however, although such an arrangement occurs frequently in practice." Eaglesfield (1945) referred to his cascades simply as multistage amplifiers, where by stages he meant the circuit portions isolated from backward coupling by buffer amplifiers.

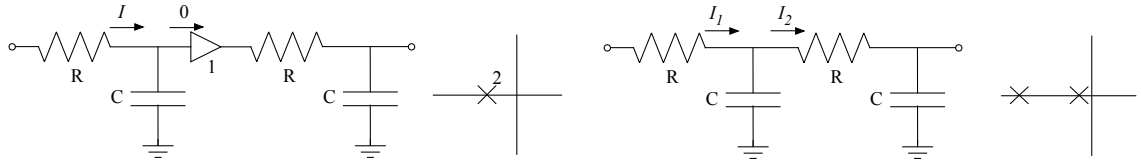


Figure 6.12: A cascade of two identical RC filters separated by a buffer amplifier (left) and without a buffer (right). The buffer (triangle with label "1" indicating its gain), takes zero input current and copies its input voltage to its output, so the second filter stage does not influence the first stage by taking current out of it, or "loading" it. The resulting transfer function H^2, the square of the one-pole stage transfer function, has two poles in the same location, at $s = -1/RC$, as shown in the left pole plot. Without the buffer, some of the current through the first resistor flows into the second, contrary to what we assumed when we computed the transfer function of the RC filter. Such mutual coupling—the second stage affecting the first stage—results in a different filter, with two different pole locations as shown in the right pole plot.

Second, as a description of the propagation of waves along a continuum or sequence of places, a cascade of filters representing each place-to-place transition is a natural description. In modeling the propagation of sound-induced waves in the inner ear, the notion of a cascade was articulated by Licklider (1953) as a description of the traveling-wave model: "...the passage of the traveling wave down the partition–fluid system, which if it is viewed as consisting of resonators, consists of many elements in cascade and not in parallel." Here Licklider's cascade concept may also have been intended to cover the then-current electrical network models of the cochlea that would actually support wave propagation in both directions; that is, he did not necessarily recognize Tucker's distinction quoted above. At least one paper on modeling waves in the cochlea has used *cascade* for simply interconnected lumped circuits (Kletsky and Zwislocki, 1981), but most of us in the auditory modeling field use the term as Tucker suggests, reserving it for models that propagate signals through stages only in the forward direction (Lyon, 1998; Sarpeshkar, 2000).

Finally, cascades are an implementation strategy for many kinds of digital filters. A typical textbook on digital signal processing will mention cascades only in this sense, as the usually most natural and robust way to break up a filter design into easily implementable pieces.

In this book, all of these senses are important, and mostly equivalent. In the case of the cascade model of wave propagation in the inner ear, multiple outputs are also very important. The "multiple-output cascade filterbank" is the natural structure for efficiently implementing a model of sound processing in the cochlea.

that connects the engineering to what's known about hearing is Hartmann's *Signals, Sound, and Sensation* (Hartmann, 1998).

Chapter 7

Discrete-Time and Digital Systems

> Since a sufficiently approximate solution of many differential equations can be had simply by solving an associated difference equation, it is to be expected that one of the chief fields of usefulness for an electronic computor would be found in the solution of differential equations.
>
> — "The use of high-speed vacuum tube devices for calculating," John Mauchly (1942)

7.1 Simulating Systems in Computers

When we build or model linear (or nonlinear) systems in computers, we use discrete time steps. The relationship between discrete-time and continuous-time linear systems, and the implementation of such systems in digital computer software, are introduced in this chapter. Fortunately, all of the representations and techniques used for discrete-time linear systems are exact counterparts of those used for continuous-time linear systems, with difference equations taking the place of differential equations, and *Z transform*s taking the place of Laplace transforms, as we explain below.

A *discrete-time system* operates with quantized time, and can be linear. A *digital* system, or *digital filter*, on the other hand, represents signals as quantized in both time and amplitude, and therefore cannot be precisely linear. Digital systems are what we can make with digital hardware or software. In the old days of digital filters (1960s and 1970s), people worried a lot about quantization effects, the manifestations of the digital number system's nonlinearity. In more modern times, however, this distinction between discrete-time and digital is typically ignored, as we economically operate with digital floating-point number systems with enough resolution to simulate linear discrete-time systems with more than enough accuracy. Still, when we say *digital filter* we're likely referring to a computer implementation that can process a signal, as opposed to a *discrete-time linear system* which refers to a more idealized analytical model.

7.2 Discrete-Time Linear Shift-Invariant Systems

Electrical, mechanical, and other physical systems, such as those in our ears, are continuous-time systems, whether linear or not. But when we build computer-based systems to process sounds, we work on *samples* of sound waveforms, measured at (usually) equally-spaced sample times. So we use the theory of discrete-time linear systems, which are in most respects analogous to continuous-time linear systems.

Mathematically, a discrete-time system is an operator that maps an input sequence $x[k]$ to an output sequence $y[k]$, where k is the integer index of the sequence (these sequences are conceptually defined for all integer indices, from negative infinity to infinity). Just as with continuous-time systems, linearity means that if the system maps the sequence x_1 to y_1, and x_2 to y_2, then it maps $ax_1 + bx_2$ to $ay_1 + by_2$. And shift-invariant

means that if $x[k]$ maps to $y[k]$, then the shifted sequence $x[k-n]$ maps to $y[k-n]$ for any constant integer shift n.

7.3 Impulse Response and Convolution

In the discrete-time case, a *unit impulse* has a sample of value 1 at time index 0, surrounded by values of 0 for the infinite past and future. The *impulse response* is the sequence $h[k]$ that comes out of a filter when a unit impulse is provided as the input.

Any sequence $x[k]$ can be expressed as a sum of scaled and shifted impulses. The output $y[k]$ of a linear shift-invariant system is then easy to express as a sum of correspondingly scaled and shifted copies of $h[k]$. The sum that expresses this relationship is known as the convolution sum, and is analogous to the convolution integral of continuous-time systems:

$$y[k] = \sum_{n=-\infty}^{\infty} x[n]h[k-n]$$

For causal systems, the impulse response is zero at negative indices. In that case, the convolution sum can be written as a sum over only nonnegative indices of the impulse response, in a form that makes it clear that the samples of h are weights applied to samples of x at the index of the output being computed (k) and earlier ($k-n$ for positive n):

$$y[k] = \sum_{n=0}^{\infty} x[k-n]h[n]$$

7.4 Frequency in Discrete-Time Systems

These systems are also amenable to analysis in terms of frequency. We tie samples to time via a sampling interval T, or sample rate (sampling frequency) $f_s = 1/T$. The eigenfunctions are again sinusoid-like: sampled complex exponentials. But we write them differently, as geometric sequences, powers of the complex parameter z:

$$x[k] = A_x z^k$$

With $z = \exp(sT)$, $x[k]$ represents discrete samples of the continuous complex exponential $A_x \exp(st)$ at times $t = kT$, for all integers k. This z is a conventional variable for a complex frequency in discrete-time systems. It is the frequency-domain variable of the Z transform, analogous to the s of the Laplace transform. Like s, z is not only a generalized frequency, but it also represents an operator that can be used algebraically, as a factor in the transform domain.

7.5 Z Transform and Its Inverse

The Z transform of a sequence $x[k]$ is defined as a function of the transform variable z:

$$X(z) = \mathcal{Z}\{x[k]\} = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$$

for sequences that are defined for all k. There is also a one-sided Z transform, with the lower summation limit at 0, for sequences defined only for nonnegative k. For transforming causal impulse responses, the two versions are equivalent. For signals that might be nonzero before time 0, including the sampled complex

exponentials that are the eigenfunctions of discrete-time linear shift-invariant systems, the two-sided version must be used.

While the frequency parameter z is lowercase, we adopt the convention of capitalizing the name of the Z transform (the alternative lowercase convention is also found in the literature). As before, we'll use uppercase/lowercase variable pairs for a signal and its transform: $x[k]$ transforms to $X(z)$. The discrete-time index (k) is in square brackets, while continuous-valued arguments (t, s, ω, or z) are in parentheses.

The transfer function of a discrete-time system is a function of the complex frequency z, and is the ratio of the Z transforms of X and Y, as well as being the Z transform of the impulse response, exactly as with Laplace transforms for continuous-time systems:

$$H(z) = \frac{Y(z)}{X(z)} = \mathcal{Z}\{h[k]\}$$

That is, the Z transform is the operator that maps from convolution in the discrete-time domain to multiplication in the complex-frequency domain. The convolution theorem, analogous to the continuous-time convolution theorem of Section 6.13, tells us:

$$H(z)\,X(z) = \mathcal{Z}\,\{h(k) * x(k)\}$$

The inverse Z transform computes the impulse response from the transfer function:

$$h[k] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} H(\exp(i\theta))\exp(i\theta k)d\theta$$

(this formula with integration around the unit circle works at least for stable systems, to which we restrict our attention to avoid consideration of regions of convergence and acausality and such complications).

Notice that, as with continuous-time systems, the inverse transform is an integral over frequencies; but here the integrand evaluates $H(z)$ at $z = \exp(i\theta)$, that is, along the unit circle in the z plane (the angle θ in the z plane is related to frequency by $\theta = \omega T$, and has units radians per sample). Compare this to the inverse Laplace transform, an integral in which $H(s)$ is evaluated at $s = i\omega$, along the imaginary axis. These paths, the imaginary axis in the s plane and the unit circle in the z plane, are very special, as they represent the sets of complex frequencies s and z that correspond to sinusoids that neither grow nor decay with time. For the filters to be stable, and for the inverse transform integrals to be meaningful as stated, there must be no poles on or to the right of the imaginary axis in the s plane, or on or outside the unit circle in the z plane. It is also possible to compute transfer functions of unstable or acausal systems, as we mentioned in Section 6.6, but we won't often need to.

7.6 Unit Advance and Unit Delay Operators

As with Laplace transforms, it will not often be necessary to evaluate or manipulate the transforms directly, since simple operator methods provide easy shortcuts. In this case, multiplication by z in the transform domain represents the *unit advance* operator. This $z\{\cdot\}$ is a peculiar thing: it represents a view into the future, looking at the next later sample of a sequence:

$$z\{x\}[k] = x[k+1]$$

Notationally, using the frequency parameter z this way in a time-domain expression does not make much sense; but treating z as an algebraic operator, and converting x to its frequency-domain version X, similar to how we used s as the derivative operator, produces in the frequency domain algebraic expressions in variable z. That is, the notation above really means that if we have a signal $x[k]$ and its Z transform $X(z)$, then $zX(z)$ is

Statistics and History Connection: Generating Functions and Z Transforms

In the field of statistics, discrete sequences are sometimes described by *generating functions*, which are essentially the same as what we call Z transforms. When the discrete sequence is the probability mass function of a discrete random variable, its generating function is called a probability-generating function.

Generating functions were described by Abraham de Moivre in 1730, in proving his formula for the distribution of the sum of several independent identically distributed random variables (such as the total number of pips showing on six dice), and were named by Pierre-Simon Laplace in 1780 (Hald, 2005).

The theory of discrete-time systems with Z transforms was worked out at the MIT Radiation Laboratory during World War II by the mathematician Witold Hurewicz (1947)—for analyzing predictors of variables such as airplane positions as part of their radar development effort, as recounted by Bennett (1993). The name was provided a few years later, in follow-up work by Ragazzini and Zadeh (1952), who discussed the relationship to generating functions in depth.

the Z transform of the advanced signal:

$$zX(z) = z\mathcal{Z}(x[k]) = \mathcal{Z}(x[k+1])$$

which is obvious from the definition of the Z transform.

The unit advance operator z is a noncausal filter: its current output is a not-yet-arrived sample of its input. That's why we more typically work with its inverse, z^{-1}, which represents a *unit delay*, that is, a one-sample delay of a signal, which is a causal filter:

$$z^{-1}\{x\}[k] = x[k-1]$$

$$z^{-1}X(z) = z^{-1}\mathcal{Z}(x[k]) = \mathcal{Z}(x[k-1])$$

Sometimes the time-domain operators are represented by different symbols, to avoid the awkward overload of the generalized frequencies; D is common for differentiation, and D or Δ is also sometimes used for a unit delay, though just overloading z as the unit advance operator seems to be most common. In the geophysics field, however, signs are reversed and z is the unit delay operator.

The unit delay is typically implemented as a register in hardware, or a variable in software, that holds a value for use in the next iteration of a digital filter. So we see z^{-1} frequently in signal-flow diagrams that represent what our digital filter hardware or software will do, as in Figure 7.1. On the other hand, we usually express transfer functions in terms of z, and we plot poles and zeros in the z plane.

7.7 Filters and Transfer Functions

Discrete-time filters are typically divided into two classes: *recursive* (or infinite impulse response, IIR) and *nonrecursive* (or finite impulse response, FIR, or transversal), depending on whether they include feedback loops such as the one that the filter on the right in Figure 7.1 has. In special cases, when a zero cancels a pole, a recursive filter, that is, one with feedback, can have a finite impulse response, so one can quibble over whether to use these alternative terms interchangeably.

A difference equation and corresponding transform-domain algebraic description for a simple nonrecursive filter, corresponding to the signal-flow diagram on the left in Figure 7.1, are:

$$y[k] = bx[k] + ax[k-1]$$

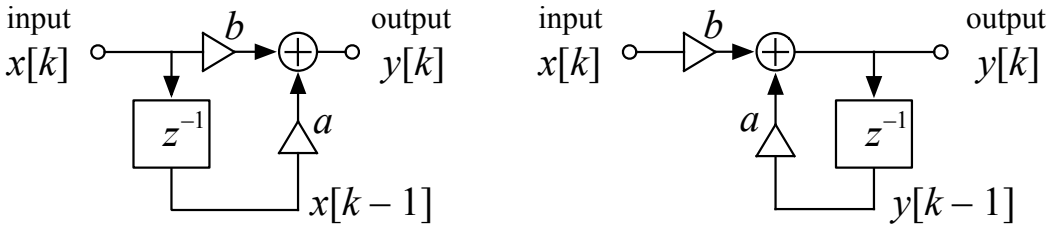


Figure 7.1: The signal-flow diagrams of two different first-order digital filters are shown here: a finite-impulse-response (FIR), or nonrecursive, filter on the left, and an infinite-impulse-response (IIR), or recursive, filter on the right. With appropriate choice of coefficients a and b, namely $a = 1 - b$, with both coefficients between 0 and 1, these are smoothing filters (that is, they suppress fluctuations and have average output equal to average input). The FIR filter on the left can only do a little bit of very local smoothing, while the IIR filter on the right behaves like the RC lowpass filter, with potentially very long time constant, when the coefficients make it a smoothing filter. The label z^{-1} in the box represents the unit delay operator. The triangles represent multiplication by the constant coefficients shown. The two filters' difference equations, $y[k] = bx[k]+ax[k-1]$ and $y[k] = bx[k] + ay[k-1]$, and z-domain operator equations, $Y = bX + z^{-1}aX$ and $Y = bX + z^{-1}aY$, are apparent from the signal-flow diagrams. Labeling the input $x[k]$ and the output $y[k]$ as here suggests the operation steps on sequences of samples, with z^{-1} being a unit delay, or memory element. We alternatively label them X and Y, which suggests interpreting the z^{-1} as a transform operator. We switch back and forth between such labels and interpretations, as we did in Chapter 6.

$$Y(z) = bX(z) + az^{-1}X(z)$$

$$H(z) = \frac{Y}{X} = b + az^{-1} = \frac{bz + a}{z}$$

For $0 < a < 1$ and $a + b = 1$, this is a smoothing filter with unity gain at DC (at $z = 1$): each output sample is a weighted average of two adjacent input samples. The causal impulse response is the sequence $[b, a, 0, 0, \ldots]$. A typical design has $a = b = 0.5$, corresponding to a 2-point moving average, which has $H(z) = 0$ at $z = -1$ (that is, at the Nyquist frequency, π radians per sample). Alternatively, when the coefficients are of opposite sign, this filter will emphasize high frequencies and suppress low frequencies—an antismoothing filter.

The recursive case is more interesting, because it can smooth over a longer time, not just across immediately neighboring points. The first-order recursive lowpass filter can be specified in terms of the signal-flow diagram on the right in Figure 7.1, or as the corresponding difference equation or *recurrence relation* as

$$y[k] = bx[k] + ay[k-1]$$

where $y[k-1]$ is the previous (unit-delayed) output value being used to compute a present output value $y[k]$, and a and b are the *coefficients* that define the filter. In the transform domain, the system description can be read directly off the flow diagram, treating the delay operator z^{-1} as an algebraic factor just like the coefficients a and b:

$$Y = bX + z^{-1}aY$$

$$Y(1 - z^{-1}a) = bX$$

$$H(z) = \frac{Y}{X} = \frac{b}{1 - z^{-1}a} = \frac{bz}{z - a}$$

This one-pole transfer function is illustrated in Figure 7.2. The figure also illustrates a two-pole discrete-

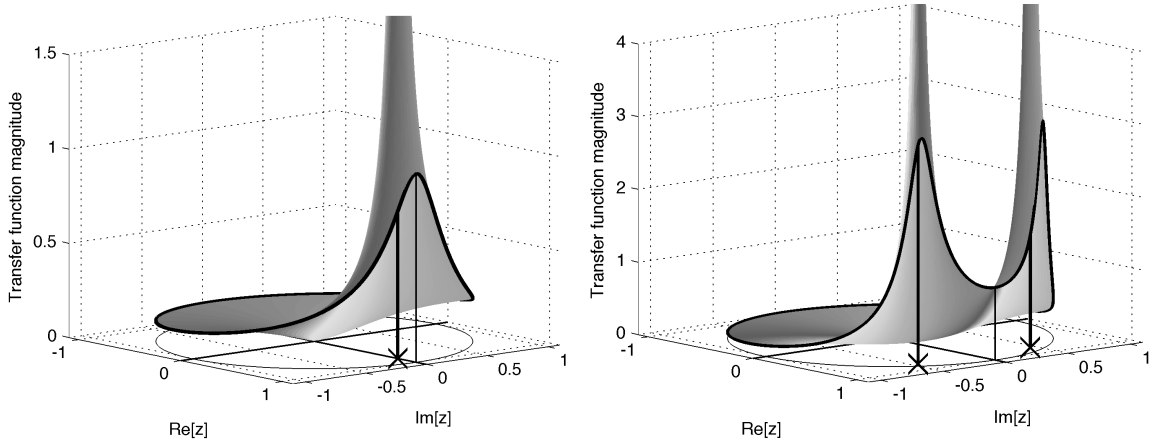


Figure 7.2: Complex transfer functions of one-pole (smoothing) and two-pole (resonator) filters, evaluated inside the unit circle of the z plane. The frequency response is the transfer function evaluated on the unit circle, shown by the dark curves at the circular cut. As in Figure 6.4, phase is mapped to hue (see the color plates); there is one cycle of hue variation around each pole.

time resonator transfer function. These are the discrete-time versions of the one-pole and two-pole transfer functions shown in a similar style in Figure 6.4 and Figure 6.9, respectively.

As with Laplace transforms, when the transfer function is a ratio of polynomials, the roots of the denominator and numerator (the system's poles and zeros in the z plane) can be used as a compact and powerful description of the system, and can be used to visualize, or graphically compute, the transfer function. Relative to continuous-time systems, the difference is that to compute the response at a particular frequency ω, one measures distances and angles from poles and zeros to the point $z = \exp(i\omega T)$ on the unit circle in the z plane, as opposed to the point $i\omega$ in the s plane. Transfer function calculations from z-plane pole–zero plots are illustrated in Figure 7.3 and Figure 7.4. The highest frequency that can be unambiguously interpreted, $\theta = \omega T = \pi$, or one-half cycle per sample, maps to $z = -1$ in such plots, since the frequency response repeats cyclically as the frequency point moves around the circle. Compare the resulting frequency response plot on the right in Figure 7.3 to Figure 6.6; in the discrete-time case, the $1/\omega$ approximate rolloff of high frequencies is not actually an asymptote, due to the circularity.

Just as s can be pushed all the way back to circuit elements, z is easily pushed into the block diagrams or signal-flow graphs that are typically drawn to represent discrete-time systems, such as in Figure 7.5. A discrete-time linear system described by a difference equation with constant coefficients can be drawn as a set of signal-flow paths with gains on them (the fixed coefficients), connecting summing nodes and delays. Each unit delay is labeled with its transfer function, z^{-1}, and each gain by its transfer function, the scalar coefficient. Then, as we did for the first-order filter above, the equations that relate input and output can be written by inspection and manipulated into the form of a rational transfer function.

7.8 Sampling and Aliasing

Discrete-time filters don't represent all the details of continuous-time signals, except under special limitations. The *Nyquist–Shannon sampling theorem* specifies one such condition: a continuous-time signal with no power at frequencies higher than a bandwidth W Hz can be exactly reconstructed from its samples spaced at sampling interval T seconds, as long as $2W < 1/T$; that is, if the sampling rate f_s exceeds twice the signal bandwidth.

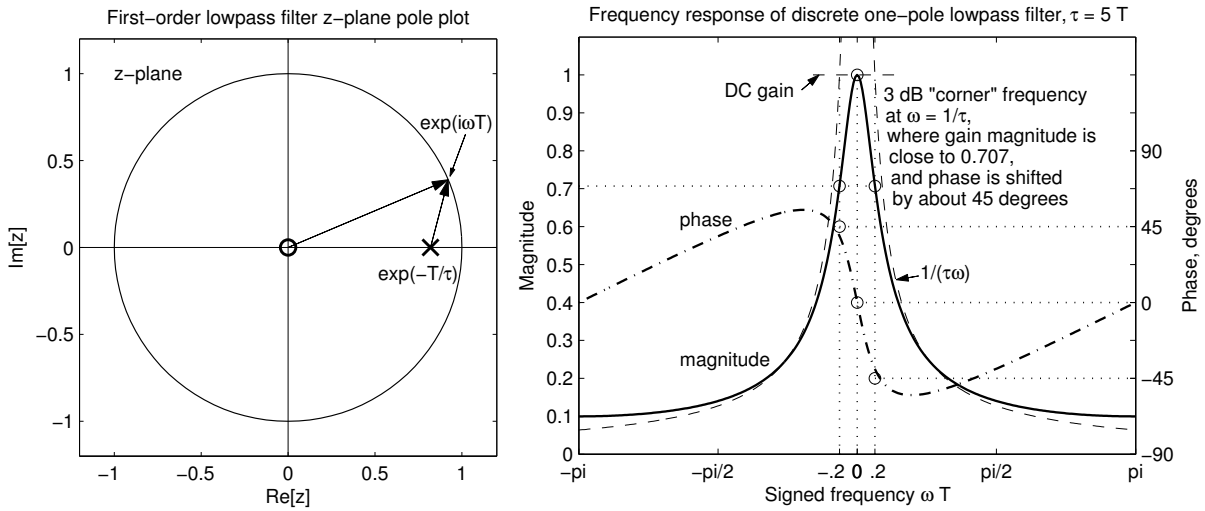


Figure 7.3: Calculating the frequency response of the example first-order discrete-time lowpass filter from a z-plane pole–zero diagram with one pole and a zero at the origin. In this example, the pole is located at $z = \exp(-0.2)$, corresponding to $T/\tau = 0.2$, a smoothing time constant of 5 samples (for example, 5 ms time constant at 1 kHz sample rate, $T = 0.001$ s). For any frequency ω, the magnitude response is inversely proportional to the length of the line from the pole to the frequency point at $z = \exp(i\omega T)$, and the phase lag is the angle between the real axis and that line. Since the zero is at the origin, its distance to the frequency point on the unit circle is always 1, so the zero does not affect the gain magnitude (any other position of the zero would affect the gain magnitude); this zero does provide a phase lead, however, which reduces the net phase lag.

Detail: Zeros at the Origin

Why does the one-pole continuous-time filter correspond to a one-pole–one-zero discrete-time filter?

$$H(z) = \frac{bz}{z - a}$$

The factor of z in the numerator of the example lowpass transfer function represents a zero at the origin. It means the output is advanced by one sample from what a filter without that zero would do, that is, compared to the case where the filter output is taken *after* the delay element in Figure 7.1. After that delay, the output would be $y[k-1]$, or $z^{-1}Y$. The pole at the origin in $1/z$ cancels the zero in z in that case, corresponding to omitting the z in the transfer function.

Being at the origin, the zero in z affects the phase, but not the magnitude, of the frequency response: the factor of z, or $\exp(i\omega T)$, contributes a phase advance of ωT at frequency ω rad/s.

When a discrete-time filter has more poles than zeros, zeros at the origin can be added, by adding factors of z, advancing the output, reducing the phase lag while keeping the filter causal. The filter is not minimum phase without them (the concept of minimum phase was introduced in Section 6.11), because the inverse filter would not be causal.

There is no corresponding concept for continuous-time systems with rational-function transfer functions; s-domain points that would map to $z = 0$ are infinitely far to the right side in the s plane, where zeros have no effect at finite frequencies.

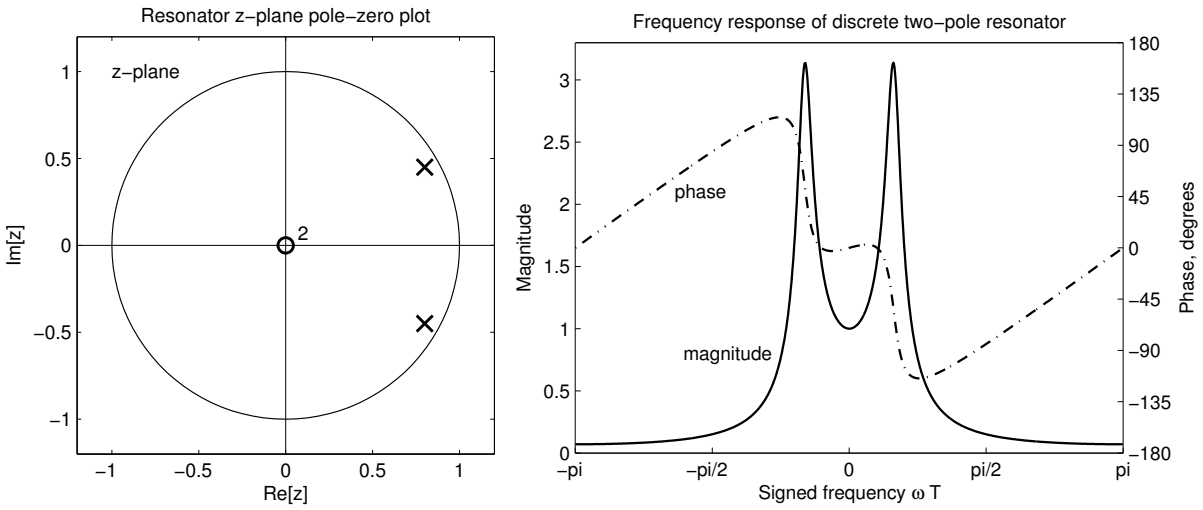


Figure 7.4: The pole–zero plot and frequency response of the example second-order discrete-time resonator of Figure 7.2. The response magnitude is proportional to the product of the reciprocal distances between points on the unit circle and the two poles, so the positive and negative frequencies that are close to the upper and lower poles both produce gain peaks. Following the process described in Figure 7.3 for each pole and zero, the log magnitude gains of the two poles add, and so do the phases contributed by the two poles, as these are the real and imaginary parts of the complex log of the complex gain. The two zeros at the origin add to the phase, but do not affect the gain magnitude.

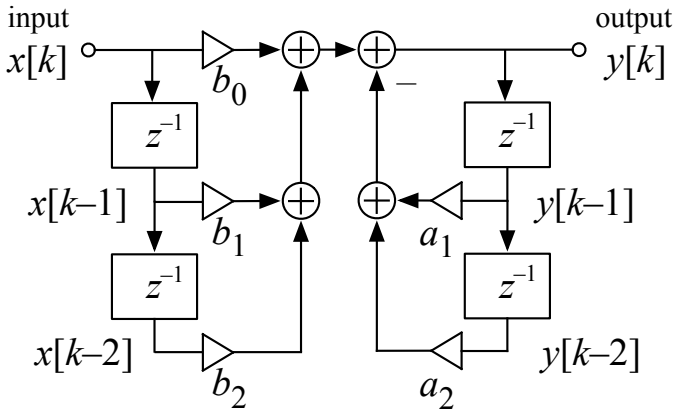


Figure 7.5: The signal-flow diagram of a second-order digital filter with three forward coefficients and two feedback coefficients. The filter's difference equation $y[k] = b_0x[k] + b_1x[k-1] + b_2x[k-2] - a_1y[k-1] - a_2y[k-2]$, which is apparent from the diagram, is also the program step that computes an output sample.

This limitation, $2W < f_s$, is known as the *Nyquist criterion*.

Shannon and Weaver (1948) expressed the sampling theorem this way:

> If a function $f(t)$ contains no frequencies higher than W cycles/sec it is completely determined by giving its ordinates at a series of points spaced $1/2W$ seconds apart.

Samples of complex exponentials $x(t) = \exp(st)$ can be written as $x[k] = z^k$, for $z = \exp(sT)$ when sampling at interval T. This mapping $z = \exp(sT)$ is not one-to-one. Due to the periodicity of the exponential function, different values of s map to the same value of z, which means that several different continuous-time signals lead to identical sample sequences. Different continuous-time signals with identical samples are known as *aliases* of each other. When we reconstruct a signal from its samples, we usually try to reconstruct the lowest frequency of the various aliases, which will match the original if the the original only contain frequencies less than half the sample rate. *Aliasing* occurs if the reconstructed signal contains some of the wrong frequencies, either at the time of sampling an input because the original had some frequencies too high for the sampling rate, or at the time of reconstructing a continuous-time output because our reconstruction process is not perfect in choosing only the lowest of the possible frequencies.

Aliasing is not limited to signals of the form described, but is most often examined in terms of frequencies. For real signals, any frequency f includes a positive-frequency component at $z = \exp(i2\pi fT)$ and a negative-frequency component at $z = \exp(-i2\pi fT) = \exp(i2\pi(f_s - f)T)$. The circularity in z and the pairing of complex conjugates in real systems means that a sampled system will not distinguish between the frequencies f and $f_s - f$ (or more generally, between f and any frequency displaced up or down by f from any multiple of f_s). A pair of such continuous-time sinusoids can lead to identical sample sequences, as shown in Figure 7.6.

For example, consider sinusoids of frequencies 2.5 kHz and 7.5 kHz ($s = \pm i2\pi 2500$ and $s = \pm i2\pi 7500$):

$$x_{2.5}(t) = \cos(2\pi 2500t)$$

$$x_{7.5}(t) = \cos(2\pi 7500t)$$

Now sample both of these at 10 kHz. Then $sT = \pm i\pi/2$ and $sT = \pm i3\pi/2$, respectively, and $z = \exp(sT) = \pm i$ for both of them. In 10^{-4} s, the phase $2\pi 2500t$ changes by $\pi/2$ or 90 degrees, while the other changes by $3\pi/2$ or 270 degrees, to give the sample sequences:

$$x_{2.5}[k] = \cos(k\pi/2) = [1, 0, -1, 0, 1, \ldots]$$

$$x_{7.5}[k] = \cos(3k\pi/2) = [1, 0, -1, 0, 1, \ldots]$$

Observe that these two sinusoids are aliases of each other when sampled at this 10 kHz rate, which is high enough to unambiguously represent only frequencies below 5 kHz.

The Nyquist criterion—first clearly expressed by Shannon in the form quoted above, but anticipated by others including Nyquist, Küpfmüller, Whittaker, Kotelnikov, Ogura, and Raabe (Meijering, 2002)—can be interpreted from the point of view of either the signals or the systems. Given a family of signals of bandwidth W, the *Nyquist rate for this family of signals* is $2W$, the lower bound on sampling rates that will unambiguously represent those signals by their samples. Alternatively, given a system with sample rate f_s, the *Nyquist frequency of the system* is $f_s/2$, the upper bound on signal frequencies that the system can unambiguously represent (that is, without aliasing to signals of lower frequencies).

We typically provide a good margin of safety between twice the highest frequency of interest and the sample rate that we use to process sounds, mainly because it is hard to guarantee that there is no power present at frequencies higher than what we are interested in. A lowpass anti-aliasing filter is typically used before sampling a continuous-time sound waveform, providing a gradual cutoff between the highest frequency of interest and the frequencies that will alias.

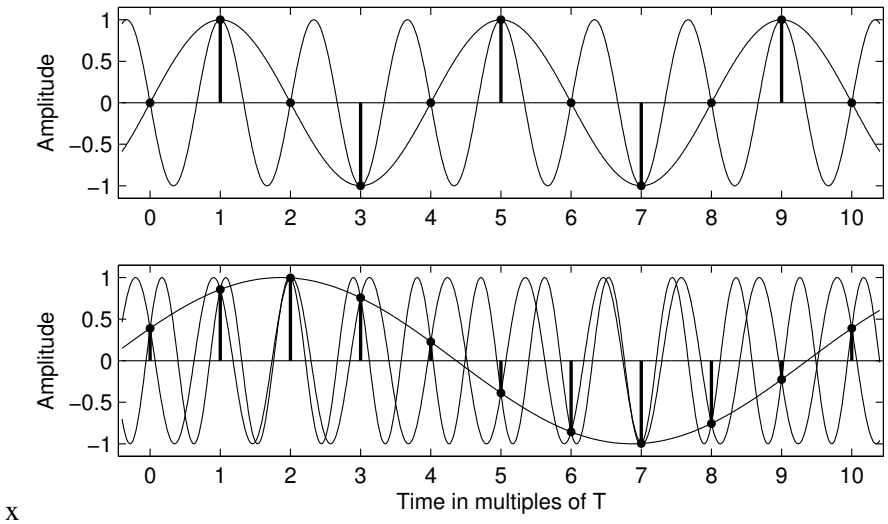


Figure 7.6: Sampled sine waves with frequencies 0.25 and 0.75 times the sampling frequency (top) can give identical sample sequences (dots), as discussed in the text. In light of their identical samples, the signals shown are *aliases* of each other. Another example, frequencies 0.1, 0.9, and 1.1 cycles per sample (bottom), shows how a low frequency has aliases slightly above and below the sampling frequency.

7.9 Mappings from Continuous-Time Systems

For our example first-order smoothing filter, we have made the digital equivalent by mapping the pole position from $s_p = -1/\tau$ in the s plane to z_p in the z plane, a point that just depends on the ratio of the filter time constant to the sampling interval:

$$z_p = \exp(s_p T) = \exp\left(\frac{-T}{\tau}\right)$$

For the RC filter with $\tau = 0.5$, if we use a sample interval $T = 0.1$ (that is, we sample at 10 Hz), the z-plane pole is then at the position illustrated in Figure 7.3:

$$z_p = \exp(-0.2) = 0.819$$

The feedback coefficient a is therefore $a = 0.819$, and we need $b = 1 - a = 0.181$ to make the DC gain come out to 1—to see this, plug in $z = \exp(i0T) = 1$ for the frequency point to evaluate the DC gain of the transfer function as $b/(1 - a)$.

The digital smoother's impulse response, the sequence $[b, ba, ba^2, ba^3, \ldots]$, shown in Figure 7.7, is a sampled version of the impulse response of the continuous-time smoothing filter that it derives from, scaled to maintain the same DC gain of 1:

$$\begin{aligned} h[k] &= bz_p^k = ba^k \\ &= \left(1 - \exp\left(\frac{-T}{\tau}\right)\right)\exp\left(\frac{-kT}{\tau}\right) \\ &= 0.181 \cdot 0.819^k \end{aligned}$$

Matching pole and zero positions via $z = \exp(sT)$ is a common way to convert continuous-time filters to nearly equivalent discrete-time filters. This method is known as the *pole–zero mapping* (Yang, 2009),

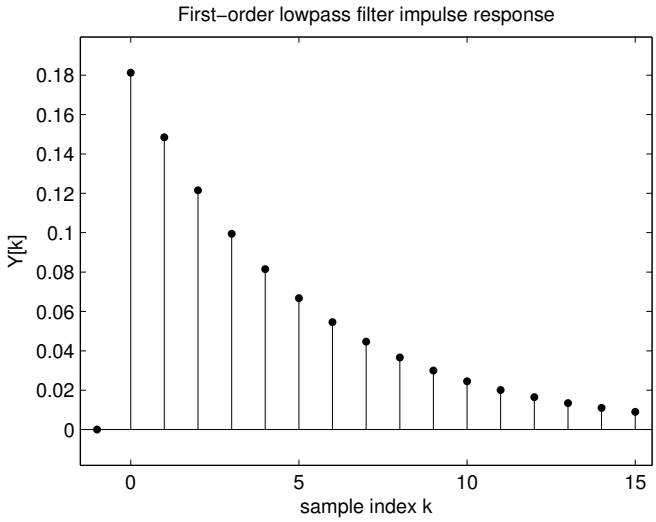


Figure 7.7: The discrete-time first-order smoothing filter's impulse response is a geometric sequence (for $k \geq 0$), a sequence of samples of an exponential decay.

pole–zero matching (Johnson, 1997), *pole mapping* (Cooke, 1993), or *matched Z transform* (Narasimhan and Veena, 2005) method.

Each pole corresponds to one term in the homogeneous response of the system; matching from continuous-time to discrete-time using $z = \exp(sT)$ makes the term in the discrete-time response equal to a sampled version of the term in the continuous-time response, but does not determine how the gain factors of the various terms will be set in the impulse response. Also matching the zeros through the same mapping will determine the relative gain factors. This leaves an overall free gain parameter, which is typically set by a constraint at zero frequency, as above, or at some other frequency of interest. To minimize the phase lag, the output is advanced as much as possible while keeping it causal; that is, enough zeros are added at $z = 0$ to bring the total number of zeros up to the number of poles.

When the continuous-time filter has no zeros, the impulse response of the discrete-time filter made this way will be a sampled version of the continuous-time filter's impulse response (or nearly so, within a gain adjustment factor or a shift). This property, known as *impulse invariance*, defines another approach (alternative to the pole–zero mapping method) for converting a filter from continuous to discrete time. When zeros away from $z = 0$ are present, pole–zero mapping does not yield impulse invariance, but is nevertheless a suitable method for our purposes. Defining the discrete-time filter by sampling the impulse response of the continuous-time filter would yield the same poles, but somewhat different zeros in such cases.

Other methods of converting between continuous-time and discrete-time systems exist, but we don't use them. We usually use pole–zero mapping, because it is simple; sometimes it is equivalent to impulse invariance.

7.10 Filter Design

In courses on signal processing, engineers typically learn techniques to design filters for applications such as separating the channels of frequency-division-multiplex communication systems, for example in telephone and radio systems. Each filter is designed to be close to an ideal goal of flat frequency response in a specified passband and near-zero response in specified stopbands, separated by very sharp transitions. Certain standard filter types that come closest to the ideal by some criterion—Butterworth, Chebyshev, elliptic, and Bessel

filters, for example—are taught as building blocks for system design (Oppenheim and Schafer, 2009).

However, nothing in the auditory system comes close to a flat passband or an abrupt cutoff. Much of what is taught in these courses, beyond the basics we just covered, is irrelevant to modeling hearing. Instead, we need to be flexible about exploring, by *ad hoc* techniques, how different structures, or pole–zero configurations, make filters that approximate what is going on in hearing. Subsequent chapters provide examples of such exploration.

We will consider simple arrangements of poles and zeros, starting with continuous-time systems, to model the filtering that the cochlea does. Then, as we did for the first-order example discussed above, we'll map those s-plane poles and zeros to z-plane poles and zeros for a chosen sample rate, and convert those pole–zero descriptions to digital filter coefficients as described in the next section.

7.11 Digital Filters

We need to know what working linear systems look like in software, not just how to describe them by impulse responses, frequency responses, and such. Fortunately, the software to run digital filters is relatively simple, compared to the design and analysis processes.

Converting a continuous-time linear system description to an approximately equivalent discrete-time system description is usually straightforward. By whatever method, such as pole–zero mapping from a continuous-time design, the linear system can be put into the form of a difference equation, such as the one we showed for the first-order lowpass filter:

$$y[k] = bx[k] - ay[k-1]$$

In general, the difference equation will specify a set of forward coefficients as weights on previous inputs, and a set of feedback coefficients as weights on previous outputs; these coefficients in the implementation are the same as the coefficients in the rational transfer function numerator and denominator. The software is then simple: each new sample of the output is a weighted sum of values that have already arrived or already been computed and are sitting in memory. The typical general formula for computing the next term of the output is:

$$y[k] = \sum_{n=0}^{N} b_n x[k-n] - \sum_{n=1}^{N} a_n y[k-n]$$

where N is the order of the filter.

The first part of this formula looks like a convolution, if the forward coefficients correspond to an impulse response as $b_k = h[k]$. When the feedback coefficients are all zero, this convolution is all there is. Such nonrecursive filters are also called *finite impulse response* (FIR) filters, since the response to an impulse is at most $N + 1$ nonzero output samples. FIR filters have zeros, but no poles (or poles only at $z = 0$, to add enough delay to make them causal).

Our first-order recursive lowpass filter does not have a finite impulse response—its exponential impulse response takes forever (in theory) to get back to zero, so we need some nonzero feedback coefficients to implement it (only one in this case—that's what first-order means). Such recursive or IIR filters, with exponentially decaying impulse responses, are what we need to model real-world systems such as the cochlea. IIR filters have poles, and often also zeros.

IIR filters have a number of well-known standard "forms," one of which is the *direct form I* shown in Figure 7.5, and another the *direct form II*, illustrated in Figure 7.8; both are illustrated for order 2. The latter is *canonical with respect to delay*, meaning that it only requires two delay elements, or state variables, to achieve any second-order response. It is also canonical with respect to multiplications, since the general

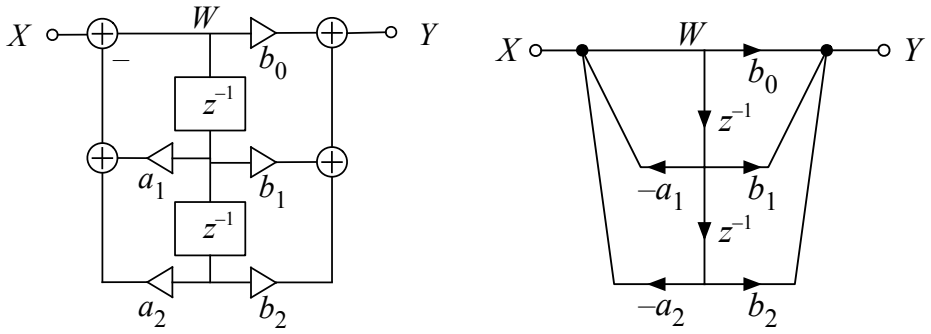


Figure 7.8: In a typical representation of a direct form II discrete-time or digital second-order section (left), the feedback block that makes the poles comes first, followed by the feed-forward block that makes the zeros. These blocks share two delay elements, delaying an intermediate signal, neither X nor Y, which is the output of the first block and the input of the second block. The directions of signal flow are implied by the coefficient multipliers (the oriented triangles), so the generous use of arrows that we saw in Figure 7.5 is typically avoided. In the even more concise form (right) arrows on lines are used to indicate multiplicative operators, including the delays, and dots represent additions.

second-order filter requires just five multiplications per time step—the five coefficients are needed to specify the two degrees of freedom for the pole pair, two for the zero pair, and one for the gain.

The illustrated forms can be extended to work for any order. However, a popular alternative way to realize higher-order filters is to factor them into cascades of *second-order sections*, each section being responsible for just one pair of poles and one pair of zeros. Historically, this factoring was popular both for numerical stability and for modularity (Karam et al., 1999).

The second-order section is often presented in terms of the noncanonical direct form I of Figure 7.5, in which there are four delay elements (two samples of x and two samples of y are stored). Using the same coefficients in the direct form II, shown in Figure 7.8, gives the same response, as can be seen by interpreting each form as two filters—one with the feedback a_i coefficients making the poles and one with the forward b_i coefficients making the zeros—cascaded in opposite order.

To write difference equations for the direct from II, we start by naming the signal being delayed; call it W:

$$w[k] = x[k] - a_1 w[k-1] - a_2 w[k-2]$$

$$y[k] = b_0 w[k] + b_1 w[k-1] + b_2 w[k-2]$$

Convert the difference equations to the frequency domain using operator notation, solve for the input–output ratio, and express as rational transfer functions:

$$W(z) = X(z) - (a_1 z^{-1} + a_2 z^{-2}) W(z)$$

$$\frac{W(z)}{X(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} = \frac{z^2}{z^2 + a_1 z + a_2}$$

$$Y(z) = (b_0 + b_1 z^{-1} + b_2 z^{-2}) W(z)$$

$$\frac{Y(z)}{W(z)} = b_0 + b_1 z^{-1} + b_2 z^{-2} = \frac{b_0 z^2 + b_1 z + b_2}{z^2}$$

Here the z^2 numerator makes a pair of zeros that keep W/X minimum phase, and the z^2 denominator makes

a pair of poles that keeps Y/W causal. When these transfer functions are expressed as rational functions of z^{-1} instead of z, these roots at the origin move out to infinity and disappear, as the expressions show, but it is more conventional to use functions of z.

Finally, eliminate W:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{W(z)}{X(z)}\frac{Y(z)}{W(z)} = \frac{b_0 z^2 + b_1 z + b_2}{z^2 + a_1 z + a_2}$$

Though we often prefer the direct form II in implementations (to avoid needing to access two sets of delayed values), writing the difference equation from the direct form I (Figure 7.5) avoids the introduction of a new signal name and makes the transfer function calculation more streamlined:

$$y[k] = b_0 x[k] + b_1 x[k-1] + b_2 x[k-2] - a_1 y[k-1] - a_2 y[k-2]$$

$$Y(z) = (b_0 + b_1 z^{-1} + b_2 z^{-2})X(z) - (a_1 z^{-1} + a_2 z^{-2})Y(z)$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0 z^2 + b_1 z + b_2}{z^2 + a_1 z + a_2}$$

One can now appreciate why the feedback coefficients are shown in the signal-flow diagram with negative signs: to keep the frequency-domain denominator polynomial in standard form $z^2 + a_1 z + a_2$.

It's easy to see how to move between the transfer function and the hardware or software implementation, since the same coefficients appear in both. To get to either the transfer function or the difference equation from specified poles and zeros, just multiply out the factored form ($(z - p_1)(z - p_2)$, for example) to get the polynomial coefficients; as long as the poles and zeros are either real or in complex-conjugate pairs, the resulting coefficients will be real. For example, given a conjugate pair of zeros z_1 and z_1^*, and a conjugate pair of poles p_1 and p_1^*, and a gain factor A, we have:

$$\begin{aligned} H(z) &= A\frac{(z - z_1)(z - z_1^*)}{(z - p_1)(z - p_1^*)} \\ &= A\frac{z^2 - 2\mathrm{Re}[z_1]z + |z_1|^2}{z^2 - 2\mathrm{Re}[p_1]z + |p_1|^2} \end{aligned}$$

The latter form obviously has real coefficients. Therefore, the implementation in terms of the difference equation is very simple, given complex-conjugate-pair pole and zero locations and an overall gain factor; just use these coefficients:

$$\begin{aligned} b_0 &= A \\ b_1 &= -2A\mathrm{Re}[z_1] \\ b_2 &= A|z_1|^2 \\ a_1 &= -2\mathrm{Re}[p_1] \\ a_2 &= |p_1|^2 \end{aligned}$$

By convention, the leading coefficient of the denominator polynomial is $a_0 = 1$; there is no place in the second-order section implementation where a different value of a_0 could be incorporated, though it's easy to use an arbitrary value for b_0, as shown, which is how the overall gain is set. When we want unity gain at DC (at $\omega = 0$), the factor A is set to the reciprocal of the gain that we get by setting $z = \exp(i0T) = 1$ and

evaluating the ratio of polynomials; that is:

$$A = \frac{1 - 2\mathrm{Re}[p_1] + |p_1|^2}{1 - 2\mathrm{Re}[z_1] + |z_1|^2}$$

7.12 Multiple Inputs and Outputs

Linear systems with multiple inputs and outputs do not introduce any difficulty. If a linear system has multiple inputs, it has a transfer function from each input to the output; it is equivalent to separate linear systems with those transfer functions, with their outputs added. A system with multiple outputs is equivalent to a set of independent systems. We commonly use the latter concept as a *filterbank*: a *bank* of filters operating in parallel on a common input. In models of hearing, each *channel* in the bank of filters is a model of the response of one place in the cochlea, or one *center frequency* in a frequency analyzer.

Terminology such as *filterbank* and *channel bank* evolved in the telephone industry, where it was common to set up "banks" of equipment in parallel to handle many voice channels.

It is also common to arrange a sequence of linear filters *in cascade*: the output of each filter being the input to the next. If the intermediate outputs are also taken to be a set of multiple outputs of the system, we have what we call a *cascade filterbank*, as opposed to a *parallel filterbank* of independent filters. This cascade filterbank is still equivalent to a set of independent filters with a common input, but as we'll see in Section 14.7, the cascade form offers significant conceptual and practical advantages when modeling the cochlea.

7.13 Fourier Analysis and Spectrograms

It is common in engineering to represent signals in the frequency domain by way of a Fourier spectrum, essentially saying how much, and sometimes what phase, of each frequency of sinusoid needs to be added up to make a signal. There are various *spectral estimation* techniques to characterize a signal by a spectrum, generally ignoring phase. One such technique is to use a filterbank: a bank of bandpass filters, with power (or energy) detectors at their outputs, followed by smoothing filters, will estimate the short-time power in each filter's frequency band.

Since filterbanks can be spectrum analyzers, the action of the cochlea as a filterbank is sometimes characterized as a Fourier analysis, a transformation to the frequency domain. When we study hearing, we need to be careful about when to use Fourier techniques and when not to. In this section, we point out some of what the Fourier transform is good for, and some of what it's not so good for.

As we showed above, a linear filter's frequency response is the Fourier transform of its impulse response. In that respect, the Fourier frequency response (including phase) is a complete description of a causal stable filter. However, to build an executable model of a linear system, we need to know the poles and zeros—or approximate it with poles and zeros—to get the coefficients, so we really need a description in the Laplace or Z domain. Getting there from the Fourier domain, especially from the Fourier-magnitude domain with no phase specification, is not always easy. When the underlying system can be described as a lumped circuit, or as a finite set of poles and zeros, or as a digital filter, or as a rational transfer function in the Laplace domain, that pole–zero description is generally much more concise than a Fourier-domain description.

Fourier showed that a periodic function can be described as a sum of discrete sinusoids; this sum is called a Fourier series, as opposed to a Fourier transform. This approach can be useful for describing periodic sound signals, for example the steady note of a musical instrument such as a woodwind, brass, or bowed string, or a steady spoken vowel. Such a description requires, however, that the sound be periodic, and that the period be

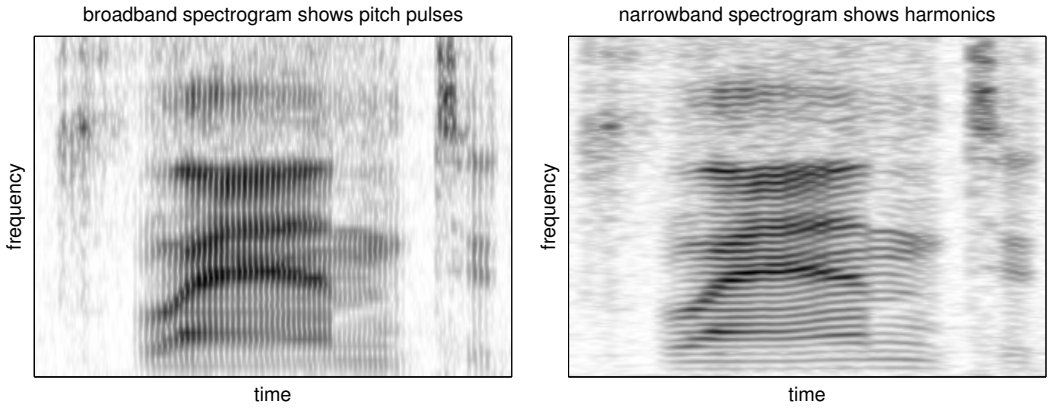


Figure 7.9: A wideband spectrogram (left) and a narrowband spectrogram (right) show the short-time power levels out of a bank of bandpass digital filters. When the filter bands are wide enough, the temporal response is fast enough to show individual glottal pulses of speech, as on the left. Conversely, when the filter bands are narrow enough, they can resolve individual harmonics of the pitch frequency, as on the right. But neither version captures the much finer temporal structure, the phase information or individual waveform peaks that the auditory nerve can represent. The spoken words analyzed here are "plan to."

known, so it doesn't extend well to changing sounds, noisy sounds, transient or percussive sounds, mixtures of sounds, etc.

The Fourier transform, as opposed to the Fourier series, is good for describing transient signals as integrals of a dense infinity of sinusoids; impulse responses are amenable to such description. To apply this technique to general sounds, however, finite short pieces of the sound signal need to be extracted, and analyzed in turn. The short-time Fourier transform (STFT) technique applies this kind of Fourier analysis to successive short pieces of a signal (the pieces are usually tapered, or *windowed*, at both ends, to reduce the spectral "splatter" that chopping the signal would otherwise cause). Long windows give good frequency resolution but poor time resolution, while short windows give good time resolution but poor frequency resolution (this is an inherent limitation of time–frequency analysis, not of the STFT itself).

Whether from a STFT technique or from a filterbank followed by square-law detectors and smoothing filters, the resulting signal descriptions in terms of power as a function of time and frequency are known as *spectrograms*. Speech signals are often represented in pictures as one of two types of spectrogram, as illustrated in Figure 7.9: wideband spectrograms with good time resolution and poor frequency resolution, which resolve voice pitch pulses in time; and narrowband spectrograms with good frequency resolution and poor time resolution, which resolve pitch harmonics in frequency. With respect to hearing, however, there is no right or ideal size for the short-time segments, or for the frequency and time resolutions of filterbanks used to describe signals as sequences of short-time spectra. Neither of these spectrogram types has much to do with how hearing works, since the auditory nerve follows the actual fine time waveforms of filtered sound signals, rather than just telling the brain about a smoothed power estimate. For machine hearing applications, we will benefit by respecting the auditory nerve, and preserving the fine time structure in the filtered channels, rather than treating the ear as a spectrum estimator.

For traditional linear-frequency-scale spectrograms and other applications with equally spaced equal-bandwidth channels, short-time Fourier techniques are appropriate and efficient; but the nearly equivalent time-domain bandpass filter technique is more flexible in terms of spacing, unequal bandwidths, full time-domain outputs, and ability to integrate useful nonlinearities.

7.14 Perspective and Further Reading

Discrete-time filtering and spectrum analysis methods are the cornerstones of modern media processing. For example, analyzing music for compression as MP3 files requires sophisticated algorithms whose fundamentals are the topic of this chapter. Besides their application to analysis tasks such as machine hearing, these methods are also at the core of speech and music synthesis, and many other fields including, via extension to multiple dimensions, image and video processing.

Digital filters are how we implement linear systems in computers. The theory of discrete-time linear shift-invariant systems is completely parallel to the theory of continuous-time linear time-invariant systems. Though the theory is based on transforms, all the needed design calculations can usually be done with nothing more complicated than algebra with complex numbers. The resulting digital filter implementations generally need only real-number arithmetic.

Linear systems are at the core of nonlinear signal processing, too. Estimating power at a filter output is a nonlinear operation, as are things like modulation and demodulation used in radio and other systems. But the nonlinearities are simple, and the system performance is usually determined by the linear filters surrounding the simple nonlinear operators. We look at nonlinear systems in Chapter 10.

In the next chapter, we look more at *resonant* linear systems, building toward models of filtering in the cochlea.

There are many great books available for more detail and depth on digital and discrete-time signal processing, and their application to audio; my favorites include Smith (2007), Oppenheim and Schafer (2009), and Gold, Morgan, and Ellis (2011).

Chapter 8

Resonators

> Another experiment should be adduced. Raise the dampers of a pianoforte so that all the strings can vibrate freely, then sing the vowel *a* in *father, art,* loudly to any note on the piano, directing the voice to the sounding-board of the piano; the sympathetic resonance of the strings distinctly re-echos the same *a*. On singing *oe* in *toe*, the same *oe* is re-echoed.
>
> — *On the Sensations of Tone*, Hermann Ludwig F. Helmholtz (1863)

> In the classes of circuits discussed in the following chapters, the pole-zero patterns show at a glance the general form of the frequency characteristics with the important features placed clearly in evidence; they display the effects of varying the circuit parameters; and they reveal the key approximations that permit certain groups of complex circuits to be treated as equivalent circuits of less complexity.
>
> — *Pole-Zero Patterns: In the Analysis and Design of Low-order Systems*, Angelo and Papoulis (1964)

8.1 Bandpass Filters

Auditory filtering in the cochlea is generally conceptualized as *bandpass* filtering, with a dense array of filters representing up to thousands of locations along the cochlear partition. A bandpass filter is a system that responds strongly to frequency components within its *passband*, and only weakly to signals of other frequencies. This concept came up in the context of critical bands, and in the context of spectrum analysis; now we proceed to the mathematical description of the concept, building on what we learned about general linear systems in previous chapters.

We often refer to the *peak* of a bandpass filter; the *peak frequency* (or center frequency) and *peak gain* describe the point of highest gain, while the *peak width* and *peak shape* describe the frequency response of the *passband*, the high part of the frequency response close to that point. The parts of the frequency response further from the peak are often described as the *skirt* or *tails* of the bandpass response—often *skirt* if it drops rapidly, as a transition to a *stopband*, a range of frequencies in which the filter attenuates strongly, and *tail* if the frequency response levels off or drops slowly. This *tail* terminology appears to be unique to the hearing field, in which the low-frequency tails of tuning curves have long been recognized as functionally important (Kiang and Moxon, 1974).

Bandpass filters have various shapes, which may be symmetric or asymmetric with respect to frequencies above and below the peak frequency. We often characterize a bandpass filter's frequency response by the

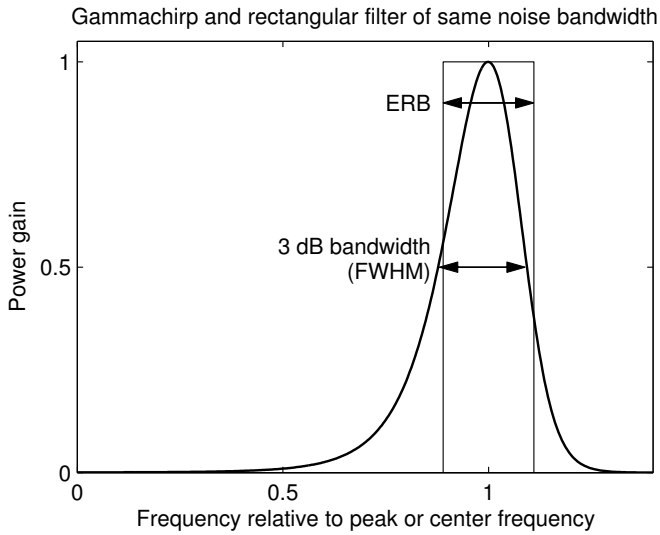


Figure 8.1: The power response (the square of the magnitude response) of an asymmetric bandpass filter, and of a rectangular filter that passes the same total noise power when the input is a white noise (that is, with the same area under the curve). The equivalent rectangular bandwidth (ERB, also known as equivalent noise bandwidth, ENB) of the asymmetric filter is the width of the rectangular filter shown. The ERB of bandpass filters that we encounter in hearing, such as this *gammachirp* filter, is typically slightly greater than the 3-dB (half-power) bandwidth (also known as the full width at half maximum, FWHM), depending on the filter shape. Both the ERB and the FWHM are popular characterizations of a filter's bandwidth, which, as shown here, are not generally equal.

square of its gain magnitude, or its *power frequency response*, defined as:

$$P(\omega) = |H(\omega)|^2$$

In particular, we summarize a filter by the *width* of its peak, by one of several measures, as illustrated in Figure 8.1. We commonly use the *3-dB bandwidth*: the width of the peak at a gain magnitude of 0.707, or −3 dB, relative to the peak. These limits are known as the half-power points, since power is proportional to the square of amplitude; this bandwidth is what physicists call the *full width at half maximum* (FWHM).

The total noise power in the output of a filter, when the input is a noise, is the integral of the product of the filter's power frequency response times the *noise power spectral density* $N(\omega)$:

$$P_{\text{total}} = \int_0^\infty P(\omega)\, N(\omega)\, d\omega$$

A simple parameter often used to describe a filter is its *equivalent noise bandwidth* or *equivalent rectangular bandwidth* (ERB), the bandwidth of a flat-topped rectangular bandpass filter of the same peak gain that passes the same total power when the input is a *white* (flat-spectrum) noise.

$$\text{ERB} = \frac{\int_0^\infty P(\omega)\, d\omega}{\max(P(\omega))}$$

The ratio of center frequency to bandwidth of a bandpass filter is conventionally called the Q, or *quality factor*, of the filter. Different bandwidth definitions lead to different Q values; variants such as $Q_{3\text{dB}}$, Q_{ERB}, and $Q_{10\text{dB}}$ are sometimes used in the hearing literature.

In many engineering applications, such as radio and other frequency-division-multiplexing applications, filters are designed to be close to rectangular filters, with flat tops and sharp transitions from the passband to the stopbands (steep skirts). In hearing, we don't find such rectangular filters, but rather more bell-shaped and asymmetric filters. These filters can be characterized by their center frequency and bandwidth, plus other shape parameters, just as probability distribution functions are characterized by mean, variance, skew (asymmetry), and kurtosis (tail weight). In fact, we end up with some of the same functional forms that statisticians use—the Pearson distributions (Elderton and Johnson, 1969).

In this chapter, we explore the *resonator*, the simplest physical system or circuit that has a bandpass-like response and can be symmetric or asymmetric. Helmholtz and others analyzed sounds in the nineteenth century using physical resonators such as those shown in Figure 8.2, and we use resonators for similar purposes. The development of an understanding of resonances and filters based on them is key to our later development of nonlinear hearing models that incorporate level dependence via movement of their poles and zeros. Poles and zeros are also key to efficient digital implementation; filters that don't have low-order rational transfer functions are computationally difficult. Therefore, we emphasize understanding resonators in terms of the connections between the pole–zero view and the frequency-response view, plus the connections to the impulse-response view that is prominent in the hearing literature.

Two-pole–two-zero filter stages, or *second-order sections* are also important because they are the natural and popular building blocks for use in implementing arbitrary linear systems as cascades of simpler sections. First-order sections with real coefficients are limited to implementing real poles and zeros. Moving to second order, or introducing resonance, provides the generality needed to implement all rational transfer functions with real coefficients, for both continuous and discrete time.

EE Connections: Alternative Resonant Circuits

We analyzed a simple two-pole resonant circuit in Chapter 6 (see Figure 6.9). Three other circuits with the same three components in series (that is, connected such that the same current goes through all of them) are shown in Figure 8.3.

Two of these new circuits, filters B and C, are of the same form as filter A, namely the generalized voltage divider form shown in Figure 6.8. Since they have the same sum of impedances (the denominator of the voltage-divider equation), we only need to update the Z2 impedances (the numerators of the voltage-divider equation) to get their transfer functions:

$$H_{\mathrm{A}}(s) = \frac{1/sC}{sL + R + 1/sC} = \frac{1}{s^2LC + sRC + 1}$$

$$H_{\mathrm{B}}(s) = \frac{R + 1/sC}{sL + R + 1/sC} = \frac{sRC + 1}{s^2LC + sRC + 1}$$

$$H_{\mathrm{C}}(s) = \frac{R}{sL + R + 1/sC} = \frac{sRC}{s^2LC + sRC + 1}$$

The first two, filters A and B, have unity gain at DC (at $s = 0$); the average output voltage will be equal to the average input voltage. The third is *AC coupled*, that is, with zero gain at DC (a zero at $s = 0$), due to the capacitor that blocks any steady current from input to output; its average output will be zero. All three filter transfer functions have the same poles, since they have identical denominators, so they have identical homogeneous solutions, including identical dynamics of energy decay when the input is not being driven. The fourth, filter D, also has the same poles and homogeneous solutions, since it is a parallel combination of filter A and a *straight-through path* without poles; we come back to this one in Section 8.6.

Using the quadratic formula to write the poles, the roots of the denominator, gives:

$$p_1, p_2 = \frac{-RC \pm \sqrt{R^2C^2 - 4LC}}{2LC}$$

We can factor filter A into a cascade of two one-pole filters having these two poles. This cascade is a pair of RC filters if the poles are real. But if the poles are complex, then such factors in isolation do not correspond to real (real-valued or realizable) systems or circuits. This is the interesting case here, as it represents resonance. Real resonant circuits have such poles in complex-conjugate pairs, as the quadratic formula suggests, so it takes a second-order filter to represent a real circuit with resonance.

Suppose we vary the resistance R when L and C are fixed, keeping R small enough that the poles are a complex-conjugate pair (that is, the quantity $R^2C^2 - 4LC$ under the square root is negative). In that case, the formula represents points on a circle of radius $1/\sqrt{LC}$ in the complex s plane, as shown in Figure 8.4. We call this radius the *natural frequency* ω_{N} of the circuit. The resistance R determines where the poles are on the circle of that radius, by setting their real part to $-\gamma = -R/(2L)$. The pair of poles can be parameterized in several ways, including these two based on natural frequency and either *decay rate* γ (gamma) or *damping factor* ζ (zeta):

$$p_1, p_1^* = -\gamma \pm i\sqrt{\omega_{\mathrm{N}}^2 - \gamma^2} = \omega_{\mathrm{N}}\left(-\zeta \pm i\sqrt{1 - \zeta^2}\right)$$

The damping factor is a nondimensional measure (between 0 and 1 for the resonant case) of how "lossy" the system is, or of the rate at which the stored energy is dissipated, relative to the system's natural frequency:

$$\zeta = \frac{\gamma}{\omega_{\mathrm{N}}} = \frac{R}{2}\sqrt{\frac{C}{L}}$$

The amplitude of resonance decreases by a factor of e in the time $1/\gamma$, corresponding to $1/\zeta$ radians of oscillation at frequency ω_{N}; the stored energy (amplitude squared) decays by a factor of e in half that time.

Resonators of Helmholtz.—The most ready way of analyzing a complex sound is that suggested by Helmholtz. He employed a series of hollow brass or glass spheres, each having a circular opening *a* to admit the vibrations of the outer air to the air in the interior of the sphere. Opposite this opening is a nipple *b*, which fits in the ear, and thus conveys the vibrations to the auditory nerves.

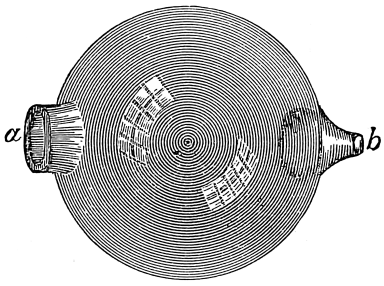


FIG. 292.—HELMHOLTZ RESONATOR.

Figure 8.2: Before we had electrical resonators, fluid-mechanical *Helmholtz resonators* were used for sound analysis. Helmholtz developed and used a set of these resonators, tuned to the frequencies of musical notes, to help him "hear out" the sinusoidal components of complex tones. These resonators are tuned by the interaction of the springiness of the air in the globe with the momentum of the mass of air in the neck. Figure from Quackenbos et al. (1891).

8.2 Four Resonant Systems

In the box "EE Connections: Alternative Resonant Circuits" and Figure 8.3, we have described four resonant or bandpass systems in terms of circuits, but from here on we work with them as abstract linear systems ("resistors are futile," I've been told). The transfer functions of the four filters differ only in their numerators, meaning they have different zeros, but the same pair of poles.

For two-pole resonant systems we typically parameterize resonators by a *natural frequency* ω_N and a nondimensional *damping factor* ζ (zeta). The s-plane pole positions in terms of these parameters, as shown in Figure 8.4, are:

$$p_1, p_2 = \omega_N \left(-\zeta \pm i\sqrt{1-\zeta^2} \right)$$

which are complex values as long as the absolute value of the damping factor ζ is less than 1, at the locations shown in Figure 8.4, as pointed out in the EE connections box.

Notice that the magnitude of the factor on the right is 1, so the poles are at a distance ω_N from the origin in the s plane. Varying ζ traces out two quarter circles, with $\zeta = 0$ mapping to poles on the imaginary axis, and $\zeta = 1$ mapping to coincident poles on the negative real axis. The damping factors we typically use in hearing are intermediate values, from about 0.1 to at most 0.4. Other fields sometimes involve resonators with very low damping factors; certain popular filter designs use higher damping factors (for example, second-order Butterworth lowpass filters use damping factors of 0.707, but such a system barely resonates). Negative damping factors, pushing the poles into the right half of the s plane, are encountered in unstable systems.

In terms of the parameters damping factor ζ and natural frequency ω_N, the transfer functions of the four filters in Figure 8.3 are:

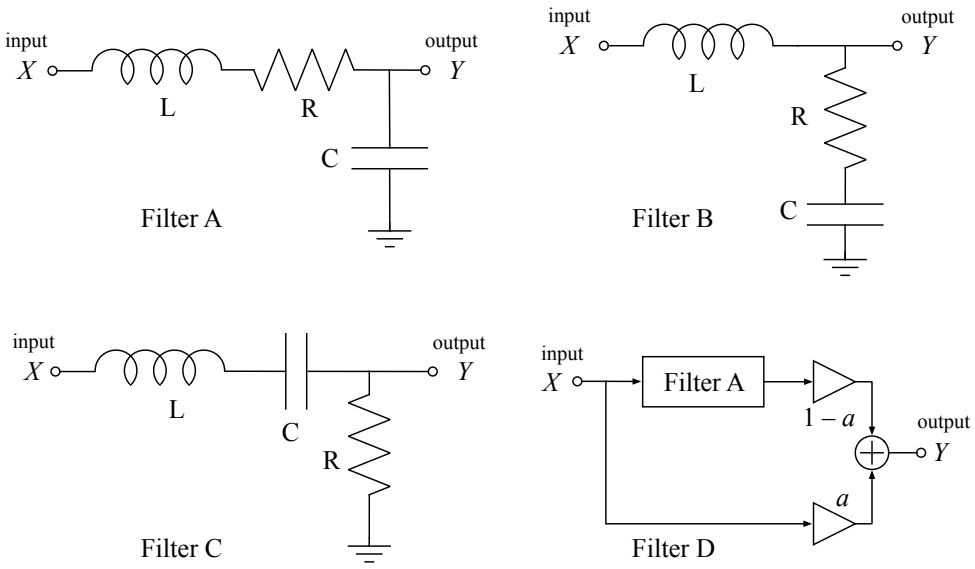


Figure 8.3: Circuit diagrams of four resonant filters. Filter A (top left) is the same as previously shown in Figure 6.9. Filter B (top right) has the resistor, the energy-dissipating element, moved from the series impedance (the impedance of the elements connecting the input to the output in the voltage divider circuit) into the shunt impedance (the impedance of the elements connecting the output to ground). Like Filter A, Filter B has unity gain at DC, since the DC impedance of the capacitor in the shunt leg is infinite. Filter C (bottom left) is a second-order filter with capacitive coupling, or zero response at DC (at zero frequency), since the capacitor is now in series. Filter D (bottom right) uses a pair of adjustable-gain buffer amplifiers (shown as triangles) to mix the output of a filter A circuit with its input. All of these filters have the same pair of poles, and all are relevant as basic building blocks and limiting cases in our study of auditory filters in subsequent chapters.

$$H_{\mathrm{A}}(s) = \frac{1}{(s/\omega_{\mathrm{N}})^2 + 2\zeta s/\omega_{\mathrm{N}} + 1}$$

$$H_{\mathrm{B}}(s) = \frac{2\zeta s/\omega_{\mathrm{N}} + 1}{(s/\omega_{\mathrm{N}})^2 + 2\zeta s/\omega_{\mathrm{N}} + 1}$$

$$H_{\mathrm{C}}(s) = \frac{2\zeta s/\omega_{\mathrm{N}}}{(s/\omega_{\mathrm{N}})^2 + 2\zeta s/\omega_{\mathrm{N}} + 1}$$

$$H_{\mathrm{D}}(s) = \frac{a(s/\omega_{\mathrm{N}})^2 + 2a\zeta s/\omega_{\mathrm{N}} + 1}{(s/\omega_{\mathrm{N}})^2 + 2\zeta s/\omega_{\mathrm{N}} + 1}$$

From the numerator polynomials, it can be seen that filter A has no zeros, filter B has a zero at $s = -\omega_{\mathrm{N}}/(2\zeta)$, filter C has a zero at $s = 0$, and filter D has a quadratic numerator and so has two zeros. These

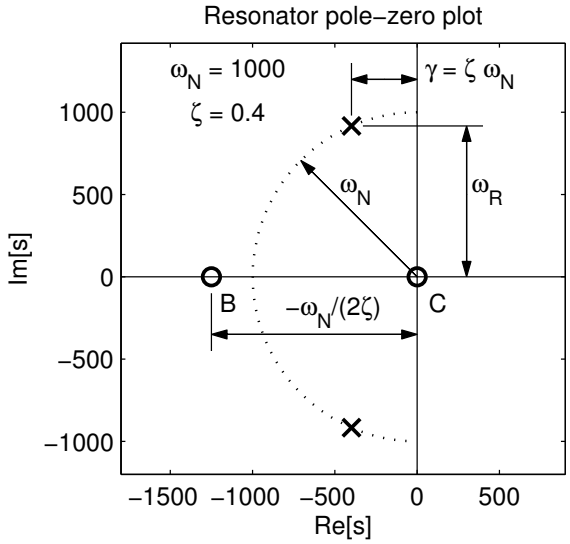


Figure 8.4: The s-plane pole–zero plot for three resonators, filters A, B, and C, illustrated for $\omega_N = 1000$ rad/s and $\zeta = 0.4$. The poles (crosses) at $-400 \pm i\,917$ rad/s (which is $-\gamma \pm i\omega_R$) are the same for the three filters. The dotted semicircle shows the locus of pole positions at radius ω_N for other values of damping factor ζ between 0 and 1: when the damping factor is near zero, the poles are near the imaginary axis, and when it is near 1, the poles approach each other at the negative real axis. Filters B and C each have a zero in addition to the poles, at the positions shown with circles.

complex transfer functions over the s plane are illustrated in Figure 8.5 (except for filter D, which is analyzed in Section 8.6).

In the log–log frequency response, or Bode plot, for filters A, B, and C, in Figure 8.7, the different asymptotic slopes at high and low frequencies are noted. A direct or inverse proportion to s or ω leads to a slope of 6 or −6 dB/octave, respectively. More precisely, the slopes are $20\log_{10}(2) = 6.02$ dB/octave); a quadratic proportionality doubles that slope. In general, the asymptotic slopes of polynomial and rational functions of frequency approach multiples of 6 dB/octave in a Bode plot.

At very high frequencies, only the leading terms (highest powers of s) of the numerator and denominator matter. Therefore, transfer functions B and C approach $2\zeta\omega_N/s$, inversely proportional to frequency at high frequency, or −6 dB/octave. Filter A's transfer function, on the other hand, approaches ω_N^2/s^2, inversely proportional to the square of frequency, or −12 dB/octave.

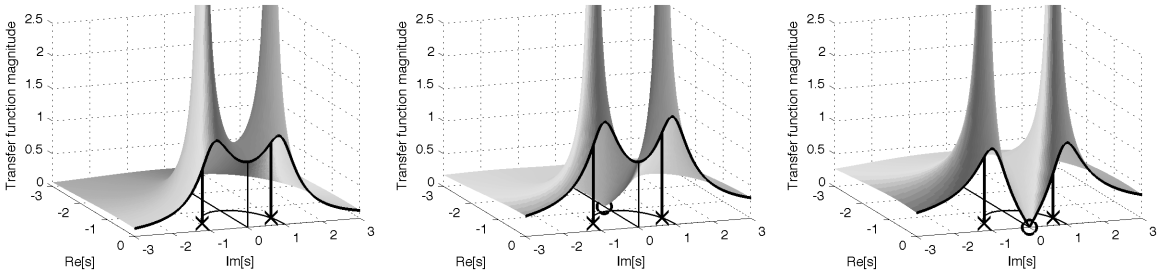


Figure 8.5: The transfer functions of the resonator filters A, B, and C, for natural frequency 1 and damping factor 0.4. See the color plates.

At very low frequencies, only the lowest-order terms matter. Filters A and B are flat (0 dB/octave) at low frequencies, while filter C approaches zero via the factor s, meaning 6 dB/octave.

Between these low-frequency and high-frequency limits, the transfer function can have a high resonant peak, depending largely on the damping factor. The closer the poles are to the imaginary axis, where damping is zero, the higher the peak.

Sometimes the pole positions are represented in Cartesian coordinates as shown in Figure 8.4, with the real part of the pole position representing the decay rate γ:

$$\gamma = \zeta\omega_N$$

and imaginary part representing the *ringing frequency* ω_R:

$$\omega_R = \omega_N\sqrt{1-\zeta^2}$$

These parameters satisfy the relation for a circle of radius ω_N:

$$\omega_N^2 = \gamma^2 + \omega_R^2$$

with the poles on this circle at:

$$p_1, p_1^* = -\gamma \pm i\omega_R$$

We say the filter "rings" when it has been excited, at the ringing frequency ω_R, the imaginary part of the pole position in the s plane, via the homogeneous response $\exp(-\gamma t)\exp(i\omega_R t)$. That is, the output oscillates, corresponding to a decaying sinusoid of frequency ω_R, a tone-like sound analogous to the ringing of a struck bell or tuning fork. The ringing frequency is most useful in the time-domain (impulse-response) description of resonators, as discussed below in Section 8.4.

It is also common to use a quality factor or Q, instead of damping factor, to parameterize the pole positions:

$$Q = \frac{1}{2\zeta}$$

This definition of the pole's Q parameter is approximately consistent with the bandpass-filter uses of Q introduced earlier, as in Section 3.6—approximately the ratio of a simple resonator's center frequency to its half-power bandwidth. At high damping, however, say $\zeta > 0.3$, the relationship to a half-power bandwidth is not precise, and the ability to even specify what is meant by a half-power bandwidth breaks down as the frequency response becomes more like a lowpass than a bandpass, as in the illustrated filter A with $\zeta = 0.4$ in Figure 8.5.

8.3 Resonator Frequency Responses

To get the frequency responses of our filters, we evaluate the transfer functions at $s = i\omega$, as shown along the imaginary-axis cut lines in Figure 8.5; their magnitudes are plotted in Figure 8.6.

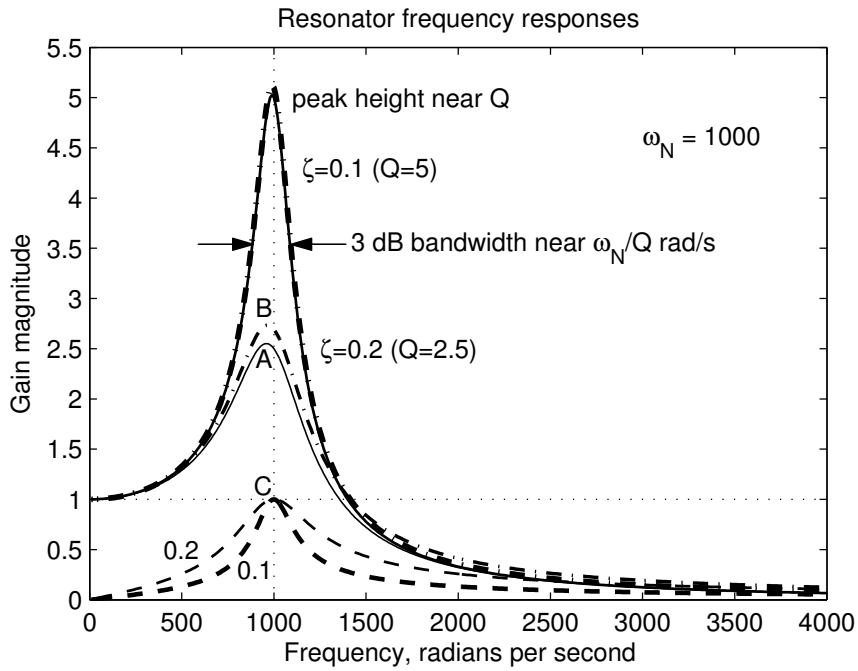


Figure 8.6: The amplitude frequency response of the three resonators, for a natural frequency of 1000 rad/s and damping factors 0.1 and 0.2. Filter A (solid curves) and filter B (dash-dot curves) have unity gain at DC, and higher gain near resonance; filter C (dashed curves) has zero gain at DC, and peaks at unity gain at exactly the natural frequency of the resonance. At low damping, the zero in filter B has little effect, so $\zeta = 0.1$ curves for A and B are very similar.

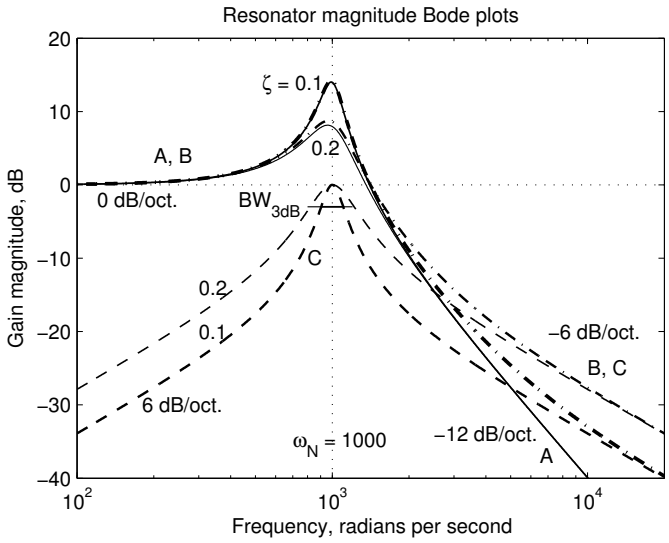


Figure 8.7: Bode plots (dB gain versus logarithmic frequency) of three resonators, for a natural frequency of 1000 rad/s and damping factors 0.1 and 0.2. The Bode plot makes it clear that filter A (solid curves) and filter B (dash-dot curves) are very similar at low frequencies; both have unity gain at DC and higher gain at resonance; in the case of low damping ($\zeta = 0.1$), their peaks overlie each other almost exactly, too. At high frequencies, filter B approaches filter C (dashed curves) which has zero gain at DC and has a peak at unity gain at exactly the natural frequency of the resonance. With twice the damping, the response at 3 dB down is about twice as wide, as marked on the filter C curves.

$$H_{\mathrm{A}}(\omega) = \frac{1}{1 - (\omega/\omega_{\mathrm{N}})^2 + i2\zeta\omega/\omega_{\mathrm{N}}}$$

$$H_{\mathrm{B}}(\omega) = \frac{1 + i2\zeta\omega/\omega_{\mathrm{N}}}{1 - (\omega/\omega_{\mathrm{N}})^2 + i2\zeta\omega/\omega_{\mathrm{N}}}$$

$$H_{\mathrm{C}}(\omega) = \frac{i2\zeta\omega/\omega_{\mathrm{N}}}{1 - (\omega/\omega_{\mathrm{N}})^2 + i2\zeta\omega/\omega_{\mathrm{N}}}$$

$$H_{\mathrm{D}}(\omega) = \frac{1 - a(\omega/\omega_{\mathrm{N}})^2 + i2a\zeta\omega/\omega_{\mathrm{N}}}{1 - (\omega/\omega_{\mathrm{N}})^2 + i2\zeta\omega/\omega_{\mathrm{N}}}$$

To get real closed forms for the amplitude gains (magnitudes of the complex gains), we can take the magnitudes of the denominators and numerators separately (square root of the sum of the squares of the real and imaginary parts) and take their ratio, thereby making it easier to see the effects of poles and zeros separately. Or omit the square roots for the power gains. To make them more concise, let's substitute a

normalized frequency: $\hat{\omega} = \omega/\omega_N$:

$$|H_A(\omega)| = \frac{1}{\sqrt{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}} \qquad |H_A(\omega)|^2 = \frac{1}{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}$$

$$|H_B(\omega)| = \frac{\sqrt{1+4\zeta^2\hat{\omega}^2}}{\sqrt{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}} \qquad |H_B(\omega)|^2 = \frac{1+4\zeta^2\hat{\omega}^2}{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}$$

$$|H_C(\omega)| = \frac{2\zeta\hat{\omega}}{\sqrt{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}} \qquad |H_C(\omega)|^2 = \frac{4\zeta^2\hat{\omega}^2}{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}$$

$$|H_D(\omega)| = \frac{\sqrt{1-(2a-4a^2\zeta^2)\hat{\omega}^2+a^2\hat{\omega}^4}}{\sqrt{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}} \qquad |H_D(\omega)|^2 = \frac{1-(2a-4a^2\zeta^2)\hat{\omega}^2+a^2\hat{\omega}^4}{1-(2-4\zeta^2)\hat{\omega}^2+\hat{\omega}^4}$$

These frequency responses for filters A through C are illustrated as Bode plots in Figure 8.7. For small ζ, the frequency response magnitudes peak near ω_N (near $\hat{\omega} = 1$), where the real part of the complex transfer function denominator is zero, though the exact frequency of the peak varies across the three filters. The gain of filter A is maximized where the denominator of the gain magnitude is minimized, since it has a constant numerator; this frequency is easily found to be $\omega_N\sqrt{1-2\zeta^2}$ (which is somewhat less than the ringing frequency $\omega_R = \omega_N\sqrt{1-\zeta^2}$). The peak or center frequency for filter C is always exactly ω_N. Filter B's peak location is more complicated, but between those of filters A and C.

8.4 Resonator Impulse Responses

For each of the example filters, there must exist corresponding differential equations that describe the relative dynamics of the input and output signals. In the absence of an input, starting from arbitrary states (e.g., arbitrary capacitor voltage and inductor current in a circuit, or arbitrary position and velocity of a mass in a mechanical resonator), the filter outputs are the homogeneous solutions of those equations. We won't bother with the differential equations, because we know that the homogeneous solutions for continuous-time differential equations with constant coefficients can be written in terms of the eigenfunctions that correspond to the poles:

$$y(t) = A_1\exp(p_1 t) + A_2\exp(p_1^* t)$$

for arbitrary complex coefficients A_1 and A_2. The zeros of the system have no effect on the homogeneous solutions.

Since the poles are in complex conjugate relationship, and since for a real system we need to find a real-valued output, the coefficients also need to be in complex conjugate relationship:

$$y(t) = A_1\exp(p_1 t) + A_1^*\exp(p_1^* t)$$

We recognize this as a decaying sinusoid (or a growing sinusoid if p_1 has a positive real part, as in an unstable system). Looking back at the value of p_1 for our resonators, we can see the decay rate is $\gamma = \zeta\omega_N$ and the ringing frequency is $\omega_R = \omega_N\sqrt{1-\zeta^2}$. Then the output can be written as a decaying exponential times a sinusoid at the ringing frequency:

$$y(t) = A\exp(-\gamma t)\cos(\omega_R t + \phi)$$

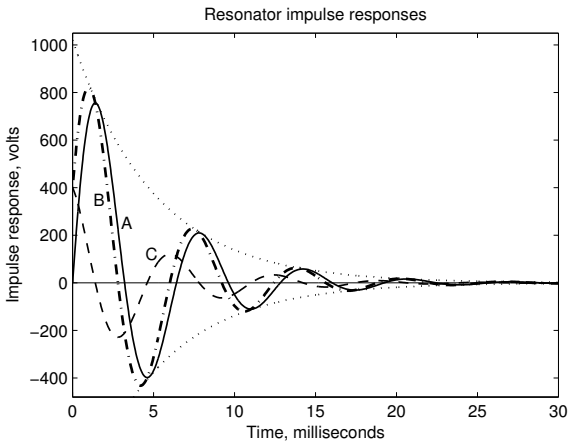


Figure 8.8: Impulse responses of the three resonators, and the exponential envelope (dotted curve) for filter A, for $\zeta = 0.2$. Filter A (solid) and filter B (dash-dot) impulse responses each have an integral of 1, while filter C's impulse response (dashed) integrates to zero. Only filter A shows no step at $t = 0$ (a step has a high-frequency spectrum falling at 6 dB per octave), therefore, only filter A's response falls at 12 dB per octave. The ringing frequency is about $1000/2\pi$ Hz, for a period of about 2π ms.

where A is twice the magnitude, and ϕ (phi) the phase, of the coefficient A_1.

The homogeneous solutions for resonant linear systems are generally of this form: an exponential decay times a sinusoid, with parameters determined by the poles. The impulse responses, however, are not all alike—not just dependent on the poles—and are determined by finding the coefficients (the amplitude and phase) that correspond to the initial conditions caused by a unit impulse. Engineers and mathematicians each have their own methods to make it easy to solve for the coefficients.

We leave it as an exercise for the reader to verify these impulse responses (valid for $t > 0$, $0 < \zeta < 1$) and to verify that h_A and h_B have unit integral and h_C has zero integral:

$$h_\mathrm{A}(t) = \frac{\omega_\mathrm{R}}{1-\zeta^2} \exp(-\gamma t) \sin(\omega_\mathrm{R}\, t)$$

$$h_\mathrm{B}(t) = h_\mathrm{C}(t) + h_\mathrm{A}(t)$$

$$h_\mathrm{C}(t) = \frac{2\zeta\omega_\mathrm{R}}{1-\zeta^2} \exp(-\gamma t) \cos\left(\omega_\mathrm{R}\, t + \sin^{-1}\zeta\right)$$

$$h_\mathrm{D}(t) = a\delta(t) + (1-a)h_\mathrm{A}(t)$$

These impulse responses are illustrated for the first three filters in Figure 8.8 and for filter A for various damping-factor values in Figure 8.9. Filter D has the scaled Dirac delta function impulse $a\delta(t)$ in its impulse response, so is not so easily plotted.

The general form for the first three, $A\exp(-\gamma t)\cos(\omega_\mathrm{R} t + \phi)$, differing only by amplitude factor A and phase ϕ, are also examples of order-1 *gammatone* impulse responses, which we analyze in the next chapter. A phase of $\phi = -\pi/2$, changing the cosine to a sine, corresponds to our filter A, with no step in its impulse response and no zero in its transfer function. This phase and its opposite ($\pi/2$) are the only phases for which

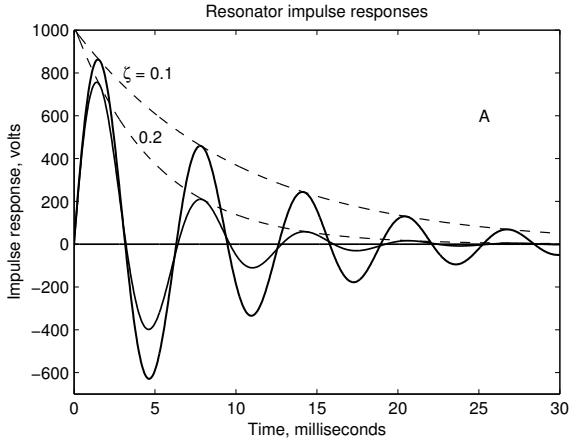


Figure 8.9: Impulse responses of filter A, for $\zeta = 0.1, 0.2$ (solid), and the corresponding different exponentially decaying amplitudes, or envelopes (dashed).

there is no zero, the only phases for which the impulse response doesn't start with a step, and the only phases corresponding to a frequency response that drops at −12 dB/octave at high frequency. For a two-pole filter, each of these special conditions implies the others. Other values of ϕ correspond to transfer functions with one zero, somewhere on the real axis in the s plane (the high-frequency slope, in general, is −6 dB per octave times the difference between the number of poles and the number of zeros). A phase of $\tan^{-1}(\gamma/\omega_R)$, or $\sin^{-1}\zeta$, puts the zero at $s = 0$; there are two such phases, corresponding to filter C and its negation, and these are the only phases for which the Bode plot has a nonzero slope, that is, +6 dB per octave, at low frequencies. These general and special cases are mentioned here because they are key to understanding the low-frequency tail behavior of the gammatone filters that we examine in the next chapter.

8.5 The Complex Resonator and the Universal Resonance Curve

The frequency responses of the resonators, shown in Figure 8.6, are not symmetric about their center frequencies, but are nearly so, not too far from the peak, especially when the damping factor is low. In reasoning with resonances, it is often helpful to have an even simpler description of the frequency response, one that is symmetric, and is "normalized" to have no free parameters at all. A one-pole complex system leads to such an approximation, which has long been used in engineering and physics. In engineering, it is known as the *universal resonance curve* (Terman, 1932; Siebert, 1986); in physics, the power-gain shape (as opposed to its square root, the amplitude gain) is known as a *Lorentzian function* (or sometimes Cauchy, Cauchy–Lorentz, or Breit–Wigner distribution) (Fornasini, 2008). The frequency response is of course exactly the same as that of the one-pole lowpass filter that we analyzed in Figure 6.6; the only difference is that its central peak represents a response to a nonzero frequency rather than to DC.

See Figure 8.10 for an illustration of how the universal resonance approximation relates to the responses of filters A and C. Near its peak frequency, a resonator's response is very close to this standard frequency-symmetric shape, if the Q is high enough (damping ζ low enough). The simple symmetric approximation to the frequency response can be found by considering a complex resonator—that is, a system with nothing but a single complex pole. The impulse response of such a system is not real-valued, but rather a decaying complex exponential. The transfer function, normalized to have unity gain at the ringing frequency (approximately at

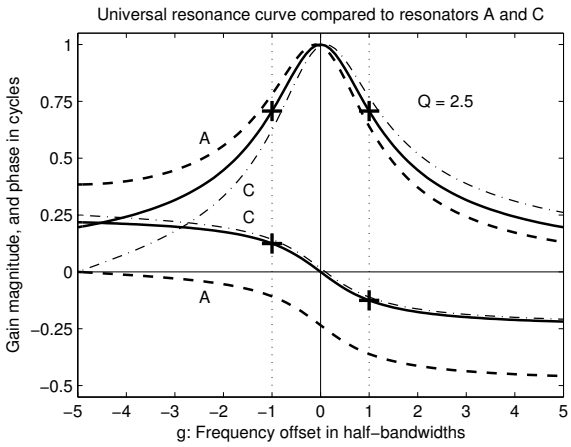


Figure 8.10: Amplitude and phase responses of the universal resonance curve (solid), on a normalized frequency deviation scale that makes it independent of Q, compared to the responses of resonator filters A and C, for $Q = 2.5$ (the A and C curves would be closer together, and the approximation much better, at higher Q than this). The peak gain of filter A has been normalized to 1 to match the other curves, but its phase has not been adjusted to match the condition of zero phase at resonance. The deviation $g = -5$, or 2.5 times the 3 dB bandwidth, corresponds to zero frequency (DC) for the $Q = 2.5$ filters. The 3-dB points of the universal resonance curve, at deviation $g = \pm 1$, gain $\sqrt{2}/2$, and phase ±45 degrees (0.125 cycle) are marked with crosses.

the peak), is:

$$H_1(s) = \frac{\mathrm{Re}(p_1)}{s - p_1} = \frac{\zeta\omega_\mathrm{N}}{s - (-\zeta\omega_\mathrm{N} + i\omega_\mathrm{R})}$$

At very low damping, we can ignore the differences between ω_R and ω_N and the peak or center frequency ω_C, with little loss of accuracy; we will assume $\omega_\mathrm{R} = \omega_\mathrm{N}$ in the above transfer function in subsequent steps. This one-pole transfer function can then be written in terms of a nondimensional frequency deviation from the ringing frequency, scaled by the decay rate $\zeta\omega_\mathrm{R}$. This scaled deviation from center

$$g = \frac{\omega - \omega_\mathrm{R}}{\zeta\omega_\mathrm{R}}$$

along with the usual $s = i\omega$ substitution, leads to this simple frequency response:

$$H_1(\omega) = \frac{1}{1 + ig}$$

This expression for H_1, when applied as an approximation to a real resonator, for frequencies not too far from the resonant peak, is known as the *universal resonance approximation*. Since it is less accurate for low-Q resonators, where ω_R differs more from ω_N, it might be better called the *high-Q resonance approximation*.

The amplitude frequency response is thus approximated as the symmetric function

$$|H_1(\omega)| = \frac{1}{\sqrt{1 + g^2}}$$

The g parameter here can be defined in terms of frequencies in hertz instead of radians per second; and

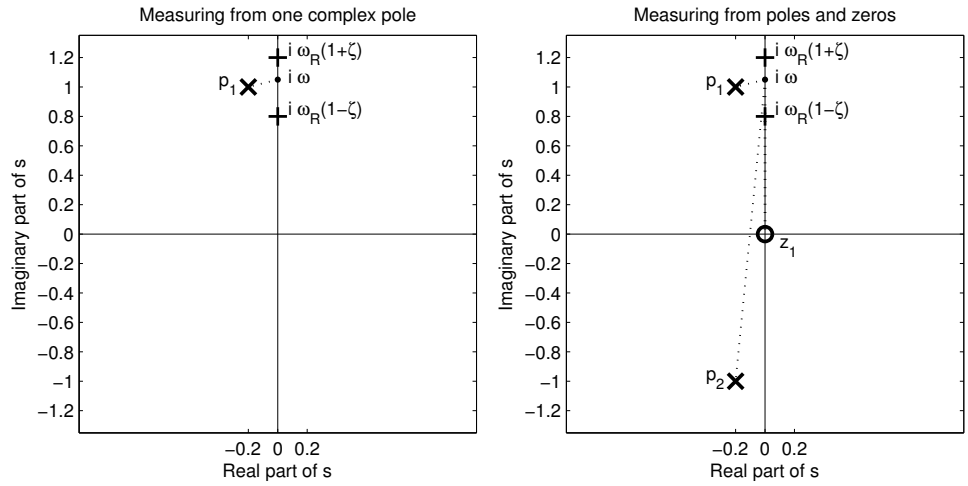


Figure 8.11: Graphical construction of the frequency responses of resonators. Using the left-hand plot with a single complex pole, the gain magnitude of the universal resonance curve, or of the one-pole complex resonator, is inversely proportional to the distance between the pole p_1 and the frequency point $i\omega$ (as a function of ω), here illustrated for ω just above the ringing frequency, ω_R. Frequencies are scaled such that $\omega_R = 1$. The cross marks "+" at $i\omega_R(1 \pm \zeta)$ indicate the points where the distance from the pole to the frequency point increases by $\sqrt{2}$ (the 3-dB points, deviations $g = \pm 1$ in the universal resonance curve, also marked by "+" in Figure 8.10). The real filters have a second pole, p_2; and may have a zero, z_1, as shown in the right plot. The relative distance to the second pole doesn't vary much near the resonance peak, but does move and tip the peak of Filter A a bit when it is included. For filter C, there is also a zero at $s = 0$; the distance to the zero appears in the numerator, so moves and tips the response in the opposite direction. The pole and zero positions shown represent $\zeta = 0.2$ or $Q = 2.5$, corresponding to the curves in Figure 8.10, where DC is at $g = -5$. Pole–zero plots such as these were once commonly used for measurement and calculation of frequency responses, but with modern computers we no longer need such aids; on the other hand, they are still very useful as tools to suggest at a glance the form of the transfer function surfaces of Figure 8.5.

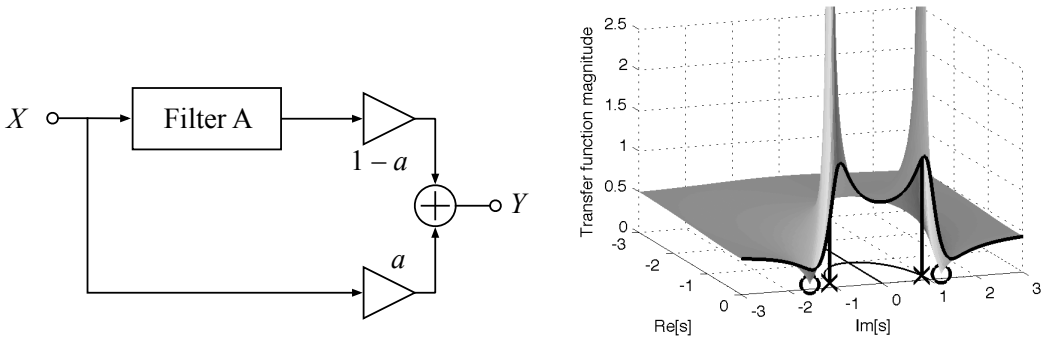


Figure 8.12: Filter D: an asymmetric resonator—schematic and complex transfer function. Adding a straight-through path in parallel to the two-pole resonator of filter A results in a strongly asymmetric peak in the frequency response, involving a complex pair of zeros in addition to the poles inherited from filter A. The ratio of the path gains sets the zero positions. The DC gain is the sum of the path DC gains; as shown, the net DC gain is 1. The illustrated transfer function is for $a = 0.5$ and $\zeta = 0.2$, half the damping of the poles in the illustrations of Figure 8.5, since the zeros near the poles would make the frequency response fairly flat with the higher damping. See the color plates.

we can use Q instead of ζ:

$$g = \frac{f - f_R}{\zeta f_R} = \frac{2Q(f - f_R)}{f_R}$$

The frequency f_R/Q is the exact 3-dB bandwidth of the universal resonance transfer function as defined here; $f_R/(2Q)$ or ζf_R is the half-bandwidth, the frequency corresponding to the decay rate, relative to which we normalize the deviation.

This one-pole approximation is an excellent universal description of resonator transfer functions, for frequency deviations small enough that the distances in the s plane from the frequency point $i\omega$ to the other pole p_1^* (or to any zeros) does not change by a significant factor from what it is at zero deviation. See the graphical computation illustrated in Figure 8.11. That means it is generally good near the peak for low damping or high Q. But this approximation is never good in the limit of low frequencies, where a resonator's other pole is as close as the one under consideration, and real zeros have their greatest effect. In hearing, where asymmetry is important, and where the low-frequency tail is often of interest, this approximation is therefore of limited use. But it is very helpful in understanding the derivation, behavior, and limitations of certain popular filter types such as the gammatone family.

8.6 Complex Zeros from a Parallel System

Consider filter D, the parallel connection of filter A with an *identity* system, or straight-through circuit path. To keep the DC gain equal to unity, apply gains to each path, summing to 1, as shown in Figure 8.12. Its transfer function is $H_D = a + (1 - a)H_\mathrm{A}$.

Consider the special case of equal path gains: $a = 0.5$. At any location in the s plane for which the transfer function of filter A is -1 (that is, magnitude of 1 and phase of 180 degrees), the gains of the two paths will sum to 0, so this parallel system will have zeros at such places. We can find those locations graphically, or by looking for zeros of the transfer function algebraically:

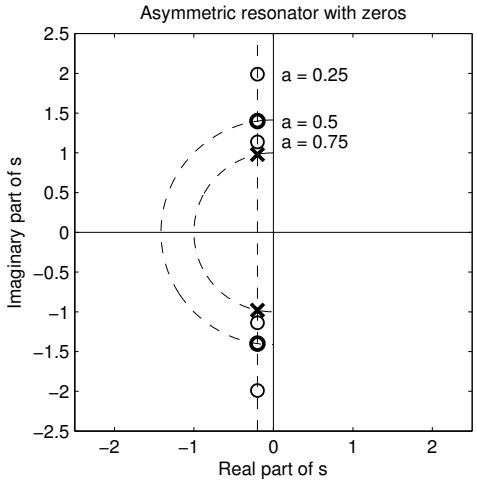


Figure 8.13: The s-plane pole–zero plot for filter D. Using mixing weight $a = 0.5$ as shown in Figure 8.12, the zeros (shown bold) are on a circle of larger radius than that of the poles, by a factor of $\sqrt{2}$, and at the same x coordinate. For other mixing weights, the zeros move to other positions at the same x coordinate; examples for weights $a = 0.25$ and $a = 0.75$ are shown.

$$
\begin{aligned}
H_D &= 0.5 + 0.5H_A \\
H_D &= 0.5 + \frac{0.5}{(s/\omega_\mathrm{N})^2 + 2\zeta s/\omega_\mathrm{N} + 1} \\
&= \frac{0.5(s/\omega_\mathrm{N})^2 + \zeta s/\omega_\mathrm{N} + 1}{(s/\omega_\mathrm{N})^2 + 2\zeta s/\omega_\mathrm{N} + 1}
\end{aligned}
$$

The numerator has roots corresponding to a natural frequency of $\sqrt{2}\omega_\mathrm{N}$ and a damping factor of $\zeta/\sqrt{2}$, as shown in Figure 8.13 (recall from Section 8.2 that natural frequency is distance of the roots from the origin in the s plane, and damping factor is the negative real part of the roots normalized by natural frequency). Graphically, we can see that these zero locations must be on the vertical line through the poles, since those are the only places with the 180 degree phase shift (90 degrees being contributed by each of the two poles on that line) that would allow the resonator output to destructively cancel the input. That observation is consistent with the zeros being on a circle of natural frequency that's larger by $\sqrt{2}$ than that of the poles, with damping that is decreased by the same factor. The resulting frequency responses are shown in Figure 8.14, and impulse responses in Figure 8.15.

More generally, for mixing parameter a between 0 and 1:

$$
\begin{aligned}
H_D &= a + (1 - a)H_A \\
H_D &= a + \frac{1 - a}{(s/\omega_\mathrm{N})^2 + 2\zeta s/\omega_\mathrm{N} + 1} \\
&= \frac{a(s/\omega_\mathrm{N})^2 + 2a\zeta s/\omega_\mathrm{N} + 1}{(s/\omega_\mathrm{N})^2 + 2\zeta s/\omega_\mathrm{N} + 1}
\end{aligned}
$$

By the same reasoning as before, the zeros stay on the vertical line through the poles, at a radius $\omega_\mathrm{N}/\sqrt{a}$,

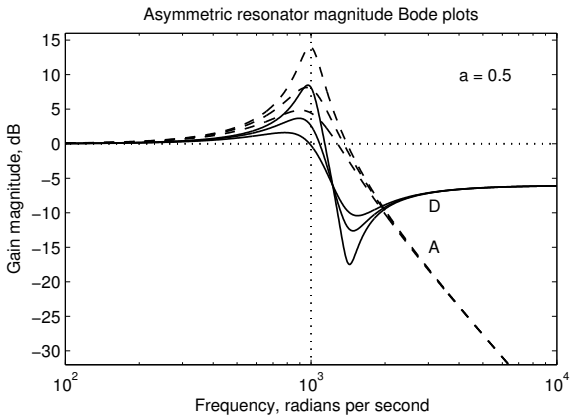


Figure 8.14: Bode plot for the asymmetric resonator (filter D, solid), compared to the all-pole resonator (filter A, dashed), for damping factors 0.1, 0.2, and 0.4, with mixing weight $a = 0.5$. The zeros cause an *anti-resonance* or *notch* about a half octave above the resonance. The high-frequency asymptote is flat, which means the impulse response will contain an impulse.

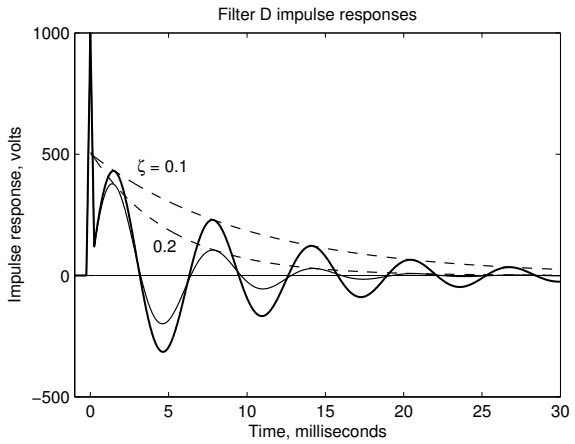


Figure 8.15: The impulse responses of the asymmetric resonator, filter D, contain an impulse of weight a (illustrated here with $a = 0.5$) at $t = 0$, representing the straight-through path. Since an impulse is infinitely tall and narrow, it is approximated in this plot by a triangle, of base width 0.001 s and height 1000, with the correct total area 0.5. Impulse responses for two dampings are shown, with the envelopes (dashed) of the exponentially decaying portions.

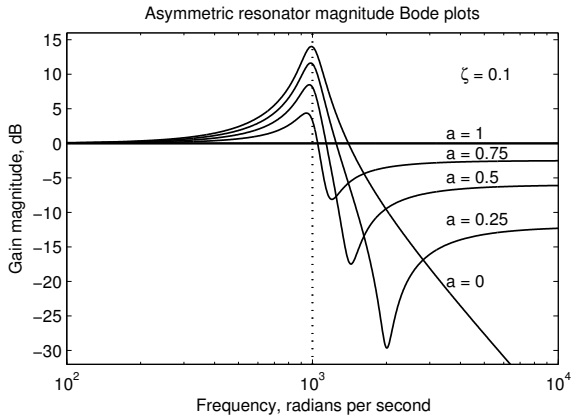


Figure 8.16: Bode plot for the asymmetric resonator with variable mixing ratio. The weight of the "straight through" signal is a, and of the two-pole resonator is $1-a$. This parameter interpolates between a flat response at $a = 1$ and the response of the simple resonator at $a = 0$, resulting in a cancellation dip above the peak.

and the high-frequency gain approaches a. The resulting frequency response gains are shown in Figure 8.16. As the straight-through weight a approaches 1, the zeros will move to cancel the poles, giving a flat response.

In the vicinity of the peak, the zero causes a pronounced asymmetry—a response much steeper on the high side than on the low side—which is why we refer to it as the *asymmetric resonator*. This simple filter gives us enough control over the shape of the frequency response to be a key part of models that we develop later. We use it as a stage in the *pole–zero filter cascade* (PZFC) auditory filter model, and, in a digital version, as a stage in our digital cochlear model, the *cascade of asymmetric resonators with fast-acting compression* (CARFAC).

8.7 Keeping It Real

Let us define a *real system* as a system in the real world, where measurements are never complex, or as a system where a real input always produces a real output. Either way, a real LTI system is one with a real-valued impulse response.

Conversely, a complex LTI system is one with a complex impulse response. Real systems can be combined in various ways to model a complex system, and complex systems can sometimes be combined to make a real system. Taking the real part of a complex system's output makes a nonlinear system in general, but there is always a real linear system, as shown in Figure 8.17, that behaves identically whenever the input is real. This relationship between taking the real part of a complex system output and an equivalent real system is useful in understanding complex resonators, and their relatives the complex gammatones introduced in the next chapter.

For general complex frequencies s, the transfer function of a real-valued LTI system will satisfy the symmetry constraint:

$$H(s) = H^*(s^*)$$

so that when the input is a real sinusoid—a sum of two complex exponentials in complex-conjugate relationship—the output will also be real due to corresponding eigenvalues (gains) being kept in a complex-conjugate relationship.

If we have a linear system that produces a complex output from a real input—for example from a differential equation with complex coefficients—we still have a well-defined impulse response $h(t)$, which will be

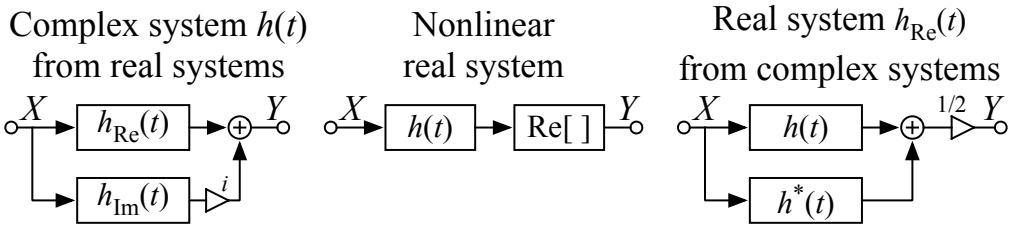


Figure 8.17: Real and complex systems: the complex system on the left, made from two real systems, can be followed by a real-part operator to make a nonlinear system, center; when the input is real, this nonlinear system is equivalent to the linear system on the right. Therefore, we can model the real system on the right via the simpler (lower-order) complex system on the left and a real-part operator as in the middle; or we can implement a discrete-time system such as the complex one on the left, using complex numbers in a computer, and take the real part of its output, as a way to implement the system on the right for real inputs.

complex, and transfer function $H(s)$ that will *not* have the symmetry mentioned above. It is sometimes useful to know what happens if we take the real part (or the imaginary part) of the output of such a complex linear system.

The real-part operator is not linear, since $\mathrm{Re}[aX] \neq a\mathrm{Re}[X]$ for complex a. For a complex system with real input (for example, the left part of Figure 8.17), taking the real part of the output (the middle part of Figure 8.17) is equivalent to making a real system (the right part of Figure 8.17) whose impulse response is the real part of the complex system impulse response. That is, the impulse response $\mathrm{Re}[h(t)] = [h(t) + h^*(t)]/2$ defines a real linear system that responds exactly the same as the nonlinear system formed by following the original linear system by the real-part operator—but only when the input is real. For complex input, this new linear system will have complex output, unlike the nonlinear system whose output is forced to always be real.

The complex-conjugate operator is not linear (for the same reason), but applying the complex-conjugate operator to a complex impulse response makes a different complex linear system with impulse response $h^*(t)$, which we used above. From the definition of the Laplace transform it is easy to show that the transfer function corresponding to $h^*(t)$ is $H^*(s^*)$.

The transfer function of the real system $[h(t) + h^*(t)]/2$ can then be found by linearity:

$$H_{\mathrm{Re}}(s) = \frac{H(s) + H^*(s^*)}{2}$$

Similarly, the transfer function from real input to the imaginary part of the output of a complex system, using $\mathrm{Im}[h(t)] = (h(t) - h^*(t))/2i$, is:

$$H_{\mathrm{Im}}(s) = \frac{H(s) - H^*(s^*)}{2i}$$

When the complex system can be described by a rational transfer function, the transfer function $H^*(s^*)$ can be obtained by flipping all the poles and zeros of $H(s)$ to their complex-conjugate locations. These equations thereby induce pole–zero plots with complex-conjugate symmetry, as required for a real system.

As an example, the transfer function of a continuous-time complex filter with a single complex pole at s_p (like the universal resonance approximation) can be written as:

$$H(s) = \frac{1}{s - s_p}$$

and the system corresponding to its conjugate $h^*(t)$ as:

$$H^*(s^*) = \left(\frac{1}{s^* - s_p}\right)^* = \frac{1}{s - s_p^*}$$

These systems produce complex outputs when given a real input. They have no circuit or other direct physical interpretation, since physical variables such as voltage, current, velocity, and displacement must be real. If we are using such a system for its mathematical simplicity, we typically want to relate it back to a nearly equivalent real system, for example by taking the real part of the output. If we take the real part of the output, for real input, applying the analysis above we get a related real linear system with these two poles and a zero:

$$\begin{aligned} H_{\text{Re}}(x) &= \frac{1}{2}\left[\frac{1}{s - s_p} + \frac{1}{s - s_p^*}\right] \\ &= \frac{1}{2}\left[\frac{(s - s_p^*) + (s - s_p)}{(s - s_p)(s - s_p^*)}\right] \\ &= \frac{s - \text{Re}[s_p]}{(s - s_p)(s - s_p^*)} \end{aligned}$$

while if we take the imaginary part of the output instead, we get a real system with the same two poles, but no zero, like filter A:

$$\begin{aligned} H_{\text{Im}}(x) &= \frac{1}{2i}\left[\frac{1}{s - s_p} - \frac{1}{s - s_p^*}\right] \\ &= \frac{1}{2i}\left[\frac{(s - s_p^*) - (s - s_p)}{(s - s_p)(s - s_p^*)}\right] \\ &= \frac{\text{Im}[s_p]}{(s - s_p)(s - s_p^*)} \end{aligned}$$

These two linear systems share the same poles, or resonant modes, but differ in their zeros, like the various resonator circuits we analyzed.

The same math applies to discrete-time systems and the Z transform, with z instead of s in the transfer functions; we apply it in the next section to understand a real digital filter made from a complex one.

8.8 Digital Resonators

The second-order digital filter forms of Figure 7.5 and Figure 7.8 can implement two-pole filters with up to two zeros. Here we introduce a few specialized variants.

The two-pole resonator with unity gain at DC (filter A) can be realized in a form that allows easy movement of the poles without a separate gain adjustment, as shown in Figure 8.18. This form can be augmented with forward coefficients, determining both the DC gain and one or two zeros, as shown in Figure 8.19; this form holds a constant DC gain as the poles are moved.

An alternative to the direct form is the coupled form (Gold and Rader, 1969), also known as a phasor

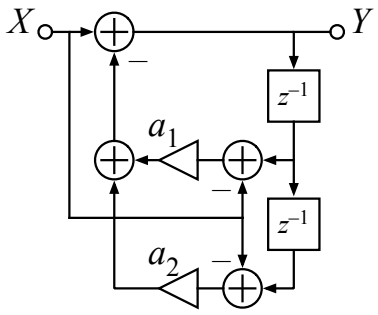


Figure 8.18: A direct-form two-pole filter stage with the input connected as shown here has unity gain at DC. It is evident by inspection that if the input is constant and the output is equal to the input, then the subtracted feedback will be zero, no matter what the coefficient values a_1 and a_2 are, due to the differences being zero where the input is subtracted from the two delayed outputs; therefore output equal to constant input is an equilibrium point, so the DC gain is 1. The transfer function to Y as shown also has two zeros at $z = 0$, and is a minimum-phase transfer function (assuming the poles are inside the unit circle, making it stable); an output taken after one or both of the z^{-1} delay elements has one or two samples of extra delay, canceling one or both zeros, so would not be minimum phase.

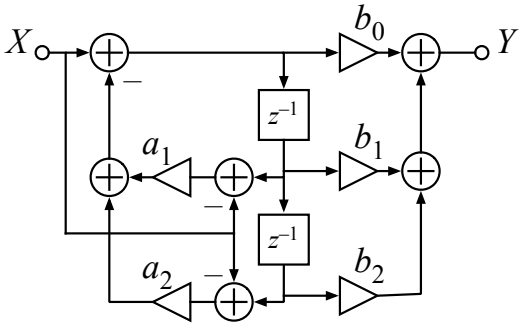


Figure 8.19: The two-pole filter of Figure 8.18 is easily modified to include zeros. In this form, the poles can be moved, while the zeros are held fixed, without the DC gain changing. The relationship of coefficient values to pole and zero locations is as discussed in Section 7.11, except that pole-location-dependent numerator of the A factor there is subsumed in the input gain $1 + a_1 + a_2$ provided by the way the input is connected here.

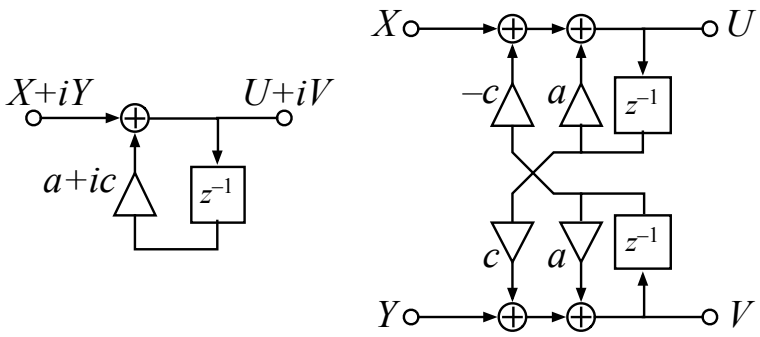


Figure 8.20: A one-pole complex-valued filter (left) is equivalent to the real-valued two-pole filter known as a *coupled-form* filter (right), if X, Y, U, and V are real, as is evident by expanding the complex multiplication into its four real terms. The system on the right can be viewed as a complex system with one pole at $z_p = a+ic$, or as a two-input–two-output linear system with real outputs whenever both inputs are real. The two-input–two-output system can nevertheless be analyzed like other LTI systems in terms of complex inputs, outputs, and eigenfunctions. If the input to the complex system on the left is real, then that system is equivalent to the one on right with the Y port unused (zero imaginary part). For that one-input system, taking only the real-part output, U, gives two poles, at z_p and z_p^*, and one real zero at $z_z = a$ (and a zero at $z = 0$ that keeps it minimum-phase). Similarly, the transfer function from the X input to the real V output has only the pole pair (and the zero at $z = 0$), as can be verified with a little algebra.

filter (Massie, 2012). A two-pole coupled-form filter is essentially a complex-valued one-pole filter, as shown in Figure 8.20. Its output can be taken in various ways, as complex or as the real or imaginary part of the one-pole complex filter.

The real-part output gives a resonator with a zero (like filter B, but not the same location of the real zero), while the imaginary-part output is a digital version of filter A. To see this, start with the transfer function of the one-pole filter with pole at complex z_p (by the same analysis as we used for the real first-order discrete filter in Chapter 7):

$$H(z) = \frac{z}{z - z_p}$$

Recall from the preceding section that the real-part operation applied to the impulse response can be expressed as $(h+h^*)/2$, and the transform of h^* is $H^*(z^*)$, which is like $H(z)$ with the poles and zeros conjugated, so we get this transfer function to the real-part output:

$$\begin{aligned} H_{\text{Re}}(z) &= \frac{1}{2}\left[\frac{z}{z - z_p} + \frac{z}{z - z_p^*}\right] \\ &= \frac{z}{2}\left[\frac{(z - z_p^*) + (z - z_p)}{(z - z_p)(z - z_p^*)}\right] \\ &= \frac{z(z - \text{Re}[z_p])}{(z - z_p)(z - z_p^*)} \end{aligned}$$

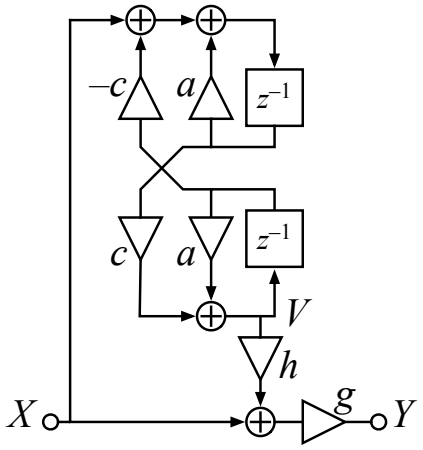


Figure 8.21: A pair of zeros is added to the coupled form by mixing the input with a minimum-phase all-pole filtered version, as in filter D.

Alternatively, taking the imaginary-part output gives two poles only:

$$\begin{aligned} H_{\mathrm{Im}}(z) &= \frac{1}{2i}\left[\frac{z}{z-z_p} - \frac{z}{z-z_p^*}\right] \\ &= \frac{z}{2i}\left[\frac{(z-z_p^*)-(z-z_p)}{(z-z_p)(z-z_p^*)}\right] \\ &= \frac{z\mathrm{Im}[z_p]}{(z-z_p)(z-z_p^*)} \end{aligned}$$

The factors of z in the numerators, corresponding to zeros at $z = 0$, are "unit advance" factors that cancel extra delay that the filter would have otherwise—leaving these factors out would correspond to cascading a z^{-1} operator, taking the filter outputs one sample later. Enough such factors to make the numerator the same order as the denominator allows the filter to be minimum phase, producing the output as early as causally possible. Notice that $H_{\mathrm{Im}}(z)$ has only one zero at $z = 0$, and a second-order denominator; that means its output comes one sample later than it would if the filter were minimum-phase. To make a minimum phase version, one could compute the weighted sum of the two delay inputs, rather than the two delay outputs as shown (assuming the Y input is eliminated). Since these zeros at the origin have a flat frequency response, filters with no other zeros, such as $H_{\mathrm{Im}}(z)$, are still typically referred to as *all-pole*; they correspond to all-pole continuous-time filters.

Figure 8.21 shows how the coupled-form filter can be connected to produce a weighted mixture of its input and its output, making the digital version of our filter D, the *asymmetric resonator* discussed in Section 8.6, which we will use in Chapter 16.

The "single-tuned resonators" described in this chapter have been fundamental to models of cochlear function since the time of Helmholtz, and they are the building blocks for all the more elaborate hearing models that we explore in later chapters. Digital versions of the resonators make simple and efficient computer functions for sound analysis.

Chapter 9

Gammatone and Related Filters

> The form $m(t)$ appears both as the integrand in the definition of the Gamma function $\Gamma(\gamma)$ and as the density function of the Gamma distribution, therefore we propose to use ... the term "Gamma-tone" or "γ-tone."
>
> — "Spectro-temporal receptive fields of auditory neurons...," Aertsen and Johannesma (1980)

9.1 Compound Resonators as Auditory Models

In this chapter, we explore a number of filter models and structures, especially the *gammatone family*, grounded in the resonators studied in the previous chapter. The gammatone and gammachirp and related filters are popular in sound analysis because they approximate auditory function better than simple resonances do, and because they are straightforward to implement. These filters provide a range of frequency-response shapes, significantly different from those of the simple resonance, but still with very simple parameterizations.

This chapter focuses on the properties and implementations of the gammatone family, a family characterized by having multiple poles at the same location. The use of several gammatone-family filters as models of auditory function is explored more deeply in Chapter 13. Although we ultimately move away from the gammatone family in making good models of the cochlea, the cascade-structured filters that we arrive at are still very gammatone-like in their structure and in their responses. The simpler gammatones are a good place to start to understand auditory filtering.

The principles of single-tuned resonators are easy to apply to systems of multiple poles. Filters with multiple coincident poles, the gammatone family, are particularly easy to analyze, and to describe and characterize mathematically. All three resonator circuits that we analyzed in earlier chapters are gammatone filters of order one—the simplest possible case, in which there is only one pole at each location. These simple resonators are also special cases of some of the other filters we will encounter.

The same mathematical techniques, starting from transfer functions and pole–zero descriptions, make it easy to analyze other classes of filters, such as the filter cascades that we use to model the filtering that is accomplished in the cochlea via a traveling hydromechanical wave, as described in subsequent chapters. To the extent that we can describe filters by poles and zeros, we can make digital filters that efficiently process sounds.

9.2 Multiple Poles

Usually (when a filter has distinct poles), we can represent the impulse response of a filter as some coefficients times the complex exponentials that correspond to pole locations. Consider the simple case of two real poles,

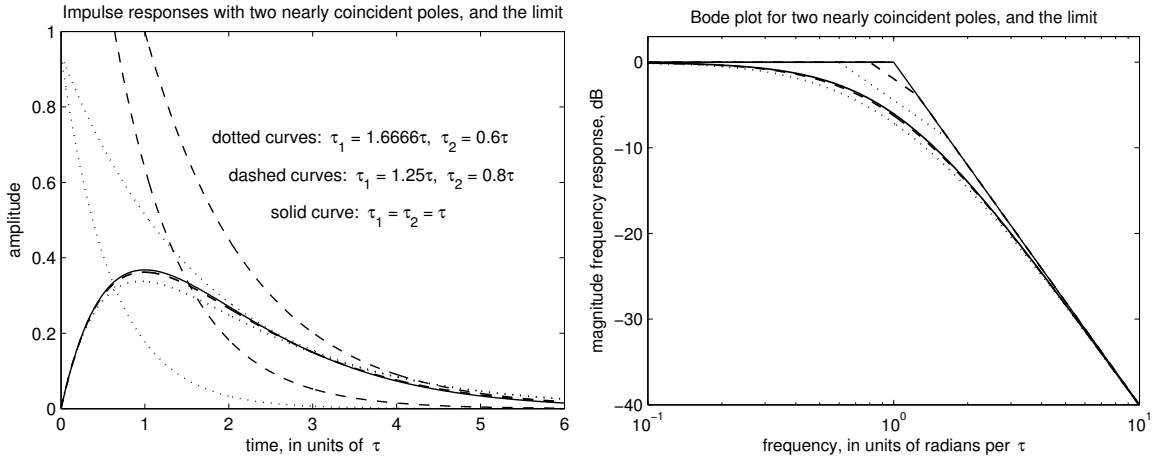


Figure 9.1: In the left panel, the solid curve is the limit of the difference-of-exponential curves that are the impulse responses of smoothing filters with two real poles with a geometric mean time constant τ. The dotted curves show the two scaled exponentials, and their difference, for pole time constants that differ by almost a factor of three. The dashed curves correspond to closer time constants. In the right panel, the same curve styles are used for the Bode plots of the corresponding amplitude frequency responses. With two poles, the filters yield −12 dB/octave (−40 dB/decade) rolloff; a short section of −6 dB/octave near the corner barely affects the shape.

as from a cascade of two first-order smoothing filters, with time constants τ_1 and τ_2. Its impulse response is, as with all linear time-invariant systems (see Section 6.5), a linear combination of the homogeneous responses, the exponential decays corresponding to the pole time constants:

$$h(t) = \frac{1}{\tau_1 - \tau_2}\left[\exp\left(\frac{-t}{\tau_1}\right) - \exp\left(\frac{-t}{\tau_2}\right)\right]$$

where the coefficients $1/(\tau_1 - \tau_2)$ and $-1/(\tau_1 - \tau_2)$ can be deduced from the observation that the impulse response of the two-pole smoothing filter must start without a step discontinuity at $t = 0$, and must have an integral of 1.0 (for unity gain at DC).

But what happens when a system has several poles at exactly the same place? The coefficients in the formula above are unbounded as the two time constants approach each other, and the formula does not work when they are equal. But the impulse response itself does approach a limit. As both time constants approach τ, the limit of the impulse response is easily found to be

$$h(t) = \frac{t}{\tau^2}\,\exp\left(\frac{-t}{\tau}\right)$$

in which the factor of t is a feature that we have not previously encountered in an impulse response. This expression is a good approximation to the impulse response even for a pair of unequal time constants, such as $\tau_1 = 0.8\tau$ and $\tau_2 = 1.25\tau$, as shown in Figure 9.1.

More generally, if we cascade N identical one-pole first-order lowpass filters, the transfer function is easy: the Nth power of the transfer function of one stage. The impulse response is less obvious, and a bit more complicated; it turns out to involve a shape known as a *gamma distribution*, which includes a power-of-t factor:

$$h_N(t) = \frac{1}{(N-1)!\,\tau^N}\,t^{(N-1)}\,\exp\left(\frac{-t}{\tau}\right)$$

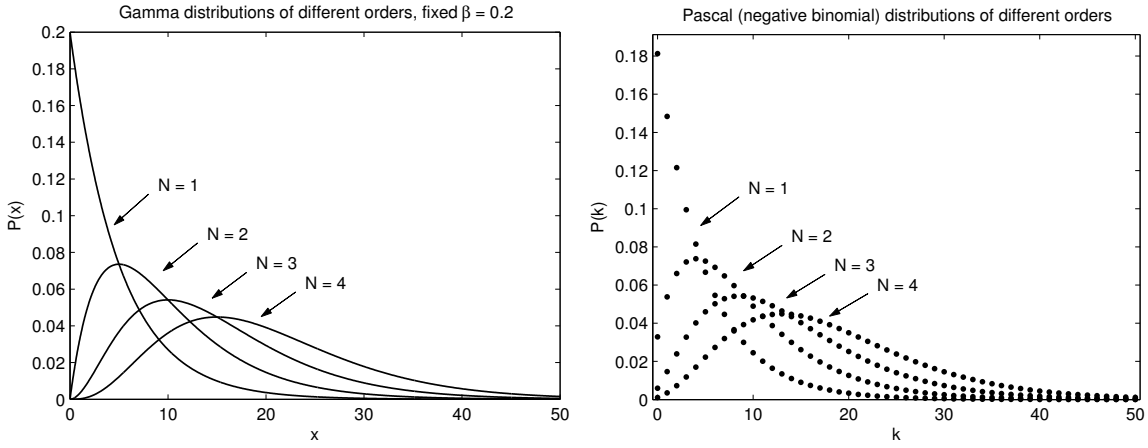


Figure 9.2: On the left are the impulse responses of cascades of N identical continuous-time one-pole smoothing filters, which are also gamma or Erlang distributions, for several values of N. On the right are impulse responses of cascades of N identical discrete-time one-pole smoothing filters with time constants of about 5 samples, which are also Pascal or negative binomial distributions of nonnegative integers.

This impulse response is illustrated, for several different N, in Figure 9.2, along with its discrete-time version. Ignoring the factor in front that normalizes it to have unity integral, the shape is $t^{N-1}\exp(-t/\tau)$, corresponding to N poles at $s = -1/\tau$. The parameter N is known as the pole order. This gamma distribution shape, or something very close to it, is a key part of the impulse response of essentially all systems that arise in modeling cochlear filtering. Various derivations and explanations of "the t factor" and the gamma distribution, and how they arise in filtering, can be found in the literature (Papoulis, 1962; Healy and Huggins, 1974; Bean, 2001). Papoulis (1962) shows that systems of multiple real poles can often be effectively approximated by systems of coincident poles; the gammatone family extends this concept to systems of complex pole pairs.

We often write filter impulse responses in terms of a decay rate parameter γ (gamma) rather than a time constant τ (that is, $\gamma = 1/\tau$); γ is familiar from resonator analysis as the negative real part of the pole location. Since the s-plane pole is at $s = -\gamma$, this parameter is preferred when talking about pole coordinates, as in subsequent sections. The lowpass impulse response parameterized by γ is:

$$h_N(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)$$

9.3 The Complex Gammatone Filter

If we shift the N-pole smoothing filter's poles up (in the direction parallel to the positive imaginary axis) by a distance ω_R, from $s = -\gamma$ to $s = -\gamma + i\omega_R$ (that is, to where the pole was in the one-pole complex universal resonance of Figure 8.11), we get a system known as the *complex gammatone* filter. Invoking the shifting property of the Laplace transform (see box "Math Connection: Shifting Property of the Laplace Transform"), we find that the impulse response, shown in Figure 9.3, is the gamma distribution (the impulse response of the N-pole system before the shift) multiplied by the complex exponential $\exp(i\omega_R t)$ that corresponds to the shift in the s plane:

$$h_{cgt}(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)\, \exp(i\omega_R t)$$

Math Connection: Shifting Property of the Laplace Transform

Laplace transforms have many interesting mathematical properties. One that we use in this chapter is the *shifting property*: a shift in the s-plane corresponds to a multiplication by a complex exponential in the time domain. That is, if $X(s)$ is the Laplace transform of $x(t)$, then $X(s-d)$ is the Laplace transform of $x(t)\exp(dt)$, for any real or complex constant d.

If $X(s)$ is a rational function, then the shifted $X(s-d)$ is a rational function with the poles and zeros of $X(s)$ shifted in the s-plane by adding d to their locations. For example, the first-order smoothing transfer function $1/(\tau s + 1)$ shifted to $1/(\tau(s-d)+1)$ has its pole moved from $-1/\tau$ to $d - 1/\tau$. The corresponding impulse response changes from $\exp(-t/\tau)$ to $\exp(-t/\tau)\exp(dt) = \exp(t(d-1/\tau))$. If d is real, this shift is equivalent to a change of time constant (and a change of gain), but still gives a first-order filter, with time constant moved from τ to $\tau/(1-d\tau)$, which is stable if the shift is not too great ($d < 1/\tau$ leaves the pole in the left half plane).

If d is pure imaginary, the pole at $1/(\tau(s-d)+1)$ is off the real axis. The impulse response retains its original decay rate, and the factor $\exp(dt)$ is oscillatory, so the filter becomes a complex resonator. We encounter both real and imaginary shifts in analyzing gammatone-family filters.

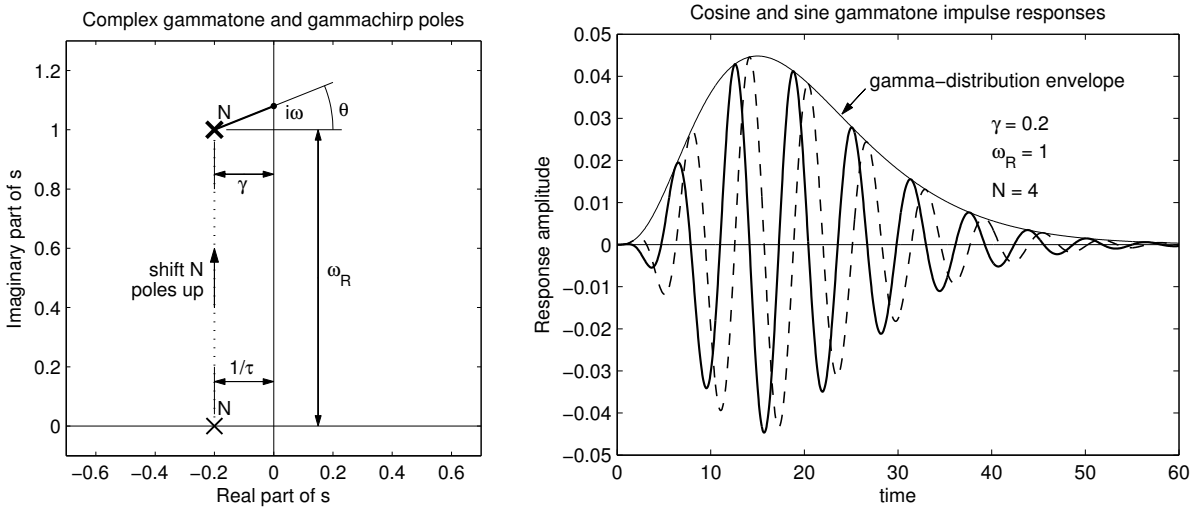


Figure 9.3: The complex gammatone is a system with N coincident poles (plot in the s plane, left), *shifted* from the position of the poles of an N-pole smoothing filter with decay rate γ. As a function of frequency ω, the magnitude frequency response is inversely proportional to the Nth power of the length of the line from the poles to the frequency point $i\omega$, and the phase lag is N times the angle θ shown, going through zero at $\omega = \omega_R$. The real and imaginary parts (right, solid and dashed curves, respectively) of the complex gammatone impulse response are found by multiplying the gamma-distribution envelope (smooth curve), the impulse response of the lowpass, by the oscillatory $\exp(i\omega_R t)$ caused by the shifting of the poles by $i\omega_R$. These real- and imaginary-part curves are real gammatones, of cosine and sine phase respectively. For these plots, the order is $N = 4$ and $\gamma/\omega_R = 0.2$.

Statistics Connection: The Gamma Distribution

In statistics, the *probability density function* (PDF) of a continuous random variable is analogous to the impulse response of a continuous-time smoothing filter in linear system theory, in that it has an integral of one and is subject to transforms and convolutions. The analogy is best for smoothing filters with nonnegative impulse responses, as opposed to those that ring, since PDFs are nonnegative everywhere. The PDF of the sum of two independent random variables is the convolution of their individual PDFs, so it is analogous to the impulse response of a cascade of two smoothing filters with the individual PDFs as their impulse responses. Causal impulse responses correspond to PDFs of nonnegative random variables.

It is common in statistics to consider the PDF of a sum of N *independent identically distributed* (i.i.d.) random variables—analogous to the impulse response of a cascade of N identical filters. When the individual PDFs are one-sided exponential distributions of mean $1/\beta$, $P(x) = \beta \exp(-\beta x)$, $x > 0$ (like the impulse response of the one-pole smoothing filter), the resulting PDF is well known as the gamma distribution (also known as the Erlang distribution or Pearson type III distribution). Its formula is:

$$P(x) = \frac{\beta^N}{(N-1)!}\, x^{(N-1)} \exp(-\beta x)$$

The Erlang distribution is applicable only to integer values of N, which is usually all we need; for the more general gamma distribution, the denominator $(N-1)!$ is typically written in terms of the gamma function as $\Gamma(N)$ to apply as well to noninteger N.

Calculations on PDFs are often done via Fourier or Laplace transforms. These transforms, known respectively as *characteristic functions* and *moment-generating functions* of the distributions (Miller and Childers, 2012), are analogous to linear system frequency responses and transfer functions. It is easy to find formulas for the distribution and its characteristic and moment-generating functions in tables. Similar formulas are also found in tables of Fourier and Laplace transforms in linear systems books, typically with slightly different terminology. In statistics, the sign convention for the moment-generating function is different; with parameter t corresponding to $-s$, the moment-generating function of the gamma distribution converges to the left of the poles in t:

$$M(t) = \frac{1}{(1 - t/\beta)^N} \quad \text{for} \quad \mathrm{Re}[t] < \beta$$

This moment-generating function corresponds to the Laplace transform of the impulse response, which converges to the right of the poles in s. Statisticians sometimes write $t < \beta$, as they usually consider only real values of the parameter t, which are enough for generating the moments of the distribution.

Statistics Connection: Scale-Space Smoothing Filters

Not all smoothing filters are analogous to PDFs, since a PDF must be nonnegative. But some important families of smoothing filters, such as those used in scale-space analysis (Witkin, 1983), do respect a nonnegativity constraint. Well-known properties of the PDFs of sums of random variables are then immediately applicable to the problem of successive smoothings at different scales. In particular, the mean of a random variable is analogous to a delay (the low frequency group delay, or how far the center of mass of an impulse response is displaced from $t = 0$); and the standard deviation is analogous to a temporal spread, a measure of the time over which signals are smoothed. For sums of random variables, the means add, and the variances add (the variance being the square of the standard deviation). Therefore, in cascades of smoothing filters, delays add and smoothing time constants combine via the square root of sum of squares; these properties hold even if the filters are not unity gain at DC, so not quite PDFs.

For the one-pole smoothing filter, the delay and temporal spread are both equal to the time constant τ, so a cascade of N of them has a delay of $N\tau$ and a spread of $\sqrt{N}\tau$. The mean and standard deviation of the gamma distribution, in terms of the rate parameter β, are correspondingly N/β and $\sqrt{N}/\beta$.

There are corresponding analogies between discrete-time impulse responses and *probability mass functions* of discrete random variables, with Z-transforms known as *probability generating functions*. When Lindeberg (1990) worked out the details of smoothing filters suitable for discrete scale-space filtering, he did so in the language of generating functions. A cascade of N discrete-time one-pole smoothing filters has an impulse response analogous to a Pascal distribution or negative binomial distribution, the distribution of a sum of N geometrically distributed integers, as illustrated in Figure 9.2.

The order-N complex gammatone's frequency response can be calculated from the pole–zero plot as shown in Figure 9.3. It is most succinctly expressed as the Nth power of the universal resonance curve:

$$|H_{\text{cgt}}(i\omega)| = \frac{1}{(1 + g^2)^{N/2}}$$

where g is the normalized frequency deviation:

$$g = \frac{\omega - \omega_{\text{R}}}{\gamma}$$

This complex gammatone has unity gain at its peak, or passband center $g = 0$, since it is a frequency shift of a lowpass filter with unity gain at DC. If a gain factor is included, the result is still known as a complex gammatone. A constant factor in front of the transfer function or impulse response, even if complex, will not affect the gamma-distribution envelope shape, nor the poles. A constant phase-only factor $\exp(i\phi)$ will not change the frequency response magnitude at all, so the impulse response of the complex gammatone described by one set of N coincident poles and the given frequency response magnitude is conventionally generalized to include a phase parameter ϕ:

$$h_{\text{cgt}}(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)\, \exp(i\omega_{\text{R}} t + i\phi)$$

Since the magnitude frequency response is an exact function of the square of the frequency deviation g, it is symmetric about its peak, independent of ϕ. This symmetric frequency response of the complex gammatone filter is an excellent approximation to that of the *real* gammatone discussed next, except in the low-frequency tail, as shown in Figure 9.4. For the sort of models that we want to build, with signal-adaptive parameters that make them good models of cochlear filtering, getting the tail right is important, so we need to look more

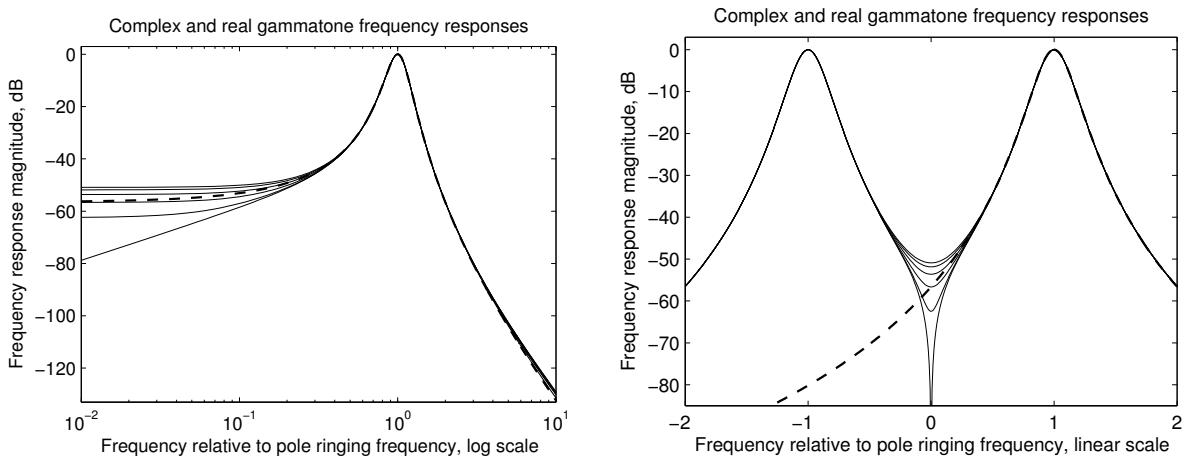


Figure 9.4: The log-magnitude frequency responses of complex (dashed curve) and real (solid curves) gammatones, with $\zeta = 0.2$, $N = 4$, and a variety of gammatone phases, on both log (left) and linear (right) frequency scales (the gains of the real gammatones have been adjusted to match the unity peak gain of the complex gammatone). The complex gammatone response is exactly symmetric about its peak frequency on a linear frequency scale, while real filters must be symmetric about zero frequency, as the right plot shows.

closely at the real versions.

9.4 The Real Gammatone Filter

The *real* gammatone filter is the system whose impulse response is the real part of the complex gammatone's impulse response. The name *gammatone*, traditionally meaning the real version, describes this impulse response, a gamma distribution times a tone (a real sinusoid) of arbitrary phase:

$$h_{gt}(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)\, \cos(\omega_{\mathrm{R}} t + \phi)$$

The *real-part* operation corresponds to half the sum of a signal and its complex conjugate. As we discussed in Chapter 8, the real-part operation on an impulse response corresponds to a real system with transfer function equal to half the sum of the original transfer function and its modification by conjugating the pole and zero locations. From the Nth-order complex gammatone, with N coincident poles at $p = -\gamma + i\omega_{\mathrm{R}}$, neglecting overall gain and phase factors, we get this real filter of order $2N$:

$$\begin{aligned} H_{gt}(s) &= \frac{1}{(s-p)^N} + \frac{1}{(s-p^*)^N} \\ &= \frac{(s-p^*)^N + (s-p)^N}{(s-p)^N\,(s-p^*)^N} \end{aligned}$$

The sum operation produces a numerator that has roots (zeros of the transfer function) that the complex filters did not have. The locations of the zeros of $(s-p^*)^N + (s-p)^N$ can be deduced by reasoning about what it takes for the two complex terms to cancel. The cancellation needed to make zeros can occur only on the real axis, which is where the two terms have the same magnitude. To cancel, the terms also need to have phases that differ by π, which can only happen for phases of $\pm\pi/2$ for each term.

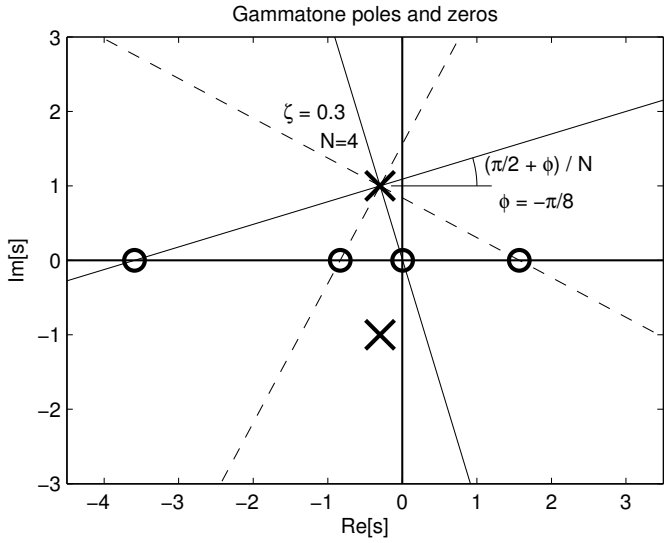


Figure 9.5: The s-plane locations of the real gammatone's N zeros can be constructed from the poles, as described in the text. Solid lines show where the top pole cluster contributes $\pi/2$ radians or 90 degrees of phase, and dashed lines show where it contributes $-\pi/2$ radians or 270 degrees. Where these rays meet the real axis, the other pole cluster provides the opposite phase, and the contributions cancel, making the four zeros shown. In this example, the gammatone phase parameter is $\phi = -\pi/8$; other phases rotate this pattern, moving the zeros along the real axis. Notice also that for these particular parameters, the damping $\zeta = 0.3$ shown puts one of the zeros very close to $s = 0$, making a very low gain in the low-frequency tail response; certain other combinations of phase, damping, and order will do the same. Other special phase values ($\phi = \pm\pi/2$) will move one of the zeros out to infinity, leaving $N - 1$ real zeros.

The phase of $(s - p)^N$ goes through N cycles, or $2N$ occurrences of these special phase values, for any loop that encircles p; half of these occurrences will lie on rays that intersect the real axis, determining the locations of the zeros, and the other half will be "around back" where they won't cause zeros. When a phase factor is included, the sum looks a bit more complicated, but calculating the zeros is basically as described, with a phase adjustment, as illustrated in Figure 9.5:

$$\begin{aligned} H_{gt}(s) &= \frac{\exp(i\phi)}{(s-p)^N} + \frac{\exp(-i\phi)}{(s-p^*)^N} \\ &= \frac{\exp(i\phi)\,(s-p^*)^N + \exp(-i\phi)\,(s-p)^N}{(s-p)^N\,(s-p^*)^N} \end{aligned}$$

Thus, the gammatone filter has N zeros on the real axis in the s plane, in places that depend on ϕ, ζ, and N; or it has $N - 1$ zeros for special values of ϕ that push one zero out to infinity. This is exactly the difference we saw between our resonator filters A and B. For particular parameters, such as the combination illustrated in Figure 9.5, the gammatone filter will have one of its zeros at DC (that is, at $s = 0$), and a 6 dB/octave slope in the low-frequency tail, like our resonator filter C.

In typical applications and analyses encountered in the hearing literature, these zeros are ignored, not calculated. Their complicated effect on the low-frequency tail may go unrecognized, or may be tolerated. Or the zeros may be removed, to make an all-pole gammatone filter (Van Compernolle, 1991; Slaney, 1993; Lyon, 1996a), as discussed in the next section.

In auditory filter applications, gammatone orders of 3 to 5 are typical (Patterson et al., 1992). Relative to real data, whether psychophysical or physiological, their advantage over simple resonators is that the skirts of the frequency response (the close-in parts of the tails) fall off much more quickly than those of the single-tuned resonance, without having to make the peak too narrow; but they don't fall *too* quickly, as the skirts of a Gaussian filter do, if the order is moderate.

Gammatones are a special case of gammachirps. Plots of the gammatone impulse responses and frequency responses are shown in Section 9.6 below on gammachirps.

The 3-dB bandwidth of gammatone-family filters is less than that of the underlying individual resonators, by a factor of about $1/\sqrt{N}$, due to the approximately parabolic shape of the resonance peaks. The effective Q of the gammatone ($Q_{3\text{dB}}$, the ratio of center frequency to 3-dB bandwidth) is therefore higher than the Q of the poles that make it:

$$Q_{3\text{dB}} \approx Q\sqrt{N}$$

9.5 All-Pole Gammatone Filters

Gammatone filters have been well studied in terms of their poles and zeros (Van Compernolle, 1991; Slaney, 1993), extending earlier analyses that started with the impulse response and approximated the frequency response in terms of the complex gammatone, which is symmetric about its peak frequency (Holdsworth et al., 1988; de Boer and Kruidenier, 1990). Various people noticed that the zeros could complicate the implementation or the analysis of the response, and that removing them to make all-pole and one-zero gammatone filters could provide useful approximations (Van Compernolle, 1991; Slaney, 1993) or even provide significant advantages in modeling auditory filtering (Lyon, 1996a; Robert and Eriksson, 1999; Katsiamis et al., 2006, 2007; Lyon, 2011a), as is discussed in subsequent chapters.

The all-pole gammatone filter (APGF) is a cascade of N identical copies of our resonator filter A, and hence both easy to analyze and easy to build. It can be regarded as a gammatone filter with the zeros removed, or the gammatone transfer function with the numerator replaced by a real constant. Neglecting overall gain, its transfer function is:

$$H_{apgf}(s) = \frac{1}{(s-p)^N\,(s-p^*)^N} = \frac{1}{((s+\gamma)^2+\omega_{\text{R}}^2)^N}$$

which if adjusted for unity gain at $s = 0$ is the Nth power of filter A (see Section 8.2):

$$H_{apgf}(s) = \frac{1}{((s/\omega_{\text{N}})^2 + 2\zeta s/\omega_{\text{N}} + 1)^N}$$

The impulse response is only a little different from the gammatone impulse response. It can be described as a sum of gammatones of different orders, or as a gamma distribution times a Bessel function, as illustrated in Figure 9.6 (Lyon, 1996b). Typical APGF and corresponding DAPGF (differentiated APGF, with one zero at $s = 0$) impulse responses are shown in Figure 9.7.

The one-zero gammatone filter (OZGF) is an APGF times a transfer function $s - z_1$ for a single real zero z_1. For the special case of $z_1 = 0$, the factor is just s, a derivative operator, or zero at DC, yielding the differentiated all-pole gammatone filter (DAPGF); this special case of the OZGF has a sloped (6 dB per octave) low-frequency tail. The OZGF (with one zero at a finite distance left of the s-plane origin) has a tail between those of the APGD and DAPGF, flattening out as very low frequency, as illustrated in Figure 9.8.

The APGF and OZGF filters are not as symmetric as the gammatones. Due to having fewer zeros, they have higher slopes on the high side: $-12N$ dB/octave rather than just the $-6N$ or $-6(N+1)$ dB/octave of the gammatone. And with fewer low-frequency zeros, they also have shallower slopes on the low-frequency side of the peak. Compare Figure 9.4 with Figure 9.8. The direction of this asymmetry is appropriate for auditory

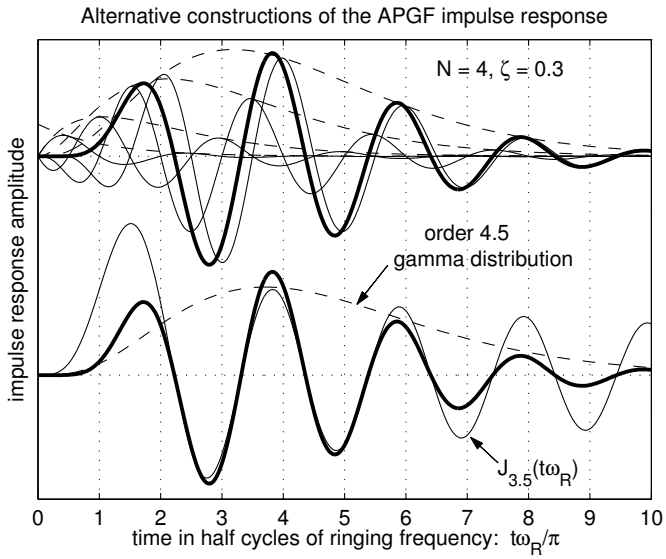


Figure 9.6: The impulse response of the Nth-order all-pole gammatone filter (heavy curves) can be constructed as the sum of gammatones of orders 1 through N, appropriately scaled and phased, as shown on the top; gamma-distribution envelopes are shown dashed. Below, the same impulse response is constructed as the product of a gamma distribution of order parameter $N + 0.5$ times a Bessel function of the first kind, of order $N - 0.5$ (for this illustration, the amplitude of the gamma distribution factor has been scaled down by a factor of 5, and the Bessel function has been scaled up by a factor of 5, for clarity). The damping parameter affects only the exponential time constants of the envelopes.

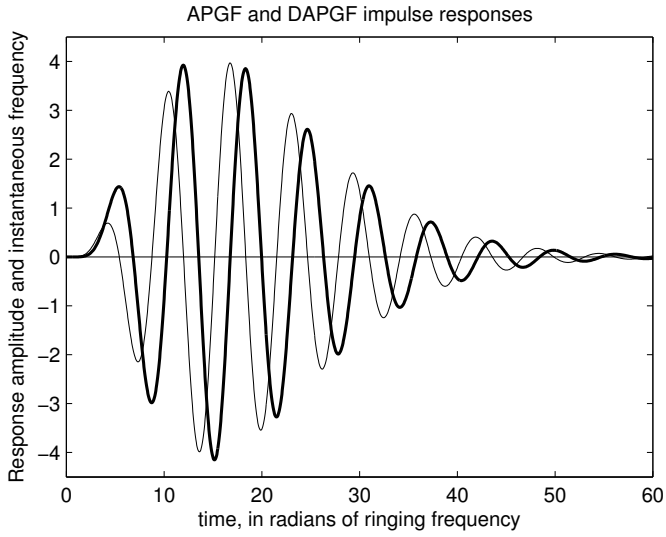


Figure 9.7: Impulse responses of the 4th-order all-pole gammatone filter (APGF, heavy curve) and its derivative, the DAPGF (light curve), with pole damping $\zeta = 0.2$ ($Q = 2.5$, $Q_{3\text{dB}} \approx 5$).

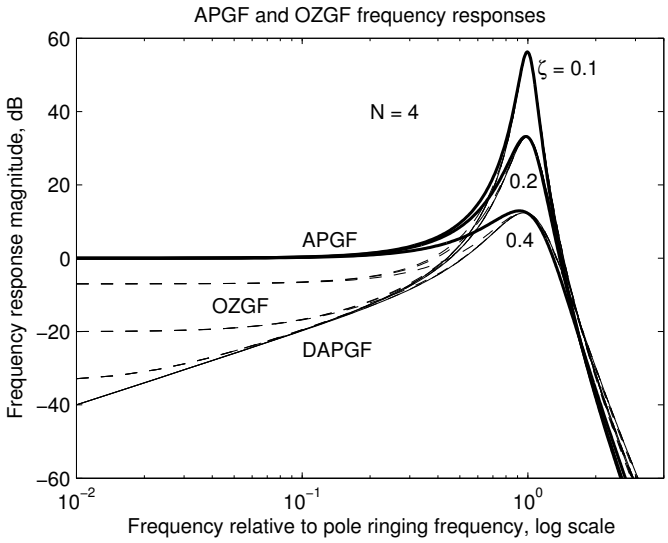


Figure 9.8: Amplitude frequency responses of 4th-order APGF and several OZGFs, including the limiting DAPGF, for three damping factors. Note that the low-frequency tails do not depend on the damping, unlike the case with (real) gammatone filters. The position of the one real zero in the OZGF interpolates between the APGF (zero at infinity) and the DAPGF (zero at $s = 0$), whether the zero is at positive or negative s. The APGF responses are the fourth powers of the resonator A responses shown in Figure 8.7, so the curves are precisely the same as in that figure, but with the dB scale changed by a factor of 4.

filters, as we will see in Chapter 13.

Compared to the modest range of gain variations of the resonator in Figure 8.7, the $N = 4$ filters of Figure 9.8 achieve a gain variation of about a factor of 256 (48 dB) when the damping and bandwidth change by only a factor of 4. This property is useful because auditory filters need a large range of peak-gain variation with only a moderate bandwidth change to model compression in the cochlea. Achieving this relationship through the combined effect of several variable-Q poles in cascade is common to the gammatones and to the filter-cascade models of the cochlea that we develop in later chapters.

9.6 Gammachirp Filters

The symmetric frequency response of the gammatones is a disadvantage for fitting experimental data on auditory function. The all-pole and one-zero variants provide a reasonable asymmetry for many situations, but not explicit control of the asymmetry. A more controllable alternative is the *gammachip*: a gammatone modified with a phase term that makes it *chirp* (Irino and Patterson, 1997). The changing-frequency tone is also known as a *glide*, especially when it is observed in the auditory system (de Boer and Nuttall, 1997; Carney et al., 1999).

The gammachirp impulse response is like the gammatone, but with an added log-time phase term. The complex and real versions are:

$$h_{\mathrm{cgc}}(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)\, \exp(i\omega_{\mathrm{R}} t + ic\log(t) + i\phi)$$

$$h_{\mathrm{gc}}(t) = \frac{\gamma^N}{(N-1)!}\, t^{(N-1)}\, \exp(-\gamma t)\, \cos(\omega_{\mathrm{R}} t + c\log(t) + \phi)$$

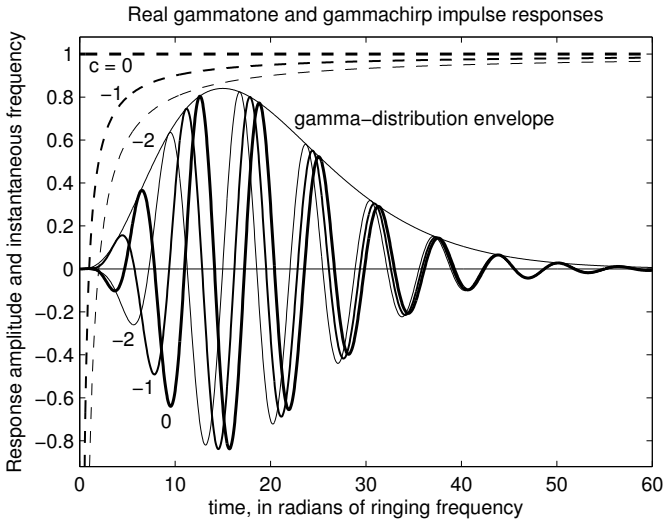


Figure 9.9: Impulse responses (lower curves) and instantaneous frequencies (upper curves) of 4th-order complex gammachirp filters with c values of 0 (gammatone case), -1, and -2. The amplitude scale is arbitrary, the frequency scale is normalized relative to ω_R, and the phases have been chosen to align near one of the later peaks. The pole Q is 2.5, which makes the effective filter $Q_{3\,\mathrm{dB}}$ about 5. The gamma-distribution envelope is also plotted. Notice that with negative c, the zero-crossing times are stretched out more near the beginning, relative to the equally spaced zero crossings of the $c = 0$ gammatone. The zero-crossing times don't change if the envelope is changed by changing the pole Q.

The log-time phase term in these impulse responses represents a changing instantaneous frequency, or rate of change of phase, equal to $\omega_R + c/t$. Although this frequency appears to be ill-behaved near $t = 0$, the amplitude is zero there (for $N > 1$), and the frequency becomes reasonable before the amplitude becomes significant, as shown in Figure 9.9. Typically, auditory filters use a negative value of c, so the instantaneous frequency starts below zero and chirps up to approach ω_R. The frequency appears to increase from a low frequency, but not from a negative frequency.

The frequency response is made asymmetric by chirping. With a negative c, the response can resemble the APGF response, less steep on the low side and more steep on the high side of the peak, which is property that we need if we want to make accurate models of auditory filtering.

Like the complex gammatone, the complex gammachirp's transfer function has N coincident poles. But it also has another factor that can't be expressed as a rational function, so it can't be completely described by poles and zeros, and has no exact implementation in circuits of lumped elements. It is still a linear time-invariant system, though. The impulse response, which completely characterizes it, is the solution of a differential equation with some factors of t in the coefficients (Irino and Patterson, 1997), as opposed to the constant coefficients needed to be able to convert to a rational transfer function.

The asymmetry of the magnitude frequency response, shown in Figure 9.10, can be represented via the symmetric complex gammatone response of Section 9.3 and an asymmetric multiplicative gain factor, the exponential of an antisymmetric function of frequency deviation (Irino and Patterson, 1997, 2001):

$$|H_{\mathrm{cgc}}(i\omega)| = |H_{\mathrm{cgt}}(i\omega)| \exp\left(c \tan^{-1}\left(\frac{\omega - \omega_R}{\gamma}\right)\right)$$

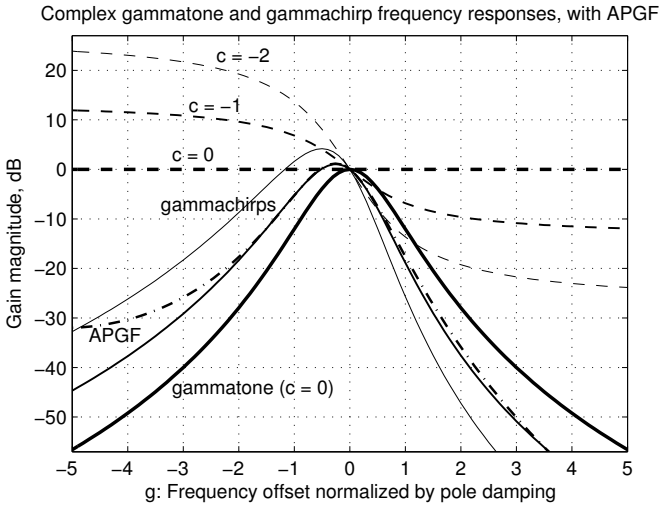


Figure 9.10: Amplitude frequency responses of 4th-order complex gammachirp filters with $c \leq 0$ (solid curves), including the complex gammatone (heavy symmetric curve). The dashed curves show the anti-symmetric log-gain function that converts the gammatone to the gammachirp. An all-pole gammatone filter (APGF) with $\zeta = 0.2$ is also shown (dash-dot line), approximately aligned with the $c = -1$ gammachirp, for comparison of their asymmetries. For real gammatones and gammachirps, the low-frequency tails can be somewhat above or below the tails illustrated for the complex filters.

which can be expressed in terms of the normalized deviation g of Section 8.5 as:

$$|H_{\text{cgc}}(i\omega)| = (1 + g^2)^{-N/2} \exp\left(c \tan^{-1}(g)\right)$$

The angle $\tan^{-1}(g)$ here is the same as the angle θ in Figure 9.3, as may be seen at the illustrated frequency point $i\omega$ by observing $\tan\theta = (\omega - \omega_{\text{R}})/\gamma = g$. The chirping can therefore be seen as a way to couple some of the complex gammatone's antisymmetric phase response into a magnitude response asymmetry.

By the same reasoning as with the real gammatone, the real gammachirp has zeros on the real axis, at locations that depend on all the parameters, and therefore a somewhat more complicated and variable frequency response in the low-frequency tail, as shown in Figure 9.11.

Rational approximations to the gammachirp have been used in digital filters for speech analysis/synthesis systems (Irino and Unoki, 1999). Four additional pole pairs and four additional zero pairs are used to approximate the filter that converts a gammatone filter to a gammachirp filter. The resulting s-plane or z-plane pole–zero diagram has 16 poles and 11 or 12 zeros. The positions of the added poles and zeros can be optimized for the damping and chirping parameters (Unoki et al., 2001).

9.7 Variable Pole Q

Gammatone-family filters have been popular in machine hearing systems and in auditory modeling in general, since their few parameters are enough to fairly accurately match most kinds of data from psychophysical and physiological experiments.

The auditory system's level dependence, an important nonlinearity that is apparent in many different experiments, can be modeled by varying the pole Q of these filters in a signal-dependent way, affecting the gain, bandwidth, and ringing time. Since the poles are all coincident, there is only one damping or Q parameter

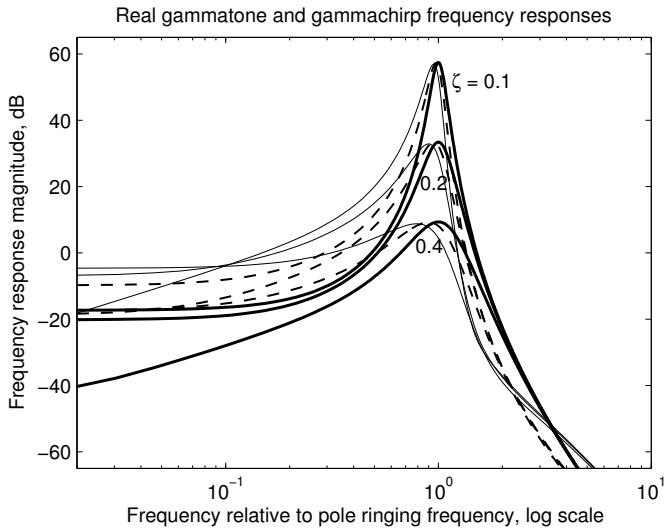


Figure 9.11: The Fourier transforms of the impulse responses are displayed here, for the three impulse responses of the previous figures, plus modifications with half and double damping values. The $c = -1$ gammachirp (dashed curves) has a nonmonotonic damping dependence in the tail, as a zero happens to move close to $s = 0$ when the damping is reduced. In several cases shown, a zero moves to near $s = 0$, pushing the low-frequency tail down; in some of the gammachirp cases, there's also a dip on the high side, due to another spurious zero that comes from interference between the complex gammachirp and its conjugate. These behaviors also have a complicated dependence on the ϕ values used; here we use the same values as in Figure 9.9.

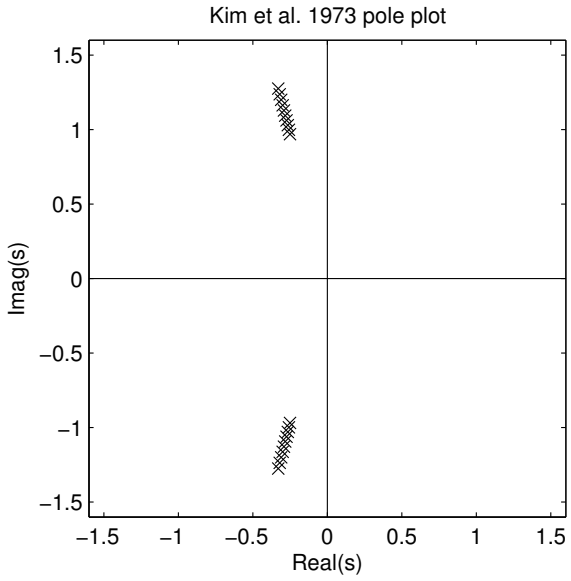


Figure 9.12: The ten pole pairs of Kim, Molnar, and Pfeiffer's model of hydromechanical filtering in the inner ear, normalized to the lowest pole natural frequency.

to vary. Data from cochlear mechanics show that the tail of the response transfer function is nearly constant (not level dependent) for very low frequencies, so matching the data requires a form of filter in which the low-frequency tail remains stable (that is, with nonvarying gain, so the system is nearly linear at low frequencies) as the peak gain changes with Q, providing high gain to weak sounds and low gain to loud sounds. It is hard to get this linear tail property when zeros on the real axis move with the Q and other parameters, which is why the APGF and OZGF are preferred over gammatones and gammachirps in this application.

9.8 Noncoincident Poles

When resonators are combined to make higher-order filters, there is no reason that the poles need to be exactly coincident to make a gammatone-like response. This observation allows us to connect the simple gammatone-family filters to a variety of filter-cascade models that we develop in later chapters.

For example, an early auditory model by Kim, Molnar, and Pfeiffer (1973) used a cascade of ten resonators with slightly staggered pole locations (their system was nonlinear, too, but in the low-level linear limit was an all-pole system). With 10 pole pairs, damping $\zeta = 0.25$ ($Q = 2$) per stage, and natural frequencies staggered by 3% per resonator stage (see Figure 9.12), this filter achieves a peak gain of 60 dB relative to its low-frequency tail gain, with a 3-dB bandwidth of 0.17 times the peak frequency (filter $Q_{3\text{dB}} = 6$), very close to what would have been expected for 10 coincident pole pairs, an order-10 APGF with $Q_{3\text{dB}}$ near $2\sqrt{10}$. The complex transfer function is illustrated in Figure 9.13.

9.9 Digital Implementations

Researchers have explored several different methods of efficiently implementing gammatones and their variants as digital filters (Holdsworth et al., 1988; Van Compernolle, 1991; Darling, 1991; Cooke, 1993; Slaney, 1993; Irino and Unoki, 1999; Van Immerseel and Peeters, 2003). The most direct and elegant method of

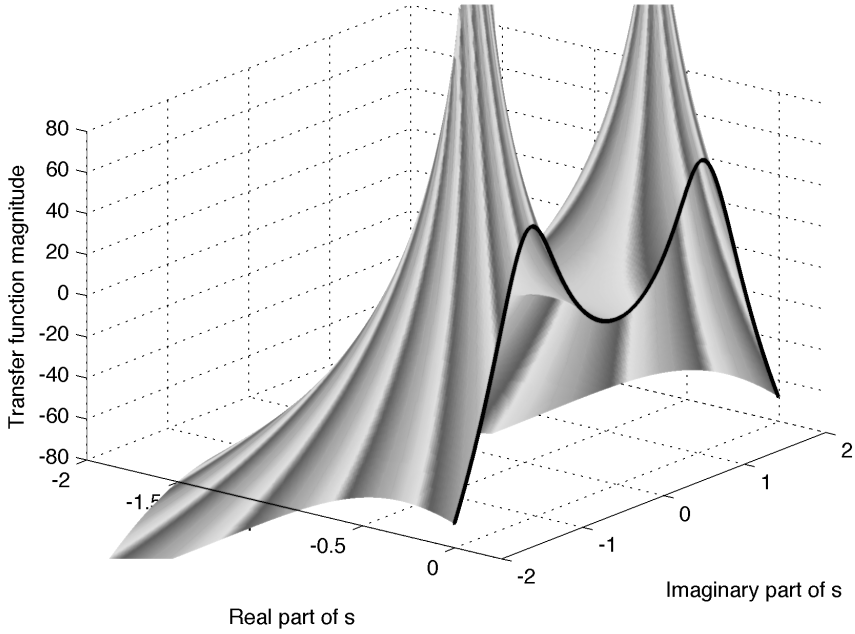


Figure 9.13: The complex transfer function for the Kim et al. filter. The cut line on the imaginary s axis shows the log-magnitude transfer function, with zero frequency in the center. The phase (hue in the color plate) goes through 10 cycles around each cluster of 10 poles.

implementing the real gammatone is probably that of Darling (1991), to cascade N identical one-pole complex filters, and take the real part of the output. Conceptually, this works for continuous-time complex filters, but with digital discrete-time filters it is more practical. The approach is illustrated in Figure 9.14, using the two-pole coupled-form stage of Figure 8.20.

The impulse-invariance digital filter design method that we introduced in Section 7.9 seeks to design a discrete-time filter from a continuous-time filter, such that the impulse response of the discrete-time filter is exactly a sampled version of the impulse response of the continuous-time filter. Using this method for an all-pole system is easy, as it corresponds to just mapping the pole locations individually or in complex-conjugate pairs, and cascading the resulting simple filters. A complex s-plane pole at $s_p = -\gamma + i\omega_\mathrm{R}$ is mapped to a z-plane pole at $z_p = a + ic = \exp(-\gamma T + i\omega_\mathrm{R} T)$ (that is, at radius $r = \sqrt{a^2 + c^2} = \exp(-\gamma T)$, and at angle $\theta_R = \tan^{-1}(c/a) = \omega_\mathrm{R} T$ in the z plane). The first-order complex filter that implements this pole has feedback coefficient $a + ic = r\cos(\theta_R) + ir\sin(\theta_R)$; expanding it into real operations gives the coupled-form second-order filter shown in Figure 8.20.

Because the transfer function to the imaginary-part output (V in Figure 8.20) has no zeros, we can make all-pole real filters with pairs of poles by taking the lower output of the coupled-form filter stage, as in the cascade structure of Figure 9.15 that makes an all-pole gammatone filter.

The real part of the output of a cascade of multiple complex stages, as in Figure 9.14, as opposed to just one stage, gives a more complicated set of real zeros, like that we found in Section 9.4 but in the z plane. That is, by cascading complex outputs to complex inputs, and taking either the real part or the imaginary part at the end, we realize the discrete-time approximation to a real gammatone, zeros and all.

The coupled-form stage with imaginary-part output V has a transfer function $V(z)/X(z)$ that we'll call

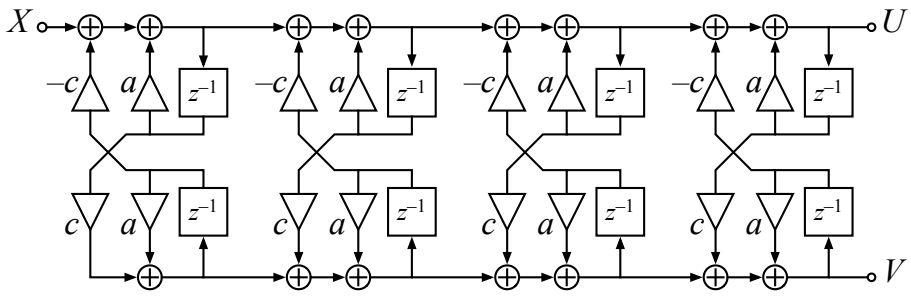


Figure 9.14: A cascade of N complex one-pole filters, with a real input and a real output, would make exactly a real gammatone filter if they were continuous-time filters. With discrete-time filters such as these coupled-form stages, it is an excellent digital approximation to the gammatone, including the zeros that come from taking the real part at the output. The outputs U and V are digital real gammatones of different phases, while $U + iV$ is a complex gammatone. A proportionate change in all of the a and c coefficients moves all the poles together, changing the damping without changing the ringing frequency.

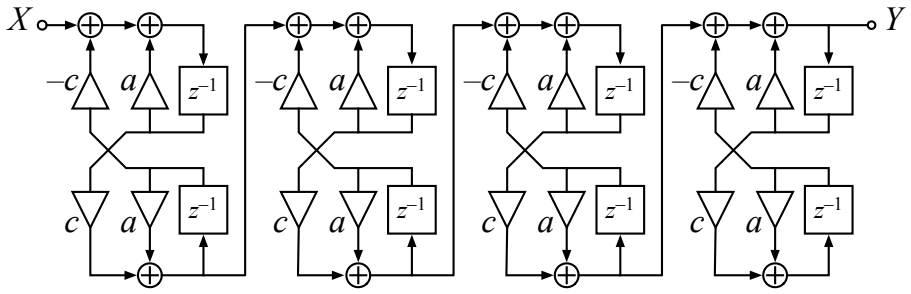


Figure 9.15: A cascade of N identical real two-pole digital filters is the impulse-invariance digital implementation of the all-pole gammatone filter—its impulse response is exactly a sampled version of the continuous-time APGF impulse response. Each coupled-form filter stage of this structure can be interpreted as a one-pole complex filter, with only the imaginary part of the output being used; equivalently, each stage is a two-pole real-valued filter with no zeros (except at $z = 0$, as explained in Section 8.8). The same structure, with graduated pole frequencies instead of identical, will implement the linearized Kim et al. (1973) filter or other all-pole filter cascade.

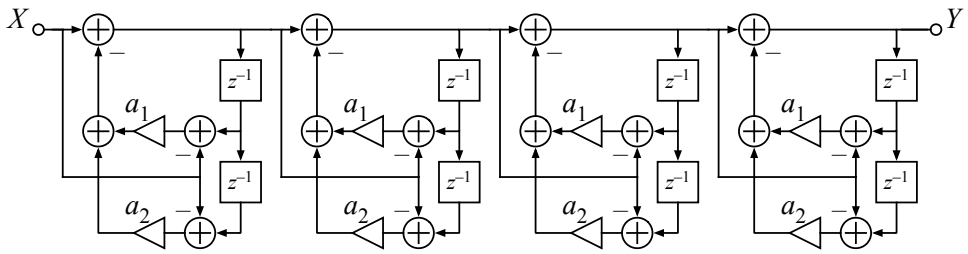


Figure 9.16: A 4th-order APGF constructed as shown, using the modified direct-form two-pole stage of Figure 8.18, always has unity gain at DC, no matter how the poles are moved via the a_1 and a_2 coefficients.

$H_A(z)$, by analogy with the all-pole resonator with transfer function $H_A(s)$ of Section 8.2:

$$H_A(z) = \frac{cz}{z^2 - 2az + (a^2 + c^2)}$$

which has a DC gain (at $z = 1$) of

$$H_A(z)|_{z=1} = \frac{c}{1 - 2a + (a^2 + c^2)}$$

This DC gain can be problematic, as it varies with pole damping, which varies a and c proportionately to each other. For that reason, we sometimes use other forms, such as a direct form in which the input signal can be connected in such a way that the DC gain is unity, independent of the coefficients. For the direct-form realization, the same transfer function, except for a gain, is written in terms of the coefficients as:

$$H_A(z) = \frac{z^2 A}{z^2 + a_1 z + a_2}$$

The DC gain here can easily be made equal to unity by setting the gain to $A = 1 + a_1 + a_2$, as illustrated in Figure 8.18, in which the input gain is applied without any extra multipliers. Then the all-pole gammatone filter (APGF) made from such stages, as shown in Figure 9.16, will always have a stable low-frequency tail gain of 1.

EE Connection: Cascades of Similar or Identical Filters

Transfer functions and impulse responses of coincident-pole filters are not unique to the hearing field, where they are called gammatones; similar functions are known in other fields, including electronics, physics, and statistics.

Papoulis points out that the gamma-distribution impulse response shape arises, very nearly, for cascaded smoothing filters even with noncoincident poles, as the result of a *causal central limit theorem*. Fitting the form of a gamma distribution to such a system's impulse response can yield noninteger N when the poles are not equal, but using the nearest integer N still gives an excellent approximation, effectively representing the system as a cascade of N identical one-pole filters (Papoulis, 1962).

Much of the mathematics of these filters has parallels in the field of statistics. Karl Pearson (1916) developed a number of probability density functions with simple parameterizations and with properties useful in statistics problems. The power-gain frequency response curve of a complex gammachirp filter is exactly a Pearson type IV distribution. Its symmetric special case, the gammatone, is a Pearson type VII distribution, or a Student's t-distribution, which has the Cauchy–Lorentz and Gaussian distributions as its low-order and high-order limits, corresponding to the universal resonance curve and the Gaussian filter.

Cascades of identical resonant filter stages were analyzed for use in wireless television and radio systems in the 1940s (Eaglesfield, 1945; Tucker, 1946), with the observation that their envelope step responses are *incomplete gamma functions* (integrals of the gamma distribution), connecting them, at least tenuously, to the current *gammatone* terminology. Tucker's cascade implementation is shown in Figure 9.17.

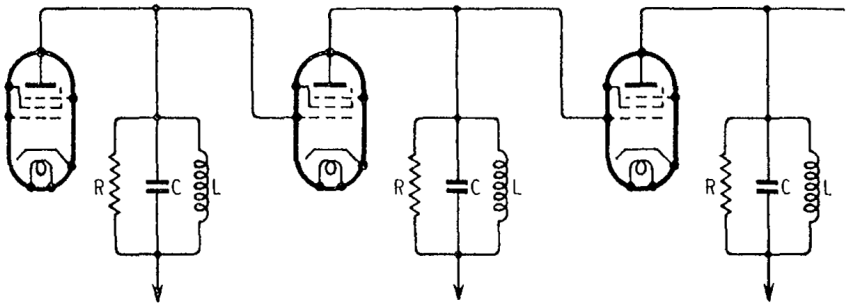


Fig. 1. *Cascade of tuned circuits without mutual coupling, and with blocking capacitors or coupling windings omitted for simplicity.*

Figure 9.17: In his gammatone-like filter cascade, made of parallel-RLC resonators with buffer amplifiers between them, Tucker (1946) represented the buffer amplifiers as vacuum pentodes. This reduction of the mathematical topic to a concrete realization would have made it easy to grasp for engineers of that time, who were familiar with such slightly-abstracted schematics. The pentodes act as transconductors, converting a grid voltage to a plate current, and the filter is the impedance of the parallel circuit, which converts the plate current out of the pentode to a grid voltage into the next pentode. At DC, the inductor shorts the current to ground, making a zero at DC, so the cascaded filters are like our filter C of Chapter 8. [Figure 1 (Tucker, 1946) reproduced by permission of SJP Business Media.]

Chapter 10

Nonlinear Systems

> These results indicate that cochlear mechanics incorporates an essential nonlinearity, so that linear superposition for neighboring spectral components does not apply even at low sound levels.
>
> — "Auditory nonlinearity," J. L. Goldstein (1967)

In this chapter, we relax the constraints of linearity and time invariance. We let systems be time varying and level dependent as a way to incorporate some nonlinear phenomena in hearing. We also touch on concepts of nonlinear system description, such as the Volterra series, that make connections between measurements on linear and nonlinear systems, and on some examples of nonlinear systems.

Nonlinear systems cannot be completely characterized by their responses to sine waves; nevertheless, they are often described in terms of their responses to sine waves of various amplitudes, and to pairs of sine waves, using several different kinds of measurements and plots. We compare several of these in terms of how the nonlinearities manifest themselves in the plots.

We also discuss how nonlinearities can complicate sampling and aliasing considerations.

Nonlinear system responses in hearing are typically described relative to a *characteristic frequency* (CF), the frequency at which the system is most responsive, or most sensitive, at low levels—the frequency with the lowest *threshold*. It is analogous to the center frequency of a bandpass filter, but the frequency of greatest response or gain can change with level, so the linear filter analogy needs to be used carefully. For any place of measurement in the cochlea or the auditory nervous system, there may be a well-defined CF, but that CF is typically *not* the frequency of greatest response, except at very low levels.

10.1 Volterra Series and Other Descriptions

The output of a linear system is its input convolved with its impulse response. If a system is not too far from linear, then using a linear convolution model plus some correction terms can be a useful description. The *Volterra series* is such a description (and the *Wiener series* is another, closely related, which we will not discuss).

The Volterra series expresses the output as a sum of terms computed from *Volterra kernels* of different orders, one of which, the first-order Volterra kernel, is the impulse response of a linearization of a nonlinear system. This first-order term is typically the most important term in a Volterra series that describes a nonlinear system, if the system is approximately linear for some range of low-level inputs.

Before the first-order kernel, or impulse response $k_1(u)$, there may sometimes be a zero-order kernel k_0, which is simply a constant. A linear system has zero output for zero input, so if one has a system with an offset in its output, the k_0 term can take care of that.

The next Volterra kernel, of second order, describes how signals can interact as products. It adds to the system output signals proportional to the square of the input, or more generally products of shifted versions of the input:

$$y_2(t) \propto x(t-u_1)x(t-u_2)$$

When an input multiplies itself this way, a sinusoid will generate a DC term proportional to the squared amplitude, as well as a double-frequency term. When multiple frequencies are present in the input, the multiplication will generate new components at sums and differences of input frequencies, as can be verified via formulae for sine and cosine of sums and differences. The second-order Volterra kernel $k_2(u_1, u_2)$ is the weighting to be applied to such terms, as a function of the two time offsets from the output time. With zero-, first-, and second-order terms included, the system output is approximated as:

$$y(t) = k_0 + \int_{-\infty}^{\infty} k_1(u)x(t-u)du + \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k_2(u_1,u_2)x(t-u_1)x(t-u_2)du_1du_2$$

Notice that if the input x doubles, the contribution of the second-order term quadruples. This squaring relationship from input to output is characteristic of a second-order nonlinearity. The third-order term has a cubing relationship from input to output, and so forth.

The second-order version is useful for modeling the rate response of auditory neurons, at least over a limited range of input level (Young and Calhoun, 2005). It has also been used for characterizing the response of higher-level neurons that don't synchronize to the phase of sound waves, as *spectro-temporal receptive fields* (Eggermont, 1993; Pienkowski et al., 2009), though approaches based on filterbank analysis can work better (Eggermont et al., 1983).

By restricting the kernels to be *delta functions*, we can get rid of the integrals, and model *point nonlinearities* (also known as *memoryless nonlinearities*) as simple polynomial functions of $x(t)$:

$$y(t) = k_0 + k_1x(t) + k_2x(t)^2 + k_3x(t)^3 + \ldots$$

This restricted form of Volterra kernel, a *Taylor series* for an instantaneous input–output function, is much more tractable, though much less powerful, than the general form (Dawson and Lee, 2004). The polynomial form is enough to explain the frequencies of the combination tones that result when the input consists of several sinusoids: the quadratic term leads to sums and differences of the stimulus frequencies, *quadratic difference tone* (QDT) frequencies $f_2 - f_1$ and $f_1 + f_2$, and the cubic term leads to *cubic difference tone* (CDT) frequencies such as $2f_1 - f_2$ for stimulus frequencies f_1 and f_2.

Though they correctly predict the frequencies of important distortion products, both the polynomial nonlinearity and the full Volterra model lead to very misleading predictions concerning the growth of combination tone level with stimulus level. They suggest that the levels of the distortion products will grow much faster than the level of the stimulus, which can be true in some systems, but is not generally true in the ear. Nonlinear curves that approach a constant or linear asymptote, instead of growing as a power of x, can be more appropriate descriptions of point nonlinearities for many purposes; modeling these by polynomials is not very effective.

A cascade of alternating linear and memoryless nonlinear systems is commonly used in radio design, and in hearing models (Pfeiffer, 1970; Duifhuis, 1976; Swerup, 1978; Duifhuis and van de Vorst, 1980; Pick, 1980; Goldstein, 1990; Eggermont, 1993; Goldstein, 1995; Meddis et al., 2001). In particular, the "LNL" (linear–nonlinear–linear) or "sandwich" configuration—a linear system, a memoryless nonlinearity, and another linear system, cascaded in that order—has been often analyzed (Middleton, 1948; Davenport, 1953) and has been shown to have the special property that when the input has a Gaussian distribution, the input–output cross-correlation function will be as if the nonlinearity were omitted, except for a constant factor (de Boer, 1976a; Korenberg and Hunter, 1986). This property allows the identification of parameters of a "bandpass

nonlinear" (BPNL) model of cochlear filtering when auditory nerve responses to Gaussian noise are recorded.

For some kinds of nonlinear behavior, especially involving slow level-dependent parameter variation, none of these methods are very useful; Harte, Elliott, and Rice (2005) concluded "that historical attempts to use functional modeling (i.e., Wiener or Volterra series) may be ill founded, as these methods are unable to represent level-dependent nonlinear systems with multivalued characteristics of this kind." For such system behaviors, models that explicitly vary the parameters of a subsystem, based on the output of another subsystem, can be effective. A classic example of such a system is the *automatic gain control* in a radio receiver, which we discuss in the next chapter and adapt to a cochlear model in later chapters.

10.2 Essential Nonlinearity

Tartini's tones (see Chapter 1) are the perceptual correlates of third-order distortion products generated inside the cochlea. The Taylor series or Volterra kernel approach suggests that doubling the input would make the contribution of such a distortion term increase by a factor of eight—cubicly. In electronic audio amplifiers that are designed to be linear, the kind of third-order distortion that limits the useful amplitude range is sometimes effectively modeled by such methods: at high levels, as the peaks of the signal waveform get distorted, the distortion rises rapidly until at some point the signal quality is not acceptable.

In hearing, however, the third-order distortion products that arise in the cochlea behave somewhat differently—the Volterra approach does not lead to a useful model of it. These cubic difference tones (CDTs) or combination tones (CTs) grow only slowly with signal level, unlike the result of a third-order Volterra term. CDTs are detectable even at very low sound levels. CDTs measured in the ear canal are even used as a diagnostic for normal cochlear function (Janssen and Müller, 2007).

This persistence of cubic nonlinear distortion down to very low levels can be seen as diagnostic of a good model. Goldstein (1967) termed this behavior *essential nonlinearity*, saying, "Essential nonlinearity is a description of the fact, not a hypothesis, that the relative level of the cubic CT is almost independent of the stimulus level."

Goldstein and Kiang (1968) make the point that the essential nonlinearity of the system means that we need to be careful about how we apply linear system concepts:

> Spectrum analysis is a key feature of all modern functional models of auditory signal processing that have been used to provide quantitative theoretical descriptions for psychophysical and physiological phenomena. Combination tones are pertinent to these models, because their properties challenge the generality of the classical assumption that auditory spectral filtering is an essentially linear, time-invariant process.

The pursuit of a realistic model of the growth of distortion tone level with stimulus level has remained an important driver in cochlear modeling (Goldstein and Kiang, 1968; de Boer, 1976b; Trahiotis and Robinson, 1979; Brown, 1993; Eguíluz et al., 2000; Ospeck et al., 2001; Duke and Jülicher, 2003; Roberts and Rutherford, 2008; Duke and Jülicher, 2008). An appropriate level dependence of distortion, as a simple emergent property, is a key feature of the adaptive nonlinear filter-cascade models that we develop in this book.

10.3 Hopf Bifurcation

Recently, explanations of the cochlea's nonlinearity in terms of a *Hopf bifurcation* in a nonlinear oscillator have been popular (Brown, 1993; Eguíluz et al., 2000; Ospeck et al., 2001; Roberts and Rutherford, 2008; Duke and Jülicher, 2008). A *bifurcation* is a qualitative change in a system behavior at some point in the variation of a continuous parameter value. Specifically, a Hopf bifurcation, or *Poincaré–Andronov–Hopf*

bifurcation, is the change in behavior of a resonator-like system at the point where the low-level damping parameter changes sign: the system is a stable filter for positive damping, and an unstable filter, or oscillator, for negative damping. In a linear system, damping less than zero characterizes an unstable system, with poles in the right half of the *s*-plane, such that the amplitude of the resonance will increase exponentially with time, rather than decay. In the context of nonlinear systems, low-level damping refers to the damping of a first-order Volterra-kernel model for linearization about a very low level; a system may be unstable for very small input and output levels, yet stabilize at higher levels, if it contains a compressive nonlinearity; it may then exhibit a stable oscillatory behavior known as a *periodic limit cycle*.

The theory is that in the ear, such systems have parameter values such that they are "poised" at the edge of stability, at the Hopf bifurcation, such that even zero or very small inputs will drive the output high enough that the nonlinearity will kick in and stabilize the system at a finite but nonzero output pattern. This behavior is typically represented using a damping factor that includes a term proportional to the square of the output amplitude or velocity, such that more output causes more damping added to the initial zero or slightly negative low-level damping. This approach gives rise to a roughly cube-root-compressive input–output level response, as well as to combination tones even at very low input levels—an essential nonlinearity.

Exactly how these oscillators integrate with a traveling-wave view of the cochlea is not usually made clear. As Eguíluz et al. (2000) say about the nonlinear-oscillator hair-cell model,

> Because the cochlea is a complex geometrical structure traversed by nonlinear waves, relating the contribution of individual hair cells to the behavior of the entire organ remains both a theoretical and an experimental challenge.

In spite of this, they argue that the single-resonance Hopf oscillator is a likely model of what's going on in the cochlea (they call three observable effects "three essential nonlinearites," but they're all effects of one mechanism):

> We have shown that tuning to a Hopf bifurcation can account for three well-documented essential nonlinearities of the ear: compression of dynamic range, sharper cochlear tuning for softer sounds, and generation of combination tones. The great advantage of the regenerative tuning strategy is that it requires a minimal number of active elements; because the tuner and the amplifier are one and the same, this mechanism is evolutionarily accessible.

However, the "minimal number of active elements" of this approach is also its weakness: it tries to explain the behavior of a distributed system via a low-order local model. In this approach, the tuning gets much too sharp at high gains, or low levels, as the filter gets all of its gain from a high-Q single-tuned resonance; response plots show this effect clearly (Eguíluz et al., 2000).

The Hopf or critical-oscillator concept has been integrated into a traveling-wave approach by Duke and Jülicher (2003), but the results still get a way-too-sharp response at low levels, as the active oscillator acts too locally. Another traveling-wave integration by Magnasco (2003) has a more distributed effect, but still has unrealistic sharp transitions. Many authors comment on the fact that the bandwidth of a Hopf-based model decreases inversely with the gain, but they seldom acknowledge that the resulting narrow bandwidths are very far from any bandwidth data observable in hearing experiments.

As described in the next section, the same kind of local damping nonlinearity, operating in a cascade of filters, can model the same set of nonlinear effects and more, even without being poised near the bifurcation at zero damping, and does not have the unrealistic narrow bandwidth problem.

10.4 Distributed Bandpass Nonlinearity

A model with nonlinearity distributed over multiple filter stages can achieve high gain without the too-narrow tuning of the Hopf models. It better integrates with how waves propagate in the cochlea, so is a better way to apply the Hopf nonlinearity idea to hearing models. In such models, the operating point never needs to be very close to the critical infinite-gain bifurcation point, since plenty of gain is available as the product of several moderate stage gains. We develop such a model in Chapter 14 and subsequent chapters of Part III.

The nonlinear filter-cascade model of Kim et al. (1973) (described in Chapter 9 without the nonlinearities) provides a good example of how nonlinearity can be incorporated in filter stages. It uses instantaneous nonlinearities embedded inside the filter stages, increasing the damping instantaneously with the squared local velocity. It includes ten cascaded stages, each essentially the same equation as a Rayleigh or Van der Pol oscillator (Duifhuis, 2012), with damping proportional to squared velocity, but with a significantly positive small-signal damping limit so it stays far on the stable side of the bifurcation. Their nonlinear second-order differential equation for stage i (for $i = 1 \ldots 10$, with $x_0(t)$ being the input) is:

$$\ddot{x}_i(t) + 2D_i\left[1 + \eta\dot{x}_i^2(t)\right]\dot{x}_i(t) + \omega_{0i}^2 x_i(t) = Cx_{i-1}(t)$$

The bracketed factor $\left[1 + \eta\dot{x}_i^2(t)\right]$, which is positive and increases as the response increases, multiplies the small-signal damping D_i in each stage. The quadratic velocity term $\dot{x}_i^2(t)$ leads to a cubic distortion nonlinearity because it multiplies the velocity $\dot{x}_i(t)$.

Kim et al. listed nine different nonlinear phenomena observed in basilar membrane and auditory nerve response that were qualitatively reproduced by this nonlinear filter-cascade model; from their abstract:

> This model, which behaves effectively linearly at low levels and nonlinearly at high levels, shows that a *single* nonlinear system is adequate to account for the following *frequency-dependent* nonlinear phenomena of the peripheral auditory system: (1) limiting of the output level; (2) decrease of Q with increasing input level; (3) decrease of the most effective frequency with increasing input level; (4) changes in phase angle of the output with input level; (5) changes in shape of the click response waveform with input level; (6) two-tone suppression with $f_1 \approx CF$ and $f_2 > CF$; (7) generation of the combination tone $2f_1 - f_2$ in response to two tones $f_1 < f_2$; (8) "amplitude" nonlinearity in response to click pairs; and (9) "temporal" nonlinearity in response to click pairs.

The cubic nonlinearity in their model stages still led to a too-fast growth of distortion with level, and no high-level linear limit, but their idea of a single nonlinear system made by *modulating the damping* in cascaded linear filter stages survives in modern models, including the ones developed in this book. Besides the memoryless nonlinearity, a slower nonlinearity that adapts the filter stage damping, based on smoothed feedback from the output level, can help to spread the nonlinearity's effect over a wider input dynamic range, and thereby help to make the combination-tone level dependence more consistent with "essential nonlinearity" observations.

The filter-cascade approach, with level-dependent control of damping, behaves a lot like the Hopf bifurcation approach, in the sense that the peak gain and bandwidth vary with the signal level, and combination tones are generated, dependent on the compressed output level. But the Hopf nonlinear oscillator is just a single two-pole stage modified to be nonlinear, so the bandwidth gets way too narrow when the gain gets high. A cascade of many stages gives a more moderate range of bandwidth variation with gain, as plots in Chapter 9 show, and relates better to cochlear traveling waves, as we discuss in Chapter 12.

10.5 Response Curves of Nonlinear Systems

In this section, we discuss the relationships between frequency–threshold curves (FTCs) and transfer functions, and other ways to show the tuning and compression of a nonlinear system. These methods mostly rely on the response to sinusoids, but at a range of different levels. Methods that consider the response to pairs of sinusoids are also introduced.

A nonlinear system, in which the response to a sinusoid is not simply proportional to its input amplitude, can be measured and characterized in a variety of ways, depending on whether the input level, the output level, or the frequency is held as a constant parameter, as illustrated in Figure 10.1. In these plots, frequencies are normalized as octaves of deviation from the CF (frequency of greatest response at very low input level).

We sometimes find interpretations of hearing data, based on sinusoidal stimuli, that would make sense if the system being measured were linear, but are not sensible in light of what's actually going on in the nonlinear auditory system measured over a large range of sound levels. For example, many experiments show rather sharp FTCs, for auditory nerve fibers or for the mechanical response of the cochlea. From such sharp curves, it is sometimes incorrectly inferred that the cochlear mechanical response as a function of frequency or place is very localized—or that there's a mysterious *second filter* needed when other measurements show that the response versus frequency or place is fairly broad compared to the FTCs (Cooper et al., 2008).

By examining the response of nonlinear filter systems, we can form a good understanding of why they look sharp when described by FTCs, but appear to be rather unsharp when described by transfer functions, or by first-order Volterra kernels, or by other linear-system or nonlinear-system characterizations.

Capranica (1992) has cautioned against the common practice of characterizing auditory neurons by their tuning curves, especially when the information of interest is more in the time domain:

> In the auditory system inhibition is ubiquitous as it is in vision and chemosensation. The fact that energy in part of a signal can reduce a neuron's excitatory response to the overall signal is a highly nonlinear operation. This should serve as an obvious warning that linear operators, such as Fourier transforms, may not be appropriate descriptors, especially for the central auditory system. When a signal is produced, it is an event in time. The one specialization that distinguishes the auditory system from all of the other modalities is time, not the power to resolve frequency. Its remarkable forte is to process rapid changes in the time domain.

The different parts of Figure 10.1 show different sets of curves that represent the same example nonlinear system. These curves are chosen to resemble, qualitatively, the shapes of response curves from a live cochlea, simplified and parameterized as three-slope transfer functions (Rhode, 1978). The frequency-based measurements and illustrations here do not address Capranica's concern about time-domain features, but do help to show that even for sinusoids the linear-system and frequency-domain views need to be taken with a grain of salt. It can be seen that the FTCs (fixed output level, iso-response curves) are sharp even when the response versus frequency (fixed input level, iso-intensity curves) are rather broad. Corresponding plot types from real cochlear measurement data are shown in Figure 10.2, where the difference in apparent sharpness of tuning is even more pronounced.

An obvious lesson from these curves is that in a compressive nonlinear system, the iso-response tuning curves are sharper, and the iso-intensity curves are less sharp, than the underlying gain-versus-frequency linearized transfer functions. This observation was invoked very soon after the discovery of nonlinear cochlear mechanical response by Rhode (1971) to explain why his response curves were still not as sharp as typical neural tuning curves. Rather than hypothesizing a sharpness-enhancing second filter, as others were doing, he says:

> An alternative explanation may be advanced on the basis of the observation that the basilar membrane appears to vibrate nonlinearly. If one plotted a transfer function for the basilar mem-

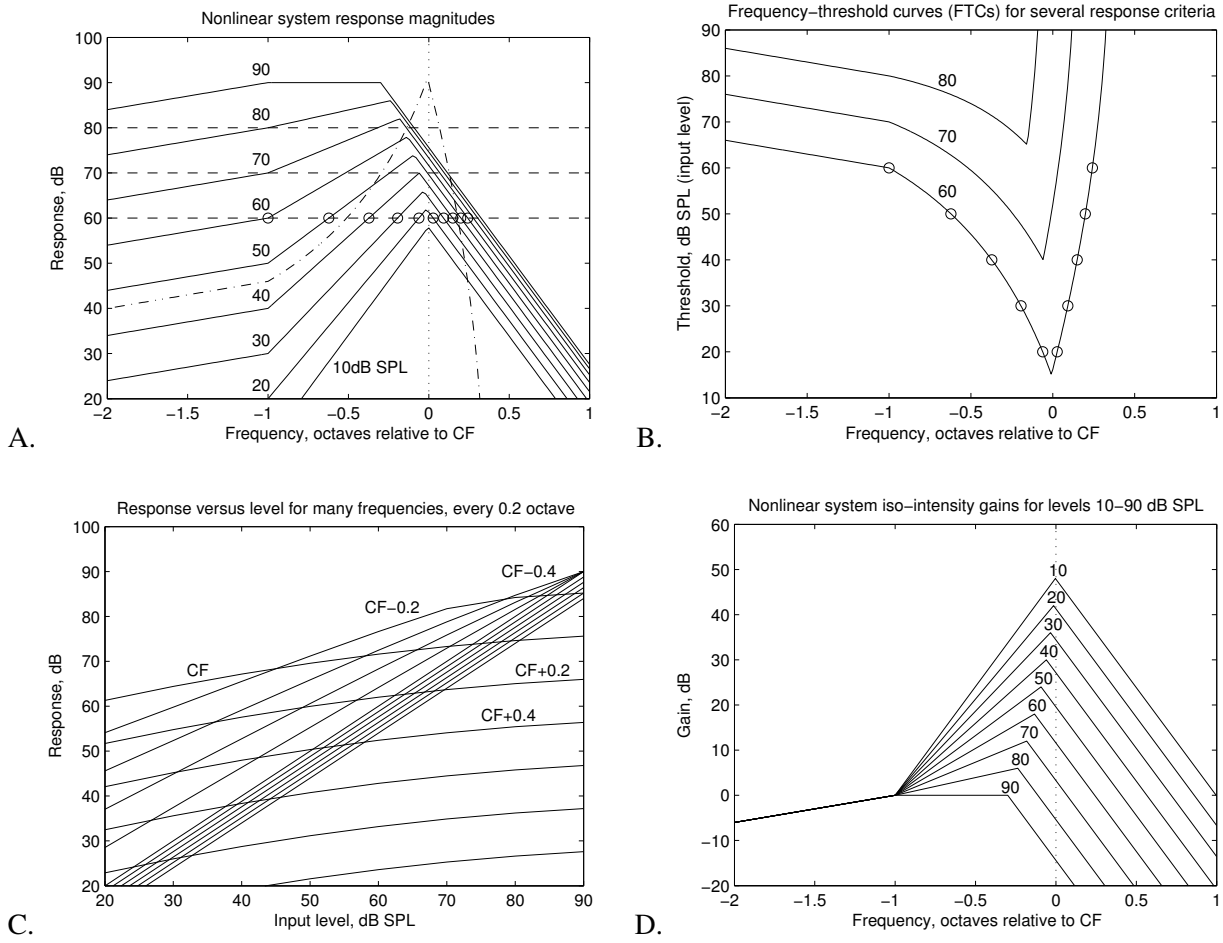


Figure 10.1: Four views of the response of a nonlinear system to sine-wave input at various levels. If we specify a system with the curves in one plot (completely enough), then the other three are generated from it mechanically.

A. Iso-level or iso-intensity curves plot the response versus frequency when the input level is held constant (at levels indicated by parameters on the curves); the dotted vertical line indicates the CF, the frequency of greatest sensitivity at low levels.

B. Iso-response curves, or frequency–threshold curves, plot the input level needed for a given output level, or response threshold, criterion; the 60, 70, and 80 dB response criteria correspond to levels shown by horizontal dotted lines in A, and corresponding points on the 60 dB curve at circled. The dash-dot line in A is a reflection of the 60 dB curve in B, to indicate how much "sharper" the iso-response curve is than the iso-level curves.

C. Iso-frequency curves plot the response level versus input level, for various frequencies. For frequencies near or above CF (CF, CF + 0.2, CF + 0.4), the system is very compressive: the curves have a low slope.

D. Iso-intensity gain curves resemble the magnitude frequency responses of linear systems, except that they are different at different input levels.

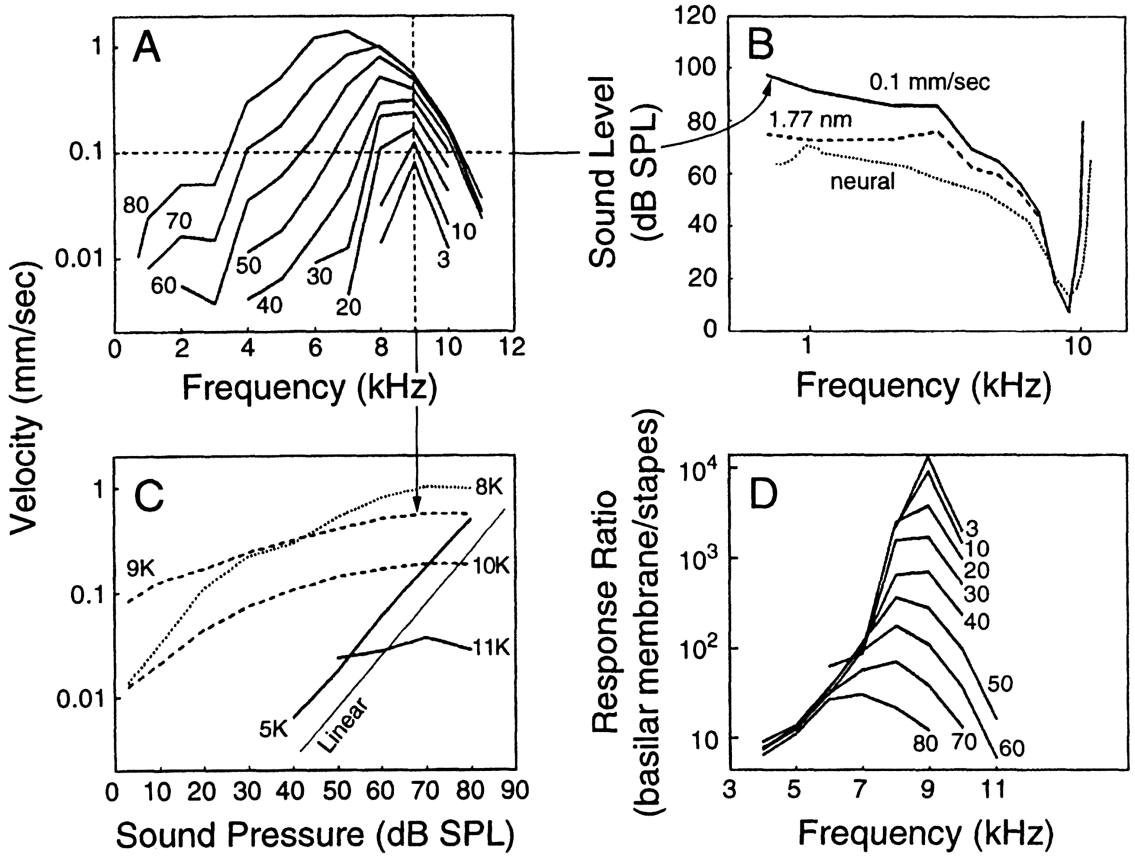


Figure 10.2: Measurements on live cochleas, plotted in the four ways described in Figure 10.1, showing a hugely nonlinear response. The response is quantified in terms of basilar membrane velocity, using laser doppler velocimeter data from Ruggero (1992). Though panel A shows a broad response region at moderate and high levels, panel B shows mechanical frequency–threshold curves (FTCs) at least as sharp as neural FTCs. Panel C shows response approaching linear for frequencies well below CF, and most compressive for frequencies above CF. Panel D shows a gain change at CF of more than 50 dB. [Figure 5.8 (Geisler, 1998) based on Figure 1 (Ruggero, 1992) reproduced with permission of Dan Geisler, Mario Ruggero, and Elsevier Science and Technology Journals.]

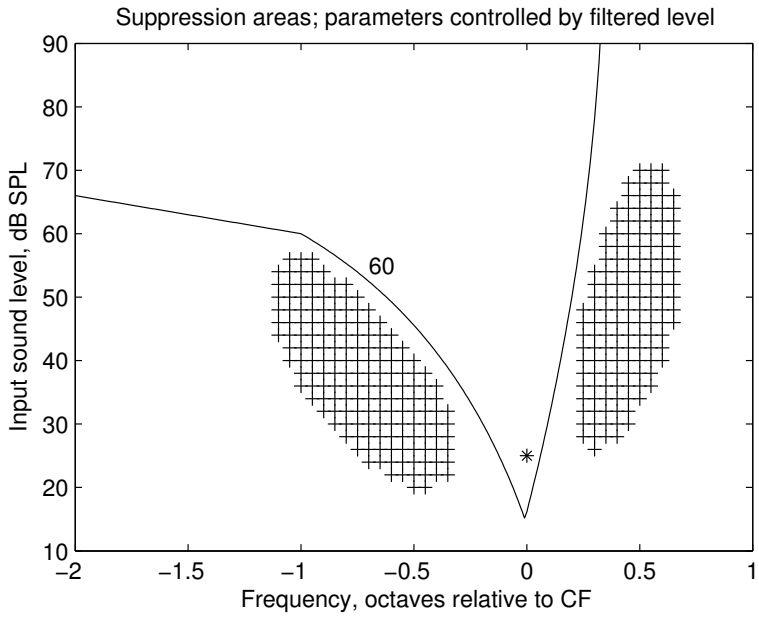


Figure 10.3: Two-tone suppression areas. When a first tone is presented at the frequency and amplitude signified by the '*' (near CF, at a low but detectable level, as shown here relative to the 60 dB response criterion), the system output can actually be reduced by addition of a second tone in the above-CF or the below-CF suppression area. The shape and size of these areas depends on the suppression criterion (here, a 1 dB drop in total output power), and on how the system's parameters depend on the spectrum of the input (here via the power detected in a broad curved filter centered at CF).

> brane using data which were collected using SPLs that resulted in a constant displacement of the basilar membrane, the high-frequency slope of the transfer function would increase dramatically to –150 dB/oct to –300 dB/oct.

In spite of Rhode's explanation, and others since then, the confusion between such curve types has appeared many times since. His use of the term "transfer function" for the "constant displacement" curves was unfortunate, suggesting the wrong connection to linear system theory, contrary to the point he was trying to make. I have tried to help resolve this confusion (Lyon, 1990), but it is still a topic of conversation today (Eustaquio-Martín and Lopez-Poveda, 2011). Almost every book on hearing has at least one inapt comparison of iso-response tuning curves to filter transfer functions.

10.6 Two-Tone Responses

Nonlinear aspects of the ear are also revealed in responses (at a given location) to pairs of tones. When two sinusoids are presented at once, the response can be surprising: the response level can be less than with one of the two tones alone, a phenomenon known as *two-tone suppression*. Tuning curves can be augmented to show *suppression areas*, as in Figure 10.3, where the addition of a second tone, above or below the CF, will reduce the response to a low-level tone near the CF, and will not itself cause an above-threshold response. Neural tuning curves typically show suppression areas both above and below CF, while mechanical tuning only shows suppression from suppressors above CF (Versteegh and van der Heijden, 2013), a difference for which no really good explanation is available.

Tones outside these areas either have little gain-reduction effect on the peak gain of the filter, or add enough response of their own to dominate the suppression of the first tone such that there's no net decrement in output level. The shape of these two-tone suppression regions provides extra information, not visible in the four sinusoid-based plot types discussed previously, about how the filter's peak gain is affected by signal energy away from CF.

A second important two-tone nonlinearity is the generation of distortion tones, or intermodulation products (such as frequencies $f_2 - f_1$ and $2f_1 - f_2$ from tone frequencies f_1 and f_2). Such intermodulation products, or combination tones, are found in the mechanical response of the cochlea, and are often audible.

Measurements with single sinusoids give little clue to an underlying mechanism or model for the nonlinear responses; two-tone measurements can add to our understanding of such a system. Consider two extreme alternative types of nonlinearities: memoryless and parametric. If the nonlinear system is implemented as a parametric linear system, where the parameters are set by the level of the sound, then it will not generate distortion tones—assuming the parameters change slowly enough to be regarded as constants. Alternatively, if the system has fixed parameters but is compressive due to one or more memoryless compressive nonlinearities, at its output or internally, then the output will likely include strong distortion products, such as cubic difference tones. When we construct models of the cochlea, we'll see that including both types of mechanism is useful in making a model that fits a wide range of data over a wide range of sound levels.

10.7 Nonlinearity and Aliasing

If we satisfy the Nyquist criterion $2B < f_s$, signals containing only frequencies less than the bandwidth B are unambiguously represented by their samples at sample rate f_s. Are we then safe from aliasing when processing such discrete signals? Unfortunately, no. For linear systems, no new frequencies are generated, so we're safe in linear filtering of such sampled signals; but for nonlinear operations, it's more complicated. In particular, a second-order Volterra kernel (such as a squaring operation) will generate new frequencies equal to sums and differences of frequencies in the input, including double-frequency terms. Where these new frequencies exceed $f_s/2$, they will alias; that is, they will appear in the output as other, lower, frequencies, potentially interfering with signals of interest. To avoid aliasing at the output of such a system, an extra factor of 2 in sampling rate is required. For more general nonlinearities, the problem is even worse. In practice, we make a compromise when implementing nonlinear sampled systems, using a sample rate higher than the Nyquist criterion would imply, and tolerating some aliasing. See the "AM Radio Demodulation" example box.

In modeling the cochlea, an important strong nonlinearity is that of the inner hair cell, roughly a half-wave rectifier (HWR), which responds to motion of the basilar membrane, points on which we model as outputs of a cascade filterbank. The resulting signal, representing what's on the auditory nerve, needs at least a few kHz of bandwidth, and should be represented without much aliasing. The HWR generates distortion products of many orders (quadratic and fourth-order especially), so a compromise is inevitably needed. Part of the compromise in typical machine hearing systems is to not try to process the entire range of audible frequencies up to 20 kHz; for telephony, a bandwidth around 3.8 kHz or so is typical. For speech recognition, up to 7 kHz or so is helpful for distinguishing different consonant sounds. For music, higher frequencies are important to listeners, and perhaps to machine hearing systems. If we make a system with 20 kHz sample rate, then fourth-order distortion products of signals up to 4 kHz, which go up to 16 kHz, will alias to frequencies above 4 kHz; third-order distortion products of signals up to 6 kHz, which go up to 18 kHz, will alias to frequencies above 2 kHz. Depending on what bandwidth we want to protect from what order of distortion product, we can find what highest frequencies we can analyze. In practice, we might tolerate a lot of aliasing in higher-frequency channels, to keep the cost down, and process signals up to around 7 kHz with a 20 kHz sample rate; second-order distortion up to 14 kHz will alias to 6 kHz and above, so it will stay above the few kHz band

that we expect to preserve, while third- and fourth-order distortion will easily alias into the low frequencies that we want to use.

Example: AM Radio Demodulation

Consider an amplitude-modulation (AM) radio receiver as an example—a desired sound signal is broadcast as variations in the amplitude of a radio-frequency carrier wave, and we want to get the sound back. Digital radios sometimes work by having an analog continuous-time front end that down-converts radio frequencies around the channel of interest to a fixed *intermediate frequency* (IF), and then sampling the IF signal and doing the rest of the processing digitally. Suppose the IF frequency (the center of the band, to which the carrier frequency is converted) is 30 kHz, and we sample at 100 kHz. For AM radio channels spaced at 10 kHz, the signal bandwidth is 5 kHz (possibly somewhat higher, but use 5 kHz as an example), and there are two sidebands, so we care about frequencies in 25–35 kHz. We can start with a digital bandpass *IF filter* to pass these frequencies of interest and remove everything else. If we then detect the modulated signal's amplitude as a power, by squaring, we generate second-order difference signals, between the carrier and the sideband components, in the frequency range 0–5 kHz, which are the demodulated audio signal components that we want. We also generate sum and double-frequency components in the range 50–70 kHz; these are aliases of frequencies in the range 30–50 kHz, but that doesn't bother us much, as we can remove those alias frequencies by using a lowpass filter, without bothering the low frequencies that we want.

But this square-law detector is not really what we want, as it distorts the audio, which is proportional to amplitude, not power. We could next take a square root (since the powers are always positive, coming from amplitudes modulated up and down from the carrier level that represents zero sound signal). But square roots are expensive. So instead of square-law, we might use an absolute value, or full-wave rectification operator. This nonlinearity generates fourth-order and sixth-order distortion, and so on, in addition to the second-order. So it makes a band at 100–140 kHz, which are aliases of 0–40 kHz, and 150–210 kHz, aliasing to 0–50 kHz, etc., thereby adding some unwanted junk into the audio band of interest. In situations like this, we may want to use a much higher sample rate, or a different demodulation technique, to get a cleaner result. The biggest aliased components are from even multiples of the carrier itself (60, 120, 180, 240, 300 kHz, aliasing to 40, 20, 20, 40, 0 kHz) which are fixed and easy to keep outside the band of interest up to eighth order, so it's not terribly bad. But if the carrier is off from 30 kHz by just 10 Hz, the tenth-order nonlinear component will alias to 100 Hz, and will make an audible hum.

These kinds of issues, particularly *intermodulation product frequencies* between sample rates and carriers, are analyzed carefully in radio design, and are sometimes relevant in machine hearing as well, especially if a high-quality reconstructed sound is needed, as in a hearing aid. Wherever a strong nonlinearity is used, it's a good idea to consider where harmonics of a signal will alias to. Softer nonlinearities such as time-varying gains, if carefully applied, are less likely to cause audible distortion and aliasing, since they generate much smaller distortion product amplitudes.

10.8 Cautions

We need to be careful when embracing linear systems concepts, sinusoidal analysis, and such. We've seen how nonlinearities complicate the simple picture that linear systems theory gives us. We'll see that the ear has important aspects in which it is nearly a linear system, and other aspects in which the nonlinearities are key to its behavior.

The field of hearing research still suffers from occasional inappropriate comparisons of iso-response and iso-intensity views of cochlear frequency response. We can break free of this confusion by explicitly recog-

nizing the importance of nonlinearities in the system description, so that the effects of level dependence will not be hidden in incompatible linearized views.

In the next chapter, we investigate the theory and implementation of automatic gain control, a key nonlinear tool in systems that must deal with a wide dynamic range of inputs.

Chapter 11

Automatic Gain Control

> In recent years, devices for the automatic control of gain have increased in importance in various areas of amplifier technology. One class of such devices is based on the following principle: a portion of the output signal current of a valve amplifier is extracted, amplified and fed to a rectifier; the resulting rectified signal voltage is then used to vary the grid voltage of an amplifier valve. In this manner an increase in output power leads to a reduction in gain.
>
> — "On the Dynamics of Automatic Gain Controllers," Karl Küpfmüller (1928)

I have long viewed the automatic gain control (AGC) function as one of the most important, and tricky, parts of modeling the function of the cochlea (Lyon, 1982, 1990). To understand or design this important level-adaptive function, one must have an appreciation for the dynamics of feedback control, in a highly variable nonlinear context.

In this chapter, I provide the basic background and analysis techniques that our cochlear models will draw on. In particular, I show how the use of output amplitude to control the damping factors of cascaded resonators can be modeled as a robust feedback control system that compresses a wide input dynamic range into a narrower output dynamic range, by examining this approach in the context of a fairly general single-channel AGC formulation.

11.1 Input–Output Level Compression

Systems that use feedback from a detected output level to adjust their own parameters to keep the output level from varying too much are called automatic gain control systems. Such systems are inherently nonlinear, with a compressive input–output function: when the input changes by some factor, the output level changes by a factor closer to 1. AGC has long been used and analyzed in wireless communication systems (Wheeler, 1928; Küpfmüller, 1928), including television (De Forest, 1942), and is an idea that has long inspired corresponding models in biological systems, including vision and hearing (Rose, 1948, 1973; Smith and Zwislocki, 1975; Allen, 1979). In hearing, AGC is an important aspect of nonlinear cochlear mechanical function, as Duck Kim (1980) points out:

> Nonlinear equivalent damping of cochlear partition, which increases with increasing response as assumed in models, produces a functionally useful automatic-gain-control effect by compressing the amplitude of cochlear-partition motion for moderate and high stimulus levels without excessive signal distortions. Such amplitude compression of the cochlear-partition motion may play a key role in achieving as wide a dynamic range of hearing as 100 dB by converting the wide

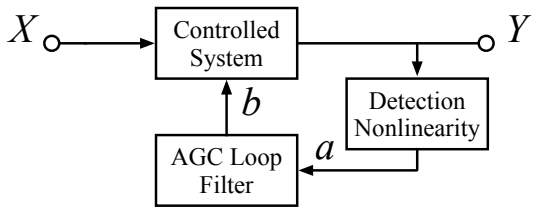


Figure 11.1: A system with automatic gain control (AGC). The loop filter output, b, can control any parameter of the controlled system that affects its gain. The loop filter, in combination with the properties of the controlled system and the detector, determines the dynamics of the response to a change in input level.

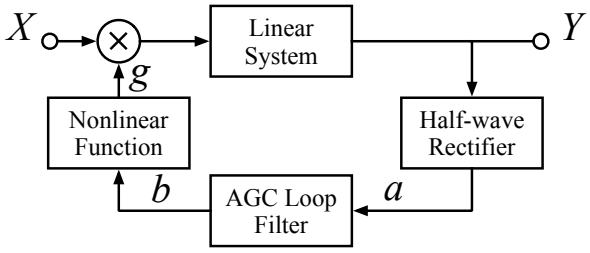


Figure 11.2: A loop with multiplicative gain is a tractable special case of the model of Figure 11.1. The "controlled system" is expanded as a linear system with a variable gain at its input, the gain g being a decreasing nonlinear function of the control parameter b. In this diagram, we also specialize the detection nonlinearity to a half-wave rectifier. The system can be approximately analyzed by linearizing the loop, treating it as a linear system with signal levels, as opposed to the signals X and Y themselves, as its input and output variables.

> range of acoustic signal amplitudes into a much narrower range of amplitudes of hair-cell cilia deformation.

The notion of "compression" can be appreciated from either of two directions: as a system with a high-level linear region with increasing gain at lower levels, or one with a low-level linear region with decreasing gain at higher levels. Engineers tend to think of a low-level linear behavior as the baseline. In the physiology of hearing, however, it is common, and sometimes more meaningful, to think of the baseline as a high-level linear limit, corresponding to a passive or dead cochlea, and an increasing gain at lower levels, corresponding to active amplification. The abstraction of level-dependent gain as an AGC system can work well from either point of view. Some systems have linear regions at both high and low levels, but this is not a requirement; some systems may have no linear region at all. In this chapter, we treat the engineer's conception, a system that approaches linear at low enough levels. When we apply this model in Chapter 19 to a cochlear model, we modify it to also have a high-level linear limit, by using a saturating detection nonlinearity.

11.2 Nonlinear Feedback Control

A fairly general system with AGC is shown in Figure 11.1; the loop filter controls a parameter b that could be a recipocal gain, but might be, for example, the damping factor of a resonator, or the damping in a distributed wave-propagation medium. It is easier to analyze AGC systems via the more specific form in Figure 11.2, where there is a particular detection nonlinearity, the half-wave rectifier, and the controlled system is modeled

On "Level"

The concept of *level*, frequently found as *intensity level* or *loudness level*, is usually expressed on a logarithmic scale, in decibels. In the 1960s, standards organizations actually began to *define* level to be the logarithm of the ratio of an intensity to a reference intensity, so that they could cast the decibel as a unit of level, making the dB behave more like a conventional unit than as a logarithm; for example, ANSI (1960) defines *level*: "In acoustics, the level of a quantity is the logarithm of the ratio of that quantity to a reference quantity of the same kind. The base of the logarithm, the reference quantity, and the *kind* of level must be specified." Most engineers have not been taught this definition of level, though, and use level more informally as a general notion of a measurement of how big a signal is, whether they represent it logarithmically or not.

In an automatic gain control loop, we typically feed back some measurement of output level to control the system gain. Some treatments in the literature assume that output level is measured logarithmically, but this model is difficult to get to work right at very low signal levels, so is more often avoided.

Wheeler (1928) speaks of "maintaining the desired signal level in the detector or rectifier," which is much like how we treat it here. That is, we let the detection nonlinearity (the rectifier) provide a signal that we take to represent level, with no prejudice about whether it is proportional to power, or amplitude, or log power, or something else.

In a real AGC system with signals representing sounds, level is a derived quantity, or even an abstraction, of what the system adapts to. A detector or rectifier produces a derived signal whose short-time average can be taken as level. But the rectified signal—whether positive part or absolute value—also contains fine temporal structure that is not part of what we call level. There may be no clean separation between the frequencies or time scales of level fluctuations and the frequencies or time scales of fine structure. But we can pretend.

as a linear system and a variable gain g, related to the loop filter output b by a monotonically-decreasing nonlinear function. Other detection nonlinearities, such as full-wave rectifier (absolute value), or square-law, or a rectifier that saturates at high amplitudes, could be used as well. The inner-hair-cell model that we introduce in Chapter 18 as part of our cochlear model is an example of the latter.

Referring to Figure 11.1 and Figure 11.2, we define the level of Y as the short-time-average value of a; that is, we let the detector define level:

$$\text{level}(Y) \equiv \text{mean}(a)$$

where the mean is taken over a time long enough to ignore the fine temporal details of signal Y, but short enough to allow for the analysis of the dynamics of the gain-control loop. Thus the output level, plus some high-frequency fine structure that we can mostly ignore, is the input to the linear loop filter in the feedback path.

Following Küpfmüller (1928), we analyze such nonlinear AGC feedback networks by first finding formulae for the steady-state relationship between the input level and the output level, and then considering small perturbations, in an analysis linearized about the equilibrium point.

11.3 AGC Compression at Equilibrium

In the case of the half-wave rectifier (positive part) as detector, as shown in Figure 11.2, the average value of a will be something less than the peak output amplitude, depending on the wave shape ($1/\pi$ for a sinusoidal output). If the linear system is more complicated than a unity gain, we need to know its transfer function and the spectrum of X to say how the levels of X and Y relate. We simplify by assuming that X and Y are narrowband signals, so we can treat the linear system as just a gain H (ignoring phase) that cascades with the

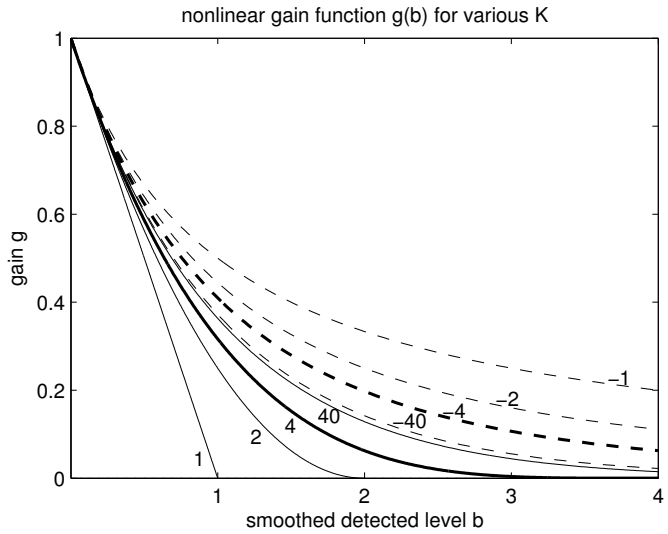


Figure 11.3: The family of nonlinear gain functions $g(b)$ used in the AGC analysis, for various values of the K parameter. All functions are near $1 - b$ at low levels, but their high-level behavior depends on the parameter. Negative K values are represented with dashed curves; $|K| = 4$ curves are bold, representing typical system choices. For high $|K|$, the functions approach $g(b) = \exp(-b)$.

gain g to relate the signals as well as their levels:

$$Y = gHX \qquad \Longrightarrow \qquad \text{level}(Y) = gH\,\text{level}(X)$$

Here we are implicitly defining level(X) as being measured the same way as level(Y), via a rectifier, and counting on the property of a full-wave or half-wave rectifier that multiplying its input by a factor multiplies its average output by the same factor; for other kinds of detection nonlinearity, or other definition of input level, the analysis can be suitably modified.

For shorthand in analyzing the AGC loop, we use bold symbols $\mathbf{x}$ and $\mathbf{y}$ to represent the corresponding levels, which can be time varying, and $\mathbf{x}_{\text{eq}}$ and $\mathbf{y}_{\text{eq}}$ for the equilibrium values of these levels, that is, when the input level is constant and the gain has settled to make the output level constant.

At equilibrium, the input and output *levels* are presumed constant, but the signals themselves are not; the fine structure of the output signal shows up in a, which will be following the positive parts of Y and will be zero during the negative parts. To follow the level, or envelope, of the output signal, the loop filter has to smooth away most of these rapid fluctuations in a, and drive the variable gain g with suitable dynamics; we analyze the dynamics below in Section 11.6.

For simplicity in the present analysis, we constrain the loop filter to have unity gain at DC, and we assume that it smooths a enough that we can treat b as constant at equilibrium, equal to the long-term mean of a: $b = \mathbf{y}_{\text{eq}}$. Using a nonlinear function $g(b)$ to control the gain, the equilibrium condition on the levels of X and Y is therefore:

$$\mathbf{y}_{\text{eq}} = g(\mathbf{y}_{\text{eq}})\,H\,\mathbf{x}_{\text{eq}}$$

which is easy to solve for input level as a function of output level even if the form of $g(b)$ is not known:

$$\mathbf{x}_{\text{eq}} = \frac{\mathbf{y}_{\text{eq}}}{g(\mathbf{y}_{\text{eq}})\,H}$$

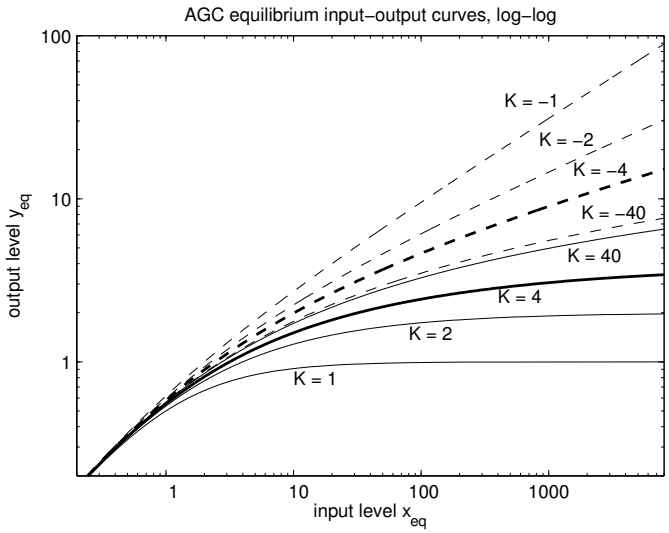


Figure 11.4: The input–output level curves for an AGC system with nonlinear gain function $g(b) = (1-b/K)^K$. Curve styles are as in Figure 11.3. At low levels, all of the curves are approximately linear; at high levels they compress to varying degrees. Negative values of K tend toward power law (root) compression, with straight asymptotes of slope $1/(1-K)$ in this log–log plot, while positive values of K cause compression toward a constant output level K (horizontal asymptotes). For high $|K|$, at the divide between positive and negative K, the high-level response approaches logarithmic compression (no straight-line asymptote).

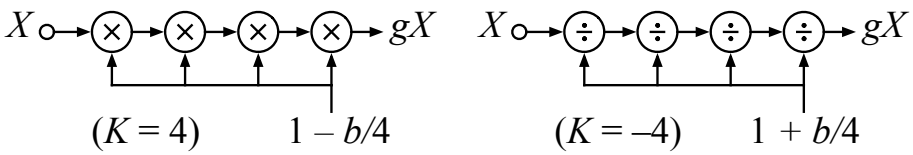


Figure 11.5: Variable-gain amplifiers (multipliers or dividers) can be cascaded to give a gain control over a fairly wide dynamic range, compared to the dynamic range of their control variable. Here the multipliers (left) correspond to our nonlinear gain function with $K = 4$, and the dividers (right) to $K = -4$.

For different nonlinear functions $g(b)$, we can solve for and plot the equilibrium input–output level relationship. For our analysis, we take the gain nonlinearity to be one of the family of curves illustrated in Figure 11.3:

$$g(b) = (1 - b/K)^K$$

for positive or negative K (but not zero). For positive K, the function is defined to be zero for $b > K$.

For this family of gain nonlinearities, the solution for input level as a function of output level is:

$$\mathbf{x}_{\text{eq}} = \frac{\mathbf{y}_{\text{eq}}}{H\left(1 - \mathbf{y}_{\text{eq}}/K\right)^K}$$

Equilibrium compression curves according to this relation, for several values of K, are shown in Figure 11.4.

History Connection: Wheeler's Automatic Volume Control

Harold A. Wheeler (1928) of the Hazeltine Corporation, a manufacturer of radios, analyzed a cascade of several variable-gain amplifiers in the automatic volume control (AVC) of an AM broadcast radio receiver. His resulting input–output level curves are as plotted in Figure 11.6.

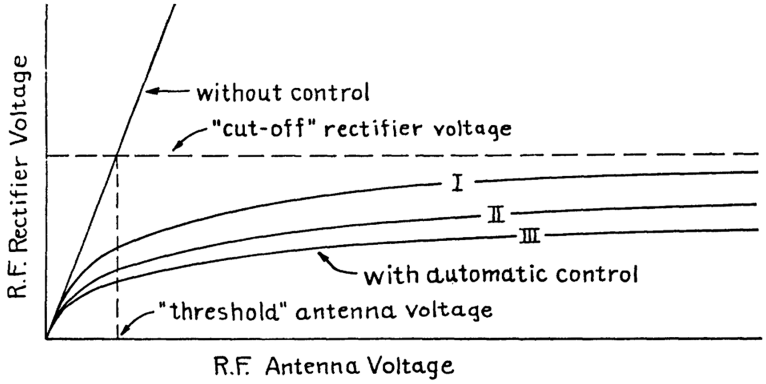


Figure 11.6: The input–output level curves of Wheeler (1928), for systems with one, two, or three cascaded variable-gain amplifiers, corresponding approximately to our model of Figure 11.5 with K values of 1 to 3. The radio-frequency output level ("R. F. Rectifier Voltage") is almost independent of input level ("R. F. Antenna Voltage") when the input level is well above the "threshold." [Figure 3 (Wheeler, 1928) reproduced with permission of the IEEE.]

Our family of nonlinear gain functions is a generalization of this idea of a cascade of several variable-gain amplifiers. The number of cascaded variable-gain multipliers is $|K|$. For example, four stages may multiply four times by $1 - b/4$, or divide four times by $1 + b/4$, as illustrated in Figure 11.5. The nonlinear gain functions are chosen to have $g(0) = 1$, and less gain otherwise (because b is nonnegative), with the initial rate of gain reduction being $dg/db = -1$, independent of the parameter K.

Wheeler's amplifier stages—vacuum pentodes—had the property that their gain could be controlled by a grid voltage, with the gain decreasing approximately linearly toward zero at a cutoff voltage. Thus with K stages his nonlinear gain function was something like $g(b) = (1 - b)^K$ in our terminology—like our $g(b)$ with positive K, and b representing the output level fed back as grid control voltage, but without the divisor of K that we use to keep the initial rate of gain reduction independent of K.

For positive K, an input level growing without bound reduces the gain $g(b)$ toward a limit of zero as the output level approaches K is our formulation, or as the output level approaches 1 in Wheeler's. Therefore his plot, reproduced in Figure 11.6, has all the curves approaching the same "cut-off" level, while our positive-K curves in Figure 11.4 approach different levels. Besides this subtle difference, our use of the divisor K also lets us generalize to negative K values, which do not lead to a finite output level bound, and which better model the compression in cochlear models.

Wheeler notes that in light of certain limitations with the amplifier tubes, "it is undesirable to reduce the amplification ratio per stage below about 1/10 of its normal value. When controlling several tubes, these limitations become unimportant." With three amplifier tubes cascaded, he therefore achieves a gain range of about a factor of 1000, or 60 dB. In cochlear models, we get similar benefits from distributing large gain changes over multiple cascaded filter stages.

11.4 Multiple Cascaded Variable-Gain Stages

When K is a positive or negative integer, we can interpret the nonlinear gain $g(b)$ as the result of cascading $|K|$ simple gain stages, as illustrated in Figure 11.5 for $K = 4$ and $K = -4$. For these K values, when $b = 1$, the gains are reduced to $0.75^4 = 0.32$ and $1.25^{-4} = 0.41$, respectively, illustrating the fact that for $b \leq 1$ the multiplicative and divisive schemes are not very different when multiple stages are cascaded (as was seen already in Figure 11.3).

For positive K, as b approaches K, $g(b)$ approaches zero, so the output level will approach K as the input level grows without bound. That is, the system compresses toward a constant output level.

For negative K, the output level grows as the $(1 - K)^{\text{th}}$ root of the input level at high levels (square root for $K = -1$, cube root for $K = -2$, etc.), rather than toward a fixed limit.

As is well known, the functions $(1-b/K)^K$ approach $\exp(-b)$ for large $|K|$, giving logarithmic compression at high levels (in the narrow gap between the $K = 40$ and $K = -40$ curves in Figure 11.4):

$$\mathbf{x}_{\text{eq}} = \frac{\mathbf{y}_{\text{eq}} \exp(\mathbf{y}_{\text{eq}})}{H}$$
$$\log(H\mathbf{x}_{\text{eq}}) = \mathbf{y}_{\text{eq}} + \log(\mathbf{y}_{\text{eq}})$$

The last term's relative contribution becomes negligible at high levels, where $\mathbf{y}_{\text{eq}} \gg \log(\mathbf{y}_{\text{eq}})$. The exact solution for $\mathbf{y}_{\text{eq}}$ as a function of $H\mathbf{x}_{\text{eq}}$ in this case is known as the Lambert W function (Corless et al., 1996); it approaches linear near zero level and logarithmic at high level.

If b is bounded, each gain in Figure 11.5 varies over a limited gain range, from 1 down to $1 - b_{\max}/K$ for $K > 0$, or down to $1/(1 + b_{\max}/|K|)$ for $K < 0$. It doesn't take very many amplifier stages to approach the high-K behavior, in which the overall gain varies from 1 down to $\exp(-b_{\max})$. As a result, nearly log-like (or more precisely Lambert-W-like) compression is the expected result at moderate levels when enough variable-gain mechanisms are cascaded, whether they are multiplicative or divisive.

This multistage approach is attractive not only for its robust compressive behavior across a wide range of input levels, as pointed out by Wheeler (1928), but also because it is a fair model of how gain control works in our models of the cochlea.

11.5 Gain Control via Damping Control in Cascaded Resonators

In auditory models made from cascaded resonators, such as the gammatone filters described in Chapter 9, varying the damping factors of the resonators will vary their gain, at least for frequencies near the resonant peak, where most of the output power typically is.

Referring back to Figure 11.1, suppose the controlled system is a fourth-order all-pole gammatone—a cascade of four resonators—and that b is used to linearly vary their damping factors, starting from a damping factor of ζ_0 in quiet:

$$\zeta = \zeta_0(1 + b/4)$$

Each resonator has a peak gain of approximately $1/\zeta$, so the peak gain of the 4-stage gammatone is:

$$g_{\text{peak}}(b) = \zeta_0^{-4}(1 + b/4)^{-4}$$

Thus, for frequencies near the peak frequency, the variable-damping gammatone filter with an output detector feeding back to control the damping this way behaves just like the multiplicative-gain loop of Figure 11.2 with $H = \zeta_0^{-4}$ and $K = -4$.

For $\zeta_0 = 0.1$, each stage starts with a peak gain of about 10. At an output level of $b = 12$, the damping

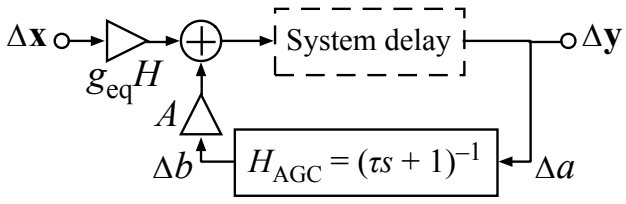


Figure 11.7: This linear system is a small-signal model of the AGC's dynamic response to changes in the input level about an equilibrium condition. The inputs and outputs of this linear model are the perturbations of input and output levels, $\Delta\mathbf{x}$ and $\Delta\mathbf{y}$. The dashed box represents the delay of the controlled system to level fluctuations; we ignore it initially, or assume that the delay is negligible compared to the time constants of the loop. The negative gain A expresses the slope with which changes in the control parameter b affect the output level at this equilibrium. For definiteness in subsequent analysis, a first-order smoothing filter is used here as the AGC loop filter.

quadruples to 0.4, increasing the relative resonator bandwidth by only a factor of 4, but changing the peak gain by a factor of $4^{-4} = 0.004$ or -48 dB (about a factor of 256 in amplitude, 65,000 in power, and approaching a fifth-root compression curve). Gains at frequencies away from the filter's peak do not change as much, so the abstraction in which we have separated the gain from the resonance is not a perfect or complete description of how the more general controlled system behaves.

If the feedback signal is limited, for example by a saturating detection nonlinearity, to $b \leq 12$, then the gain range in the example is limited to 48 dB. High enough input levels will drive the system to a high-level linear region with damping pinned at 0.4.

11.6 AGC Dynamics

Next, consider the dynamics, where the K value again plays an important role. We model the AGC dynamics by linearizing the loop around an assumed equilibrium operating point, using the model shown in Figure 11.7, where variables in the linearized loop are small level changes, and the feedback is additive (linear) instead of multiplicative.

Starting from equilibrium at some input level, with equilibrium gain $g_{eq} = g(\mathbf{y}_{eq})$, consider the effect of a small increment $\Delta\mathbf{x}$ in the input level in Figure 11.2. The increment propagates immediately to the output (that is, we assume that the forward linear system is fast compared to the AGC loop dynamics), in the same ratio as the input–output ratio (assuming a half-wave or full-wave rectifier): $\Delta\mathbf{y}/\Delta\mathbf{x} = \mathbf{y}/\mathbf{x} = g_{eq}H$. The resulting increment $\Delta\mathbf{y} = \Delta a$ is filtered, smoothly changing Δb and thereby Δg in $g(b) = g_{eq} + \Delta g$, the effect of which on the output $\Delta\mathbf{y}$ closes the loop. In the linearized analysis, the nonlinear effect of moving toward a new equilibrium is ignored, and only first-order effects of small changes in level and gain are considered. Ideally, Δb will respond on a reasonable time scale, and not overshoot or oscillate.

To compute the model's closed-loop transfer function to level changes, we need to know the relationship between Δb and the resulting $\Delta\mathbf{y}$, via Δg. In the linearized model, the gain A from Δb to $\Delta\mathbf{y}$ is taken to be the derivative at equilibrium, which depends on the input level and on the nonlinear function $g(b)$. With $g(b)$ being a decreasing function of b, A will be negative:

$$A = \left.\frac{\partial \mathbf{y}}{\partial b}\right|_{b=\mathbf{y}_{eq}} = H\,\mathbf{x}_{eq}\left.\frac{\partial g}{\partial b}\right|_{b=\mathbf{y}_{eq}}$$

The derivative of g is:

$$\frac{\partial g}{\partial b} = \frac{\partial (1 - b/K)^K}{\partial b} = -(1 - b/K)^{K-1}$$

This derivative depends on the equilibrium level about which the model is linearized; at any given equilibrium we evaluate it at $b = \mathbf{y}_{\text{eq}}$. Plugging in our previously derived expression for $\mathbf{x}_{\text{eq}}$ in terms of output level, from Section 11.3, gives the linearized gain parameter A that completes the loop at that equilibrium:

$$\begin{aligned} A &= \frac{-H\,\mathbf{y}_{\text{eq}}}{H\left(1 - \mathbf{y}_{\text{eq}}/K\right)^K}\left(1 - \mathbf{y}_{\text{eq}}/K\right)^{K-1} \\ &= \frac{-\mathbf{y}_{\text{eq}}}{1 - \mathbf{y}_{\text{eq}}/K} \end{aligned}$$

Notice that H affects the relation between the input and output levels, but not the loop dynamics for a given output level.

Recall from Section 6.14 that for a feedback system, the closed-loop transfer function is:

$$H_{\text{closed}} = \frac{H_{\text{forward}}}{1 - H_{\text{loop}}}$$

so for the linearized AGC model, with forward gain $g_{\mathbf{eq}}H$ and loop gain $A\,H_{\text{AGC}}$, we have:

$$H_{\text{closed}} = \frac{g_{\mathbf{eq}}H}{1 - A\,H_{\text{AGC}}}$$

To interpret this linearized closed-loop transfer function, it helps to have a definite transfer function for the loop filter, H_{AGC}. A typical simple case is to use a one-pole lowpass filter, as shown in Figure 11.7:

$$H_{\text{AGC}} = \frac{1}{\tau s + 1}$$

in which case the closed-loop transfer function is:

$$H_{\text{closed}} = \frac{g_{\mathbf{eq}}H}{1 - \frac{A}{\tau s+1}}$$

Let us define the factor $M = 1 - A$; since A is negative, M is greater than 1. We can manipulate the closed-loop transfer function into a form that makes it clear that the response has a zero at the loop filter's pole position (at $s = -1/\tau$), and a real pole at $s = -M/\tau$:

$$H_{\text{closed}} = \frac{\frac{g_{\mathbf{eq}}H}{M}(\tau s + 1)}{\frac{\tau}{M}s + 1}$$

That is, the closed-loop response is faster (higher pole frequency, or shorter time constant) by a factor of M than the loop filter's response.

At low signal levels, where A is near zero and M is near one, the pole and zero approximately cancel, and the system simply passes input changes to the output with gain $g_{\mathbf{eq}}H$. Considering the limit for large s, the same expression gives the high-frequency gain of the closed-loop transfer function—that is, at frequencies of level fluctuation that are too fast to have any feedback gain effect compressing them—for any level (any M). For low enough frequencies and $M > 1$, the closed-loop gain magnitude is less then $g_{\mathbf{eq}}H$, indicating that the feedback gain control is reducing the output level fluctuations.

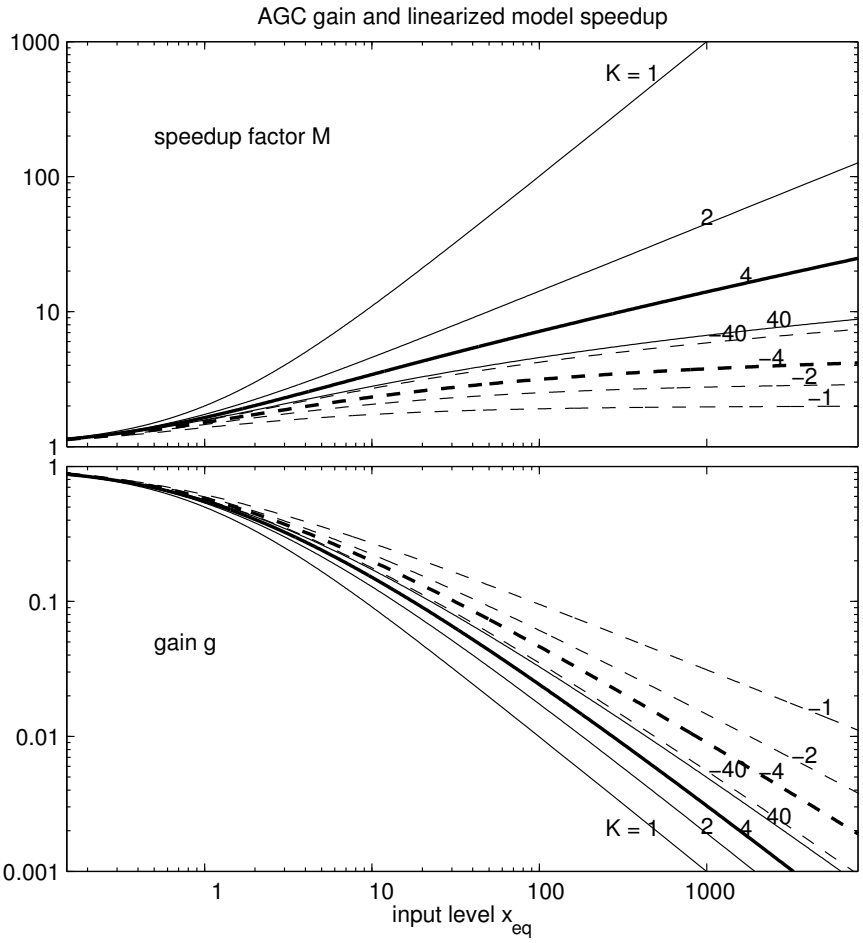


Figure 11.8: As the equilibrium input level increases, the gain g (lower panel) decreases, and the linearized closed-loop system gets faster by a speedup factor M (upper panel), for the example AGC loop with $H = 1$ (curve styles and K values as in Figure 11.3). Both sets of bold curves, $|K| = 4$, show moderate speedups. Lower positive K values lead to very low gains and corresponding very high loop speedups—a challenge to stability in the real system that includes additional delay in the loop. The cochlear models that we build are not quite as simple as this model, but their AGC behavior is approximately modeled by $K = -4$ (heavy dashed curves), with well-controlled loop time constants and compressed but not tightly controlled output level.

At any level, the low-frequency limit, or DC gain, of the closed-loop response to level change is $g_{\mathbf{eq}}H/M$, a reduction by a factor of M. This means the step response settles until the level difference is reduced by the same factor as the loop *speedup factor*. That is, an AGC loop that controls the output level tightly necessarily responds very quickly.

Nolle (1948) gives the name *flatness factor* to M, to describe the input–output compression in an automatic volume control (a.v.c.) amplifier, and shows that the same factor acts as a speedup factor in his loop analysis. As in some modern treatments (Pérez et al., 2011), Nolle had assumed a detector and variable gain that were linear on a dB level scale (logarithmic level detector and exponential gain nonlinearity), which led to linear loop dynamics independent of level; he summarized:

> The behavior of an a.v.c. amplifier, following a sudden change of input level, is analyzed on the basis of the following assumptions, which are justified in the case of many practical a.v.c. amplifiers: (1) the open-circuit voltage developed by the rectifier is a linear function of the decibel output level of the amplifier; (2) the decibel gain reduction in the controlled stages is a linear function of the gain-control voltage; (3) only one resistance–capacitance filter section is important in delaying delivery of the rectifier output voltage to the gain-control points. It is shown that the last condition is desirable from the standpoint of stability. The analysis shows that, following a sudden change of input level, the fraction $(1 - 1/e)$ of the decibel gain change required to reach a new equilibrium occurs in $(RC)/M$ seconds. RC is the time constant of the filter section specified in assumption (3), while M, a dimensionless "flatness factor," is defined as the decibel change of input level required to produce a 1-dB change of output level, under equilibrium conditions.

Our more general analysis shows that the basic result, coupling flatness to speedup through the factor $M = 1 - A$, doesn't depend on those assumptions, but is approximately correct for any loop with one-pole feedback filter, linearized about any equilibrium. Our result shows how loop dynamics depends on level in a real system, where the detector remains physically plausible down to low levels (unlike a logarithm) and where the gain nonlinearity is realistic and flexible. The level dependence is captured by the dependence of M on the equilibrium output level:

$$M = 1 - A = 1 + \frac{\mathbf{y}_{\text{eq}}}{1 - \mathbf{y}_{\text{eq}}/K}$$

which is plotted in terms of input level in Figure 11.8. As the plot shows, the dynamics can show a big difference between the two forms of nonlinear function (positive versus negative K) for small $|K|$. But if the gain is broken up over several amplifier stages (for example, $|K| = 4$ or more) then the curves are closer together, meaning the form of the nonlinearity—multiplicative versus divisive—is less important.

Figure 11.9 and Figure 11.10 show an AGC system simulation using $K = -4$, and its linearized model about the point of 30 dB gain reduction; assuming a 20 kHz sample rate, the simulation can be interpreted as being based on a 1250 Hz signal tone and a loop filter with 10 ms time constant, for a closed-loop response time constant of about $\tau/M = 3$ ms, which is at the fast end of the sensible range for AGC in hearing.

With the parameters used in the simulation, the loop filter does not do a great job of suppressing the signal's fine time structure from the control variable b, as can be seen in Figure 11.10. As a result, the gain that multiplies the signal includes components at the signal frequency, and twice the signal frequency, and more, which will generate small second-order and higher-order distortion products at the multiplier.

11.7 AGC Loop Stability

As with feedback control systems in general, stability is a concern in the design of an AGC loop; the nonlinearity makes it harder to analyze than a linear system, so the conditions for stability may not be as simple.

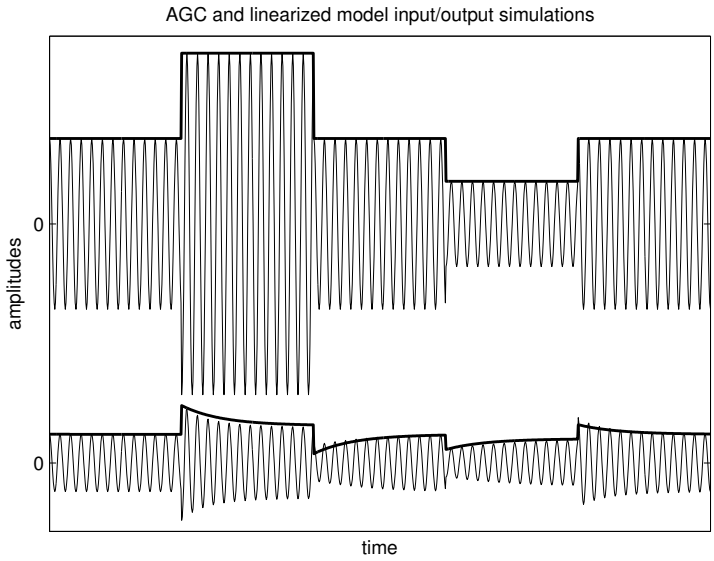


Figure 11.9: Inputs (top) and outputs (bottom) of a simulation of the multiplicative AGC system of Figure 11.2, and its small-signal linear model (see Figure 11.10 for parameters), as the input amplitude is stepped up and down by factors of 2. The amplitude-modulated sinusoids are the inputs and outputs of the nonlinear AGC simulation, while the bold curves are the inputs and outputs of the linearized model added to the equilibrium levels, scaled up by a factor of π for comparison to the peaks of the sinusoids. The fact that the output curves do not perfectly match is an indication that the linearization about the equilibrium condition is not a perfect model of the nonlinear dynamics. The equilibrium gain from input to output is $g_{\text{eq}}H = 0.34$, which is why the output curves are so much smaller than the input curves.

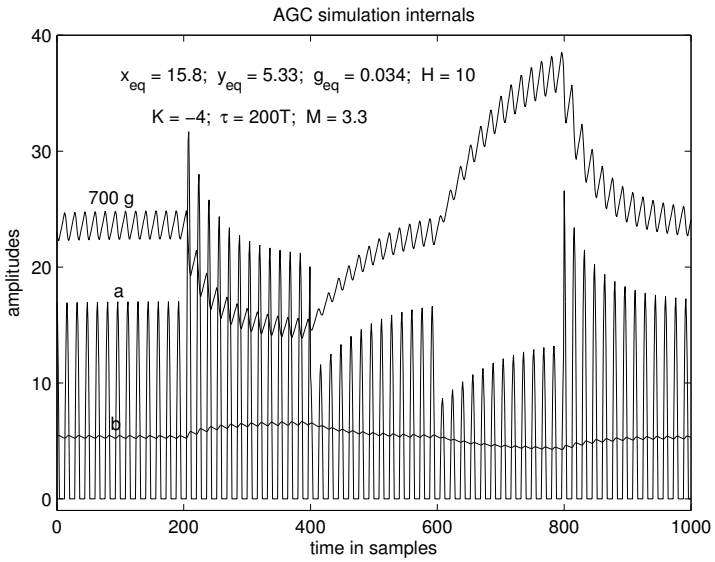


Figure 11.10: The internal signals a, b, and g of the simulation shown in Figure 11.9, of the AGC system of Figure 11.2, show how the half-wave-rectified output is smoothed, imperfectly, resulting in a gain g with considerable ripple. The parameters of the simulation are indicated on the plot; the g signal has been scaled up to reduce confusion. Notice that the speedup factor (in the linear approximation) is a moderate $M = 3.3$, even though the equilibrium gain reduction is fairly severe at $g_{\text{eq}} = 0.034$ (about −30 dB relative to the gain at low level).

Generally, if the loop doesn't regulate the system too quickly, and the loop filter doesn't have excess phase shift, stability will not be a problem. The variants with bounded speedup factor (with negative K) are preferred for that reason.

As examples, consider $K = 4$ and $K = -4$ for a system with $H = 1$ and input level $\mathbf{x}_{\text{eq}} = 1000$. The equilibrium output levels are $\mathbf{y}_{\text{eq}} = 3.06$ and $\mathbf{y}_{\text{eq}} = 9.00$, respectively, corresponding to speedup factors $M = 14.1$ and $M = 3.8$. That is, the positive-K version has its output at about 3/4 of its asymptotic limit of K, and is compressing small level changes 14:1, with a response time 14 times faster than the loop filter; the negative-K version has a 3 times higher output, is compressing only 3.8:1, and is responding only 3.8 times faster than its loop filter's natural response. The increasingly fast response of the tightly compressing positive-K version could lead to stability problems at some level, if there is any delay in the loop. The negative-K version is more robust, but controls the output level more loosely (that is, it lets the output level vary more with input level changes).

The small-signal model that we derived is unconditionally stable, but the real system may not be. When the forward system has delay (the dashed box in Figure 11.7) comparable to τ/M, the added delay will make the loop overshoot and ring. A simulation with $K = -4$ and an added delay of τ/M is shown in Figure 11.11; it remains fairly well behaved, but exhibits moderate gain overshoots at input-level steps.

In a positive-K version with delay, big enough step increases in the input will even drive b high enough to make $1 - b/K$ negative, driving the gain to zero and causing the system to have zero output until it recovers; this is well outside the region where the small-signal analysis applies. Lower-bounding the gain to be not less than some small positive value (or upper-bounding b to a value less than K) is one way to help ameliorate that problem. The lower bound on gain will lead to a large-signal linear regime, where the input level is so high that the output level is high enough to hold the gain fixed at its lower bound.

A simple condition for stability in a linear feedback loop is that the magnitude of the loop gain drop below

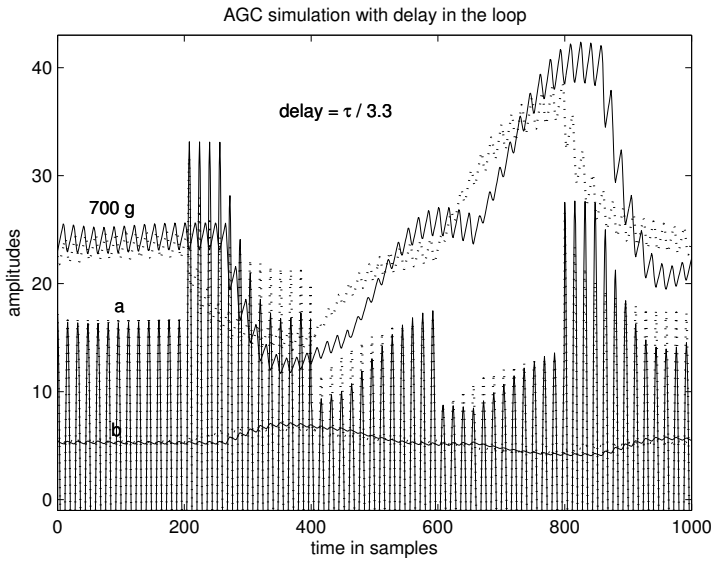


Figure 11.11: The simulation shown in Figure 11.10 has been repeated with a delay equal to $\tau/3.3$ in the forward path. Solid curves show the new simulation, while the dotted curves are copies of the response from the previous figure, from the simulated system without delay. This delay corresponds to τ/M using $M = 3.3$ from the linearization about the initial equilibrium. With the delay, the smoothed output estimate b is slow to react, and the gain adjustment overshoots. The output level excursions are slightly greater. With this much delay, even though it is a fairly small fraction of the loop filter's time constant, a larger speedup factor M would lead to considerably worse behavior.

1 before the phase shift around the loop reaches 180 degrees. If we ignore the system delay in Figure 11.7, the phase shift around the loop is bounded by 90 degrees, so the loop is stable. The magnitude of the loop gain decreases to 1 near $\omega = |A|/\tau \approx M/\tau$. The additional system delay D that will lead to instability is approximately the delay that adds another 90 degrees at this frequency, or $D = \pi/(2\omega) = \pi\tau/2M$. Thus, the higher the speedup factor, the less delay can be tolerated in the loop, or the longer the loop filter time constant must be to stabilize the loop.

We might alternatively model the level-response dynamics of the controlled system by a rational transfer function, such as a one-pole smoothing filter, rather than as a pure delay, so that the loop can again be analyzed in terms of poles. A one-pole smoother of time constant τ_s to represent the system will have a low-frequency group delay of τ_s, and a phase shift bounded by 90 degrees, so the resulting poles will never quite go unstable. But the loop can still overshoot and ring excessively if the delay condition derived above is approached. Therefore, the system delay should be kept small compared to the sped up loop time constant.

Techniques known as *compensation* can be used to help stability; basically, they reduce the phase shift of the loop filter to reduce ringing or to tolerate more delay in the loop. In our CARFAC cochlear model, as described in Chapter 19, we use a loop filter with less phase shift to keep the level-response dynamics more stable even while responding quite rapidly.

For more on stability and compensation, a control-theory text should be consulted. Dorf (1974) includes an explicit treatment of linearized automatic gain control loops.

11.8 Multiple-Loop AGC

Nested gain-control loops can be useful in adapting to a wide dynamic range, helping to achieve a high compression while staying far from ringing and instability. An outer slow loop can reduce the dynamic range that an inner loop has to handle, except at abrupt onsets. Increasing the loop gain in the outer loop causes it to compress to a level that looks relatively low to the inner loop, keeping the inner loop's equilibrium at a lower M than it would be otherwise, so it will be further from the point where delay in the loop will bother it.

Figure 11.12 shows two approaches for combining an outer slower loop with faster inner loops, in which we arbitrarily use a factor of 2 higher loop gain on each enclosing loop. The doubling of gain on each enclosing loop is typical of what we use in our cochlear-model AGC. The second approach can be interpreted as exactly the multiplicative AGC analyzed in this chapter, except with a different loop filter. This loop filter has multiple real poles and zeros, as analyzed in Chapter 19, resulting in a lower phase shift through a wide frequency range, and therefore better dynamics.

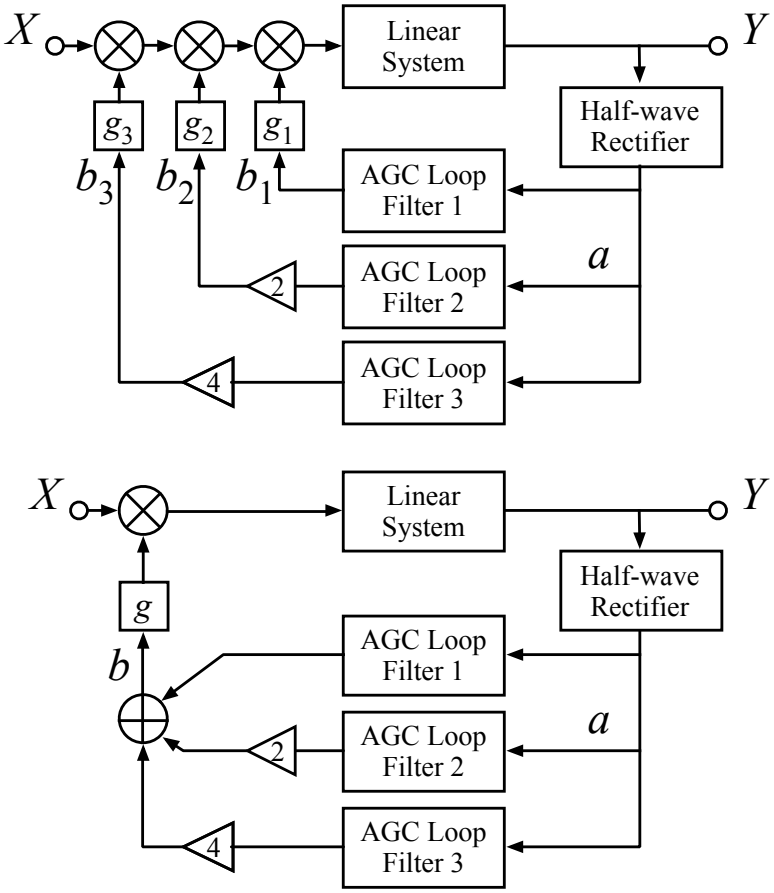


Figure 11.12: An AGC with multiple filters in the feedback path, each controlling a separate gain through separate nonlinear gain functions (top), and an alternative way to configure a multi-time-constant loop with the outputs of the loop filters summed to control a single gain (bottom). If the outermost loop is very slow, it can reduce slow level variations, leaving the more inner loops to deal with faster and smaller level variations. If the nonlinear gain functions g_i are approximately exponential, even with moderate $|K|$, the effects of the two schemes will be similar.

Chapter 12

Waves in Distributed Systems

> The movement of waves down the basilar membrane is analogous to the propagation of light waves in a medium of continuously changing index of refraction. While the velocity of light varies as it travels through the substance, substantial reflections will not occur as long as the index of refraction changes slowly enough.
>
> — "The cochlear compromise," Zweig, Lipes, and Pierce (1976)

The cochlea is not a system of lumped elements, but rather a *distributed* system. Distributed systems can be linear or nonlinear; before we tackle the cochlea, we need to study the simpler linear case. Linear system theory is still applicable to distributed systems, as it is to systems of lumped elements—but it gets a bit more complicated because the transfer functions are not as simple as ratios of polynomials, and because signals are functions of location, not just of time.

The spatially distributed state of a distributed system is typically described as a wave, a function of continuous time and space, rather than in terms of inputs and outputs. We can conceptually look at the system response (transfer function or impulse response or frequency response) at an infinite number of outputs at a continuum of locations, or at a finite set of outputs at discrete locations, with the wave at some unique location as input.

Systems of lumped elements, having a finite number of degrees of freedom, are described by ordinary differential equations that relate the rate of change of each state variable to the state and inputs of the system. In a distributed system, the motion (displacement, velocity, pressure, current, voltage, or whatever) of every point along a continuum of one or more dimensions is part of the state, so such systems are described by partial differential equations, in which rates of change with respect to both time and space are involved. In this chapter, we do not involve ourselves much with partial differential equations directly, but rather discuss what their solutions tell us about waves.

We explore the mathematical description of traveling waves, starting with uniform media, in which the distributed parameters are the same everywhere, and show that the responses of the medium at a succession of points, as a wave propagates in one direction through the medium, are equivalent to the responses at the outputs of the stages of a cascade of identical filters. Then, we consider nonuniform media, in which the parameters change with location, as in the cochlea, and show that the same structure, a cascade of filters representing short segments of the medium—but not identical stages in this case—still makes a good model. Further, we discuss approximating the cascade stages in terms of poles and zeros, or low-order linear systems of lumped elements. As a result, we have a very general way to approximate the responses in a nonuniform distributed system such as the cochlea by the responses of a cascade of simple filters that are easy to realize efficiently in computers.

Example: Delay Lines and Moving-Average Filters

A *delay line* is a linear system whose output is a copy of its input from an earlier time, with delay T:

$$y(t) = x(t - T)$$

The transfer function is not representable as a rational function, and has no poles or zeros:

$$H(s) = \frac{Y(s)}{X(s}) = \exp(-sT)$$

and the frequency response is just a phase lag proportional to frequency:

$$H(i\omega) = \exp(-i\omega T)$$

In the 1940s, J. Presper Eckert built delay lines for use in a radar moving target indicator, and then as memory systems for early digital computers (Galison, 1997). The delay lines in the UNIVAC used acoustic compression waves in cylindrical tubes of mercury, each storing 720 distinguishable pulses (10 words of 12 6-bit characters each) (Bell and Newell, 1971); that is, the system had a least a 720-dimensional state space. It would have taken thousands of parts to approximate such a system with lumped electrical elements. A hundred such delay lines constituted the 1000-word recirculating memory of the UNIVAC.

Consider a continuous-time moving-average filter: the output at any time is the average value of the input over an interval of length T ending at that time. This system cannot be described in terms of ordinary differential equations, nor implemented in terms of lumped circuit elements, because its state must represent all the details of the input over the preceding interval of duration T, as via a delay-line memory. If the input functions of time are band-limited—have a bound on the highest frequencies present in them—then the moving average can be well approximated by lumped circuits, or by discrete-time systems using samples of the signal to be smoothed. Or such systems can be built as physical analogs, using wave-propagating devices or magnetic-tape delay loops or some other distributed mechanism to hold the continuous-time distributed state. Whether by a distributed delay line or a lumped approximation to one, a moving-average filter can be implemented as shown in Figure 12.1.

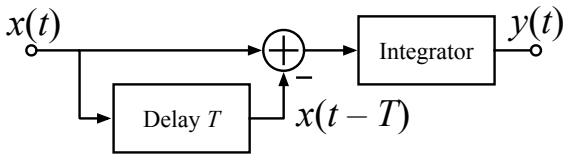


Figure 12.1: This diagram represents a linear time-invariant system that computes a moving average: $y(t) = \int_{t-T}^{t} x(t)dt/T$. The delay operator holds the details of the function $x(t)$ over an interval of duration T, so cannot be described via a finite number of state variables. Due to the subtraction $x(t) - x(t - T)$, the integrator outputs the difference between the integral up to time t and the integral up to time $t - T$ (except for a potential constant offset that can be dealt with by resetting the integrator state to zero when the input has been at zero long enough). In a physical implementation, the delay might be implemented by propagating waves through a lossless uniform medium, or by an approximation such as a suitable length of coaxial cable or a string under tension.

12.1 Waves in Uniform Linear Media

The propagation of sound as a pressure wave in air (or in a mercury delay line) is an example of wave behavior in a distributed linear system. The propagation of waves on the strings of a musical instrument is another.

Whether we care about waves in just one direction, or in multiple directions, or standing wave patterns, to the extent that the system is linear we can treat one direction at a time. When considering only one direction of propagation, each point can be considered as an input, and any later (or "downstream") point as an output, related by a linear system between them. This observation, extended to nonuniform (spatially varying) media, underlies our cochlear modeling approach.

In a distributed system, just as in a system of lumped elements, the signals that hold their shape between an input and an output, or between any two locations, are sinusoids, or more generally complex exponentials. The general real sinusoid (sinusoidal variation in time) propagates as a sinusoid in space, moving at a constant speed, if the system is *uniform* and *lossless*. That motion is therefore described by a sinusoidal function of a linear combination of time and space coordinates:

$$W(x,t) = A_1 \cos(-kx + \omega t - \phi)$$

where W is a distributed (wave) response, k is the *wavenumber* (which depends on properties of the medium and on the frequency ω), and x is the space coordinate in the direction in which the wave is propagating. This wave description has no notion of input, output, or causality, but we can incorporate those ideas by picking a point in space to call the input, and another to call the output.

For the wave $W(x,t)$ as described, the relationship between wavenumber k and frequency ω is the same everywhere (that is, it does not depend on x), which is what we mean by *uniform*, the restricted class of wave propagation media that we are considering at present. Air of constant temperature is a uniform medium for sound propagation; so is a stretched guitar string of uniform mass per unit length. A canal of constant width and depth is a uniform medium for water waves running in the direction of its length, while a canal of varying width or depth is not. In a nonuniform system, a sinusoidal function of time no longer leads to a corresponding sinusoid in space, so the description of waves will need to be generalized.

The wavenumber is the *spatial frequency* of the wave, or rate of change of phase with distance, in radians per meter in SI units. It is the spatial analog of temporal frequency, ω, the rate of change of phase with time, in radians per second. These frequencies appear in the cosine's phase argument with opposite signs (for a wave propagating in the $+x$ direction) since the wave at some positive x corresponds to what the wave was doing at $x = 0$ at an earlier t.

From the phase of the wave description, we can see the trajectories of constant phase in space and time. For example, consider the point of zero phase: $0 = -kx + \omega t - \phi$, which can be solved for position as a function of time: $x = (\omega t - \phi)/k$. Similarly, any point of constant phase moves in the $+x$ direction at a velocity (dx/dt) known as *phase velocity*:

$$v_\phi = \frac{\omega}{k}$$

Knowing that the cosine has a period of 2π, a wavelength, or spatial period, is apparent:

$$\lambda = \frac{2\pi}{k} = \frac{2\pi\, v_\phi}{\omega}$$

where wavenumber, wavelength, and phase velocity can be written as $k(\omega)$, $\lambda(\omega)$, and $v_\phi(\omega)$ when we want to make explicit their dependence on frequency.

As we did with systems of lumped elements, we use complex exponentials as a way of combining phase and amplitude effects into complex amplitudes. Using the definition of the cosine in terms of complex exponentials we rewrite W as the sum of two complex waves of opposite phase, and then pull the phase effects out

EE Connection: Linear Electrical Transmission Lines

It is common to model hydrodynamic wave systems such as the cochlea by electrical circuit analogs. Figure 12.2 shows a *ladder filter* of series inductors and shunt capacitors, approximating an electrical transmission line's inductance and capacitance per unit length. A lossless transmission line is essentially a pure delay (due to its wave equation, developed below), and its discretization into this circuit of inductors and capacitors acts approximately as a delay up to a bandwidth near the section resonant frequency $1/\sqrt{LC}$.

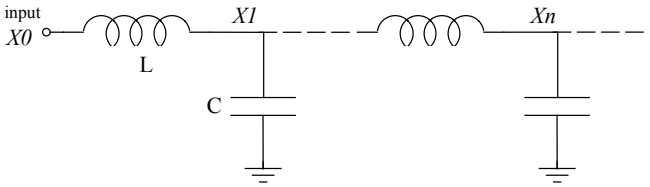


Figure 12.2: Waves travel at speeds somewhat less than the speed of light along wire transmission lines, including telephone lines, coaxial cables, power lines, etc. Such lines can be approximated or modeled by LC delay lines, circuits of iterated lumped elements as shown here.

For low enough frequencies, the responses at points Xn in the circuit of lumped elements are very much like the responses at a set of points on the distributed line, if those points are separated by comparable total amounts of series inductance and shunt capacitance. At higher frequencies, near $\omega = 1/\sqrt{LC}$, where $k(\omega)$ is large and the wavelength $2\pi/k(\omega)$ is not long compared to the section spacing, the approximation breaks down. Due to the local resonance, such high frequencies will not propagate through the lumped circuit, which is why a lowpass filter can be made this way. Such filters were traditionally called *electric wave filters* (Campbell, 1922; Zobel, 1924).

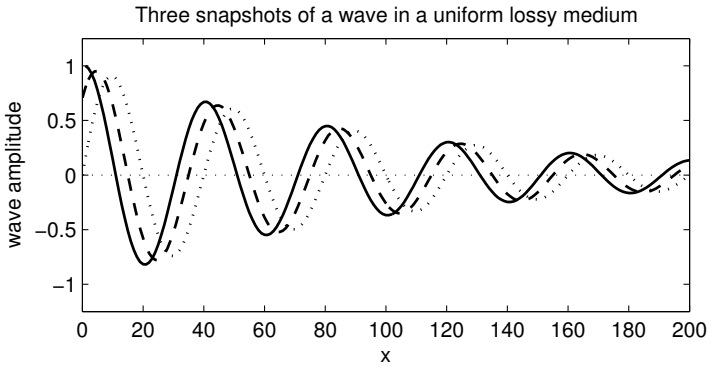


Figure 12.3: Three snapshots of a wave in a uniform lossy medium: the real part of $\exp(-ikx+i\omega t)$. The wave is sinusoidal in time, and a decaying sinusoid in space (due to the complex wavenumber k of the medium). The three snapshots (solid, dashed, and dotted curves) are separated by time intervals Δt corresponding to 1/8 cycle of the sinusoid, or $\pi/4$ radians (45 degrees) of phase (the period in time is $T = 8\Delta t = 2\pi/\omega$). The wave has a wavelength ($2\pi/k_{\mathrm{Re}}$) of 40 distance units (real part of k is therefore $k_{\mathrm{Re}} = 2\pi/40$ radians per distance unit). The wave amplitude is decaying as the wave propagates, by a factor of e per 100 distance units (imaginary part of k is therefore $k_{\mathrm{Im}} = -1/100$).

into the leading complex constants:

$$W(x,t) = \frac{A_1}{2}\left[\exp(-ikx + i\omega t - i\phi) + \exp(ikx - i\omega t + i\phi)\right]$$

$$W(x,t) = A\exp(-ikx + i\omega t) + A^*\exp(ikx - i\omega t)$$

where $A = (A_1/2)\exp(-i\phi)$. For simplicity, we usually work with complex waves, understanding that we can always get back to real signals by adding the complex-conjugate wave:

$$W(x,t) = A\exp(-ikx + i\omega t)$$

If the linear medium is lossy, the wave gives up some of its energy as it propagates, and the amplitude decreases exponentially in x, as shown in Figure 12.3. The exponential decay with x is described by a complex k, corresponding to an additional factor in the wave formula:

$$W(x,t) = A\exp(k_{\mathrm{Im}}x)\exp(-ik_{\mathrm{Re}}x + i\omega t)$$

where the parameters k_{Re} and k_{Im} are the real and imaginary parts that define a complex wavenumber $k = k_{\mathrm{Re}} + ik_{\mathrm{Im}}$. Using this complex wavenumber, we can again write the wave simply as a constant times an exponential, even though it is decaying (or growing) as a function of location x:

$$W(x,t) = A\exp(k_{\mathrm{Im}}x - ik_{\mathrm{Re}}x + i\omega t)$$

$$= A\exp(-ikx + i\omega t)$$

Sign conventions vary; here a lossy wave moving in the $+x$ direction is represented by a k with a negative imaginary part. Similarly, if k_{Im} is positive, the wave grows with x. We call such growth of the traveling wave *active amplification*. For physical systems in which the wave decays ($k_{\mathrm{Im}} < 0$), the wave energy will typically be dissipated as heat. Alternatively, in a physical system where waves grow ($k_{\mathrm{Im}} > 0$), the active

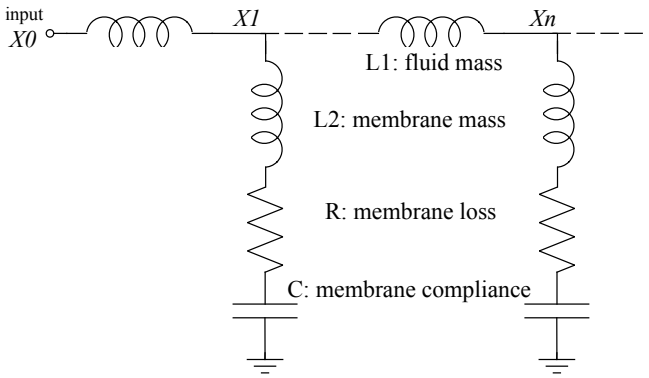


Figure 12.6: A transmission line of the type that has often been used for modeling wave propagation in the cochlea, starting with Wegel and Lane (1924), assuming a basilar membrane with significant mass and frictional loss. The mass and loss, modeled by inductance and resistance in the shunt-admittance legs, have little effect at low frequencies, where the membrane compliance, modeled by capacitance, limits the shunt current and makes the system act like the delay line of Figure 12.2; at higher frequencies this model exhibits local resonance and loss.

amplification requires an energy source; a limited power supply imposes a limit on the range over which an active amplifying system can remain linear.

The waves we have considered so far are steady sinusoids in time, but everything still works if instead of $i\omega$ we use the general complex frequency s, the Laplace transform variable. These steady, decaying, and growing complex exponentials make up the entire set of eigenfunctions of continuous-time linear systems—not just of the lumped-element systems we studies in Chapter 6, but also of the distributed wave-propagating systems of this chapter.

Physics Connection: Plane Waves in Multiple Dimensions

In uniform systems of more than one dimension, we can represent *plane waves* in any direction by a simple generalization: replace the dimension x by the *space vector* (location in 2D or 3D space) $\mathbf{x}$, the wavenumber k by the *wave vector* $\mathbf{k}$, and their product by the *dot product* (sum of products of coordinate dimensions) $\mathbf{k} \cdot \mathbf{x}$:

$$W(\mathbf{x}, t) = A \exp(-i\,\mathbf{k} \cdot \mathbf{x} + i\omega t)$$

The wave vector points in the direction of wave propagation. Planes perpendicular to this direction are called wavefront planes. Moving from a position $\mathbf{x}_1$ to a position $\mathbf{x}_2$ changes $\mathbf{x}$ by the vector difference $\mathbf{x}_2 - \mathbf{x}_1$. If this difference is orthogonal to the wave vector $\mathbf{k}$, then the phase at a given time does not change, because the dot product $\mathbf{k} \cdot \mathbf{x}$ does not change. Such positions are within a wavefront plane of the plane wave (it is also possible to have waves that are not plane, such as spherical waves emerging from a point source, but that's beyond what we'll need to consider here).

This generalization of k to a vector, plus the generalization to complex k, is the reason that waves are more often written in terms of wavenumber than in terms of wavelength.

EE Connection: More General Transmission Lines

In the cochlea, the series inductance L in the model is analogous to fluid *mass*, while the shunt capacitance C is analogous to membrane *compliance*. Compliance is springiness, displacement per force or strain per stress—the reciprocal of stiffness. For more general media, such as lossy transmission lines and active cochleae, more general series impedances and shunt admittances are used, as shown in Figure 12.4, to represent combinations of mass, springiness, and loss or gain.

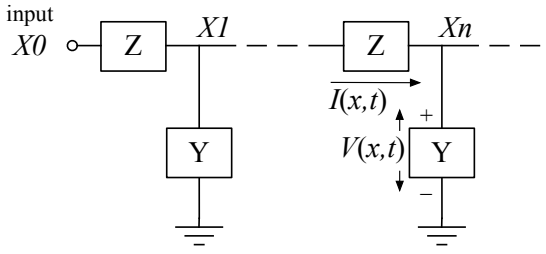


Figure 12.4: The general transmission-line model uses series impedances Z proportional to the distributed line's series impedance per unit length, and shunt admittances Y proportional to the line's shunt admittance per unit length. The electrical signals used to analyze it are the currents through the series elements and voltages across the shunt elements, signed as shown. The elements Z and Y are more general than the inductors and capacitors of Figure 12.2; each may contain series or parallel connections of several lumped elements, for example.

When the impedance of the series and shunt elements of a transmission line—or more precisely the series impedance $Z(\omega)$ per unit length (ohms per meter) and shunt admittance $Y(\omega)$ per unit length (siemens per meter)—are known and fixed (not varying with location), the wavenumber and characteristic impedance (ratio of wave voltage and current amplitudes) are simple functions of Z and Y.

The wavenumber solutions are found by solving Heaviside's *telegrapher's equations*, or *coupled time-harmonic transmission line equations*, that follow from elementary circuit analysis (Steinmetz, 1910; Mohamed, 2006):

$$\frac{dV}{dx} = -ZI, \quad \frac{dI}{dx} = -YV$$

These coupled equations imply a pair of similar wave equations for voltage and current waves, as can be seen by substituting each into a derivative of the other:

$$\frac{d^2V}{dx^2} = ZYV, \quad \frac{d^2I}{dx^2} = ZYI$$

Since the spatial second derivative of $\exp(-ikx + i\omega t)$ provides a factor of $-k^2$, the wave equations are satisfied by sinusoidal voltage and current waves of frequency ω whenever the wavenumber satisfies:

$$k(\omega)^2 = -Z(\omega)Y(\omega)$$

The ratio between the resulting wave amplitudes is the line's characteristic impedance Z_0, given by:

$$\frac{V}{I} = Z_0(\omega) = \sqrt{Z(\omega)/Y(\omega)}$$

In the case of a pure LC line, with $Z = i\omega L$ and $Y = i\omega C$ per unit length, k satisfies the relation $k^2 = \omega^2 LC$, so k is proportional to ω, signifying a frequency-independent velocity of $v = \omega/k = 1/\sqrt{LC}$, or a pure delay of $\sqrt{LC}$ per unit length.

EE Connection: Single- and Double-Ended Lines

The transmission lines that we have seen so far are known as *single-ended* lines, as a single wire carries the wave signal as a voltage relative to ground. Transmission lines are often built or drawn as *balanced lines*, as in Figure 12.5, with equal amounts of series impedance in two paths—like the 300-ohm twin-lead TV antenna wire some of you may be familiar with. The two chambers of the cochlea make a balanced or *differential* structure, too. In the analysis of such transmission lines, however, it is common to transform to a single-ended line, with only one line of series impedances, and to take voltages relative to a *ground* rather than differentially between the two sides. The resulting single-ended line is equivalent, for the differential wave; that is, for any wave that is antisymmetric about the center.

The same transformation is used in analyzing cochlear hydrodynamics, and in converting that analysis to an electrical equivalent, by assuming the pressures and longitudinal velocities (like currents) in cochlear fluid–membrane waves are opposite on the two sides of the cochlear partition (this is a good approximation as long as the wavelength is long compared to the dimension of asymmetries of the cochlear partition).

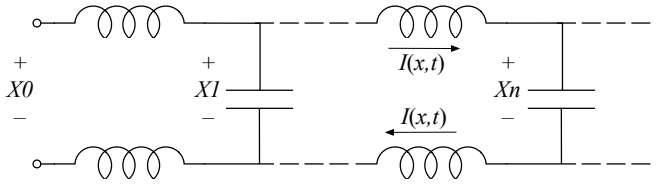


Figure 12.5: In a *balanced* or *differential* delay line, the signal of interest is the difference between the voltages on two sides, rather than with respect to a global *ground* potential. The symmetry allows such systems to be reduced to equivalent *single-ended* lines. The approximate symmetry of the hydrodynamic system of the cochlea across the cochlear partition is usually treated as good enough that the cochlea can be modeled by a single-ended transmission line circuit.

Physics Connection: Dispersion Relations and Bidirectional Waves

The relationship between wavenumber k and frequency ω is known as the *dispersion relation* for the medium. For a given frequency, it will typically have two (sometimes more) solutions for k, representing waves traveling in the $+x$ and $-x$ directions. We sometimes ignore the latter, the wavenumber for the backward-traveling wave, and represent k as simply a function of ω. Some systems will also have multiple *modes* of wave propagation, each with its own wavenumber for a given frequency, representing multiple solutions of a single dispersion relation (Watts, 2000); again, we sometimes ignore such complications, but keep in mind that they may become important second-order effects when evaluating some models of the cochlea.

Suppose a medium propagates waves in the $+x$ direction. For mechanical or hydrodynamic media, such as media that sound waves propagate through, these waves may be characterized by displacements and velocities (for example, of points on a guitar string, or points in air). Displacement and velocity patterns propagate as sinusoidal waves $W(x,t)$ of the form described above. For electrical lines, waves are usually described in terms of voltages and currents, which propagate similarly:

$$V_+(x,t) = V_1 \exp(-ik(\omega)x + i\omega t)$$

$$I_+(x,t) = \frac{V_+(x,t)}{Z_0(\omega)}$$

where ω is the frequency being considered, $k(\omega)$ is the propagation constant or wavenumber, V_1 is the complex amplitude of the voltage wave propagating in the $+x$ direction (assuming k is positive or has a positive real part), and $Z_0(\omega)$ is the characteristic impedance of the medium.

A wave can also propagate in the other direction, corresponding to the other solution of the dispersion relation; consider the electrical line's dispersion relation:

$$k^2 = -Z(\omega)Y(\omega)$$

$$k = \pm\sqrt{-Z(\omega)Y(\omega)}$$

When we treat $k(\omega)$ as a (single-valued) function of frequency, the reverse wave typically has wavenumber $-k(\omega)$, as implied by the square-root form of the dispersion relation (in other media than this simple transmission line model, multiple solutions of the dispersion relation may not be so simply related).

Besides negating the wavenumber in the phase expression, the reverse wave negates the relationship between voltage and current (since we want to continue to measure current consistently as flowing in the $+x$ direction):

$$V_-(x,t) = V_2 \exp(+ik(\omega)x + i\omega t)$$

$$I_-(x,t) = \frac{-V_-(x,t)}{Z_0(\omega)}$$

When waves exist in both directions at once, these currents and voltages add linearly to make consistent solutions.

Physics Connection: Reflections and Standing Waves

When waves of one frequency are propagating in both directions (for example, due to reflections from the far end of a finite line), the resultant waves in the medium are just sums of the forward and backward wave voltages and currents, which we can write in a way that emphasizes that the temporal pattern is everywhere sinusoidal:

$$V(x,t) = \exp(i\omega t)\left[V_1 \exp(-ik(\omega)x) + V_2 \exp(ik(\omega)x)\right]$$

$$I(x,t) = \exp(i\omega t)\frac{\left[V_1 \exp(-ik(\omega)x) - V_2 \exp(ik(\omega)x)\right]}{Z_0}$$

Due to the sum in one and difference in the other, the voltage-to-current ratio is no longer simply the characteristic impedance Z_0, as it is for the forward wave alone, nor $-Z_0$ as it is for the backward wave alone.

When the amplitudes V_1 and V_2 of the forward and backward waves are equal, the result is a pure standing wave, with sinusoidal time variation at every point, but with nonmoving sinusoidal spatial envelopes, showing zero current at the voltage envelope maxima and zero voltage at the current envelope maxima:

$$V(x,t) = |V_1| \exp(i\omega t)\cos(k(\omega)x - \phi)$$

$$I(x,t) = \frac{|V_1|}{Z_0}\exp(i\omega t)\sin(k(\omega)x - \phi)$$

for some ϕ that depends on the phases of V_1 and V_2.

Standing waves can come from reflections in lossless media. More generally, partial reflections lead to unequal forward and backward wave amplitudes.

A transmission line is not usually infinite; if it is terminated by a load resistance or impedance, or a short or open circuit, that termination will constrain the voltage-to-current relationship at that point, making a boundary-value problem that the V and I of the sum of forward and backward waves must satisfy. The solution will give the compatible amplitudes and phases of forward and reflected waves. Only in the case of termination by the characteristic impedance will there be no reflected wave: the forward wave will transfer all of its energy into the termination impedance, just as if it was propagating its energy down an infinite transmission line. Conversely, if the termination impedance is lossless (a short or open circuit, or a purely reactive impedance), all the energy will be reflected, and the backward amplitude will be equal to the forward amplitude, with just a phase shift that depends on the impedance, making a standing wave. In between these cases, some energy will be transferred into the termination and some will be reflected.

A similar analysis applies at boundaries between media, electrical or otherwise. For example, where sound waves in the ear canal hit the ear drum, some energy is transferred in and some is reflected; when the middle ear pushes on the cochlear fluid, some energy is transferred in and some is reflected. A high efficiency of energy transfer corresponds to a matching of characteristic impedances of the wave propagation media. In an electrical system, a *transformer* is used to change the voltage–current ratio to interface efficiently between different impedances; in the ear, the leverage of the middle-ear bones does this job.

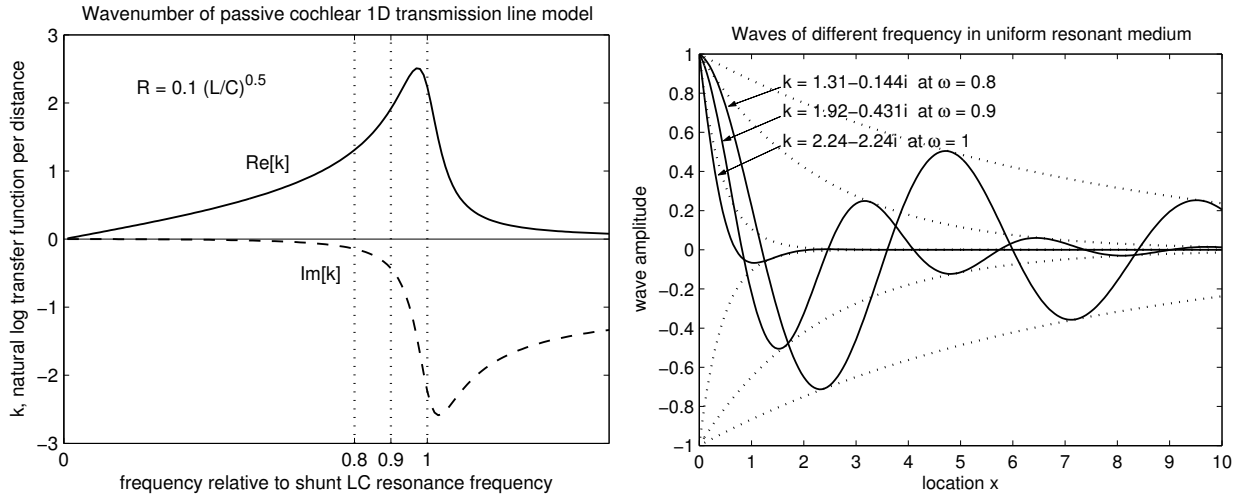


Figure 12.7: The complex wavenumber of the transmission line of Figure 12.6 is plotted in the left panel (real part solid, imaginary part dashed). As discussed in this chapter, these plots can also be interpreted as the log transfer function of a unit-length segment of the line (phase lag per distance solid, log gain dashed). Snapshots of waves at the marked frequencies (0.8, 0.9, and 1.0 times the shunt resonance frequency) are plotted in the right panel, and annotated with corresponding numerical values of k, to illustrate the relative phase shift of $\mathrm{Re}(k)$ radians per unit distance, and attenuation by a factor $\exp(\mathrm{Im}(k))$ in a unit distance, at these frequencies; spatial envelopes are shown dotted. Lower frequencies, where the imaginary part of k is negligible, propagate easily, while frequencies near and above resonance are strongly attenuated.

12.2 Transfer Functions from Wavenumbers

As described in the boxes, electrical transmission lines are a type of well-understood one-dimensional distributed medium that is commonly used as an analogy for wave propagation in other systems. If we've made the right analogies, the solution to an electrical transmission line will tell us the solutions to a corresponding hydrodynamic system, in terms of forces (analogous to voltage) and fluid flows (analogous to current); from these we can get membrane displacements, kinetic and potential energies, power flows, etc.

Considering waves propagating only in the $+x$ direction, the transfer function between any two positions x_1 and x_2 separated by a distance Δx is just the ratio of waves at those positions:

$$H(\omega) = \frac{A\exp(-ik(\omega)\,x_2 + i\omega t)}{A\exp(-ik(\omega)\,x_1 + i\omega t)} = \exp(-ik(\omega)\,(x_2 - x_1)) = \exp(-ik(\omega)\,\Delta x)$$

In general, this transfer function or frequency response is not a rational function. A key observation is that the distributed parameter values that describe the medium appear in the transfer function in the argument of an exponential function, rather than in the coefficients of polynomials as parameters do in lumped-element systems. The reader who is interested in how the wavenumber relates to the underlying physics or circuit models will find an introduction to that topic in the boxes. Others just need to know that a complex frequency-dependent $k(\omega)$ can be found when the physics of the medium is known in terms of partial differential equations or equivalent descriptions, and that such descriptions usually have only a few parameters.

If the parameters are such that the spatial frequency $k(\omega)$ is real and proportional to temporal frequency, then velocity $v = \omega/k$ is constant, and the transmission line is a pure delay. The transfer function of a length of such a line is just a phase lag proportional to frequency, of $k\,\Delta x = \omega\,\Delta x/v$ radians. But the transfer function of a pure delay, $\exp(-i\omega\,\Delta x/v)$, is not a rational function, so it does not have poles and zeros, and we can't

build the equivalent filter from a finite number of lumped elements. Similarly, the transfer functions of most other distributed systems cannot be exactly expressed as rational functions.

If k has a negative imaginary part, then the transfer function's magnitude, $\exp(k_{\text{Im}}\,\Delta x)$, is less than 1, and the transmission line is lossy. Conversely, if the imaginary part of k is positive, the system amplifies.

As an example, consider the wavenumber solution shown in Figure 12.7 for the transmission line of Figure 12.6. At low enough frequencies, it is nearly lossless and acts like a delay. At frequencies near the resonant frequency of the shunt admittance, the wave energy is significantly absorbed by the resistance, rather than propagating down the line. The plot of wavenumber is also exactly a plot of the transfer function between two points, if the y axis is scaled as gain in dB (for the imaginary party of k) and phase (for the real part of k) for the given length of transmission line.

If we know the dispersion relation (wavenumber k versus frequency ω, which comes from the differential equations describing the physics) for a uniform distributed system, then we can model wave propagation to multiple points in that system, in one direction, as a cascade of filters representing the segments from each point to the next, with segment transfer functions that are exponentially related to the wavenumber. Extending this approach to more general nonuniform media is our next problem.

12.3 Nonuniform Media

The cochlea is a distributed system that supports wave propagation and is subject to modeling by an electrical analogy—but it is not uniform. Properties such as scala (cochlear chamber) dimensions, and membrane width, mass, and stiffness, change as a function of the cochlear place dimension x. Waves that are sinusoidal in time are not sinusoidal in space, and the simple analysis just described does not quite apply. Nonetheless, there is a well-founded analysis method that leads to wave propagation being describable with only a little extra complication. It amounts to treating local segments as if the system were spatially uniform. This approach motivates the filter cascades that we introduce for modeling the cochlea.

If the wave medium is not uniform—such as a canal of varying water depth, or a cochlear partition of varying stiffness—we can mostly characterize it, for waves in one direction, by a complex wavenumber $k(\omega, x)$ as a function of frequency and location. The basic method as described by George Green (1837) was summarized in the 1911 Britannica article "Waves," in discussing long waves in canals (Britannica, 1911):

> ... The theory was further extended by G. Green (1837) and by Lord Rayleigh to the case where the dimensions of the cross-section are variable. If the variation be sufficiently gradual there is no sensible reflection, a progressive wave travelling always with the velocity appropriate to the local mean depth. There is, however, a variation of amplitude; the constancy of the energy, combined with the equation of continuity, require that the elevation η in any particular part of the wave should vary as $b^{-\frac{1}{2}}h^{-\frac{1}{2}}$, where b is the breadth of the water surface and h is the mean depth.

Here, "the velocity appropriate to the local mean depth" means the velocity $v = \omega/k$ corresponding to the local wavenumber (in the case of long waves in canals of depth h, the wavenumber is proportional to $h^{-1/2}$). In more general nonuniform media, possibly certain regions amplify a particular frequency while others attenuate it (that is, the imaginary part of k can change sign). Under reasonable conditions, however, each point in such a medium (i.e., along the x dimension) can be characterized by a local dispersion relation, or wavenumber as a function of frequency, as though it were part of a uniform medium.

Consider the transmission line of Figure 12.4, specialized to a lossless delay line with $Z = i\omega L$ and $Y = i\omega C$. The wavenumber is $k(\omega) = \sqrt{-Z(\omega)Y(\omega)} = \omega\sqrt{LC}$. If L is fixed, and C increases with position x, then k increases; that is, the wave slows down, and wavelength gets shorter, locally consistent with a local wavenumber calculation as if the medium were uniform. Fixed C and increasing L can lead to this same

spatially dependent wavenumber solution. But the waves would be rather different in these cases, as can be explained by the concept of characteristic impedance, the ratio between the voltage wave and current wave. The characteristic impedance is $Z_0 = \sqrt{Z/Y} = \sqrt{L/C}$, which can increase, decrease, or stay the same with distance. Since the LC line is lossless, the total power being conveyed by the wave at any location should remain constant, even as the wavenumber changes. Since the wave power in terms of the voltage is V^2/Z_0, and in terms of the current is $I^2 Z_0$, one of I or V may increase with position while the other decreases, depending on how Z_0 varies. Thus different positions in the medium may have both different wavenumbers and different amplitudes of the wave solution, depending on how the parameters of the medium vary. If the parameters change slowly enough with position, these local properties can be connected into an approximate complete solution for the wave. The wavenumber solutions determine the phase and velocity characteristics of the waves, while the characteristic impedance leads to an amplitude correction factor to assure conservation of energy. For systems that have more dimensions than the transmission line's one distance dimension, the parameters that affect the wave heights can more be complicated that just an impedance.

Examples of a varying wavenumber and a corresponding wave are shown in Figure 12.8, based on a spatially varying version of the transmission line shown in Figure 12.6, analyzed as a function of place for a fixed frequency (as opposed to Figure 12.7, an analysis of a uniform medium as a function of frequency).

To characterize what happens between points far apart, the approximate methods for solving the differential equations effectively break the medium into infinitesimal segments of length dx, and multiply together all the transfer-function factors for those segments, $\exp(-ik(\omega, x)\,dx)$ (and possibly also an amplitude correction factor). The factors are exponentials, and the product of exponentials is the exponential of a sum, so the resulting product is the exponential of an integral along the x dimension:

$$H(\omega) = \exp\left(-i \int_{x_a}^{x_b} k(\omega, x)\,dx\right)$$

This frequency-dependent gain and phase factor $H(\omega)$ is the approximate transfer function between points x_a and x_b in a nonuniform medium—a generalization of the exact transfer function $H(\omega) = \exp(-ik(\omega)\,\Delta x)$ that characterizes a stretch of a uniform medium.

This formula was proposed by Schroeder (1973) as the approximate solution to "an integrable model for the basilar membrane"; but his solution ignored the effect of the changing characteristic impedance, leading to a violation of conservation of energy as waves propagated from one part of the medium to another. Zweig et al. (1976) showed how to bring Schroeder's solution closer to physical reality with a spatially varying amplitude correction factor, which in the case of a wave of pressure difference across the partition and constant series impedance per unit length (fluid mass) is a factor proportional to $k^{-1/2}$. For the transfer function, the ratio of this factor between points x_a and x_b is the desired correction factor. In a cochlear transmission line, the capacitance, representing membrane compliance, is what is usually assumed to increase with x, resulting in increasing k and decreasing pressure amplitude (transfer function magnitude less than 1). The increasing compliance also results in increasing displacement per pressure difference, so transfer functions for displacement waves can be greater than 1. Both pressure and displacement waves follow Schroeder's approximation above, but need different corrections to conserve energy. For displacement amplitude, $k^{+1/2}$ might be more appropriate (or k^{+1} even, depending on system parameters). Schroeder effectively used k^0, so Zweig's correction almost seems opposite of what was needed, even if it was correct for pressure difference.

In modern physics, this method, including the correction, is known as the Liouville–Green (LG) or Wentzel–Kramers–Brillouin (WKB) method. The spatial integral of a varying local k is characteristic of the LG/WKB method, which is also known as the phase-integral method (because $k\,dx$ is a phase when k is real). This method for approximate wave solutions of nonuniform partial differential equations appears to have been first proposed by Carlini in 1817, and has been reinvented several times (Fröman and Fröman, 2002).

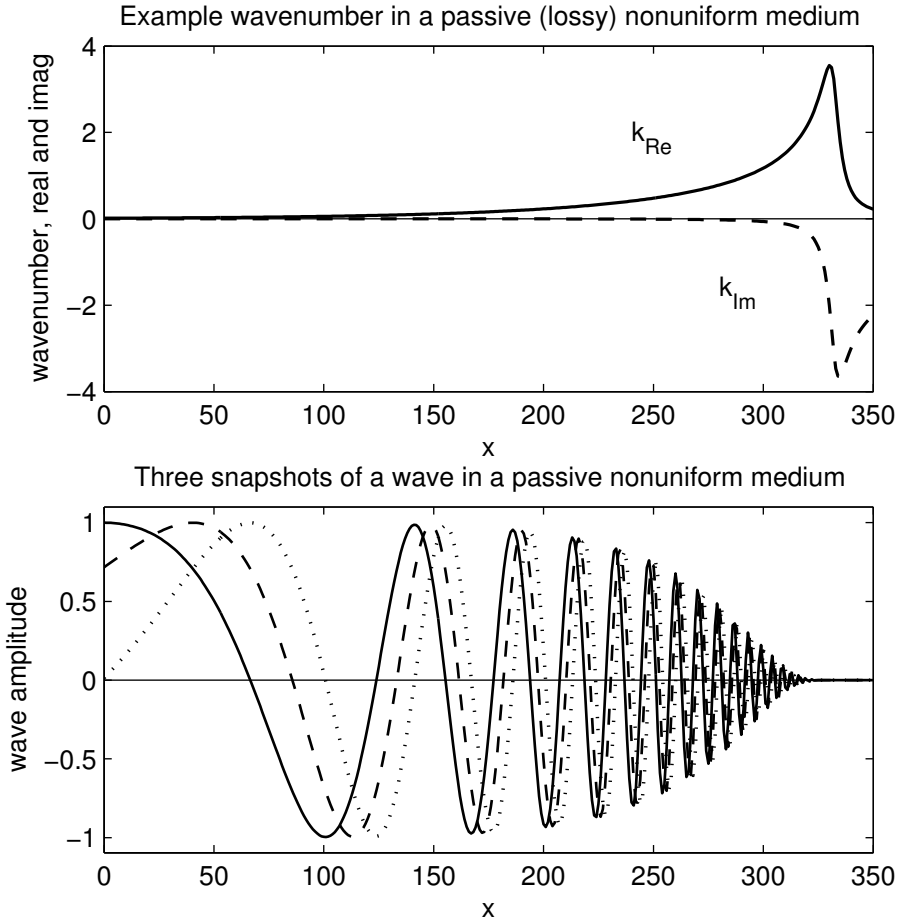


Figure 12.8: The wavenumber real and imaginary parts (top curves, solid and dashed respectively) of a hypothetical nonuniform medium, for a given frequency corresponding to the resonant frequency of the medium near its right-hand end, and corresponding wave snapshots (lower curves). The wave is sinusoidal in time, but not in space, due to the spatially varying properties of the medium that reduce the resonant frequency by a factor of 2 every 50 distance units. The wavenumber curves resemble those of Figure 12.7, since the same form of underlying model is used here, but with spatially varying parameters. The three snapshots (solid, dashed, and dotted curves) are separated by time intervals equal to 1/8 cycle of the sinusoid, or $\pi/4$ radians (45 degrees) of phase. By comparing the positions of peaks or zero crossings, it can be seen the wave is moving rapidly in the region of low k (left side), and slows down as k increases (right side). As in Figure 12.7, the attenuation becomes high when the wavelength gets down to about 6 distance units (the wavenumber gets up to about 1). The wave energy is fully dissipated shortly before it reaches the location with shunt resonance frequency equal to the wave frequency. The wave power is proportional to the square of the amplitude shown; no amplitude correction has been applied to account for either the slowing of the wave or the varying physical parameters of the system.

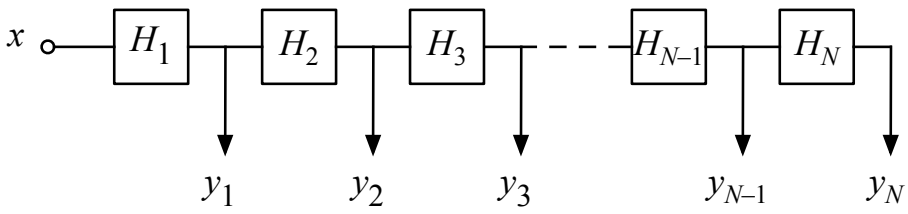


Figure 12.9: A cascade filterbank models wave propagation in a distributed system if each stage transfer function is a good approximation to the effect of the system's local dispersion relation, via the relation $H_j(\omega) \approx \exp(-ik(\omega, x_j)\,\Delta x)$

These approximate methods are known to be fairly accurate until a wave goes somewhat past the position of greatest response, into the region where it is strongly attenuated. In this region, an extension of the method to the *mode-coupling Liouville–Green approximation*, to couple energy into multiple modes corresponding to multiple solutions for k from the dispersion relation, can be used if more accurate results are needed (Watts, 2000).

12.4 Nonuniform Media as Filter Cascades

The WKB method provides a mathematical approach to finding an amplitude correction factor versus place, but it may be easier to derive it from physical conservation of energy arguments (Zweig et al., 1976; Lighthill, 1981). We sometimes ignore this gradual correction, and just assume that our model systems use a variable that has been adjusted such that power and amplitude maintain a consistent relationship across places. That is, the wave amplitudes that our filters propagate may be regarded roughly as square root of power, as opposed to a physical variable such as pressure, membrane displacement, or velocity.

The frequency response $H(\omega)$ discussed above corresponds to a linear system transfer function $H(s)$, though it will not be a rational function. Its place dependence comes from the dependence of the complex wavenumber on frequency and place, which can be derived from a physical model or fitted to observed response data. Furthermore, we can factor this filter into a product, or cascade, of several filters by splitting the interval of integration (from x_a to x_b) into N smaller steps:

$$H(\omega) = \prod_{j=1}^{N} \exp\left(-i \int_{x_{j-1}}^{x_j} k(\omega, x) dx\right)$$

Any number and size of steps leads to a factorization. If the steps are small enough, then each individual filter will be well approximated from a local wavenumber by $\exp(-ik\Delta x)$, where $\Delta x = (x_b - x_a)/N$ is the step size, making it easier to tie the filters to a local model of the underlying wave mechanics:

$$H(\omega) \approx \prod_{j=1}^{N} \exp\left(-ik(\omega, x_j)\Delta x\right) \approx \prod_{j=1}^{N} H_j(\omega)$$

Therefore, independent of the details and dimensionality of the underlying wave mechanics, the responses of the medium (for example, the cochlear partition) at a sequence of places are equivalent to the responses at the outputs of a sequence of cascaded filters $H_j(\omega)$, of the structure illustrated in Figure 12.9. The LG/WKB method constrains the design of those filters when the underlying physics is known (de Boer and Viergever,

1982).

Finally, approximating the filters H_j by rational functions of s can lead to a cascade filterbank that is easy to implement and at least qualitatively a good approximation to the original distributed system. If we know only the amplitude, or only the phase, from measurements or data of some sort, we can still use that to constrain the design of a filter model, in conjunction with causality and minimum-phase assumptions. Even for nonlinear and time-varying wave mechanics, we can reasonably assume that a nonlinear and time-varying filter cascade will be a useful structural analog and a fruitful modeling approach: modeling local behavior with local filters.

12.5 Impulse Responses

In studies of the cochlear traveling wave, impulse responses are often inferred from either mechanical or neural measurements, at one or more points in the cochlea. A feature of such impulse responses that has been commented on by many authors is a "chirp" or "glide": a changing instantaneous frequency of the resonant impulse response, during the brief duration of the response. This "frequency modulation" has posed a puzzle for modelers (Møller, 2003):

> ... examination reveals another difference between these impulse response functions and impulse response functions of ordinary bandpass filters. The frequency of the damped oscillations changes along the time axis. This becomes evident by observing the intervals between zero crossings of the individual waves of the damped oscillation, which are not the same in the beginning as at the end of the oscillation. ... The reason for this frequency modulation lies in the traveling wave nature of the basilar membrane motion. The functional implication of frequency modulation is, among other things, that the motion of the basilar membrane *cannot* be correctly modeled by ordinary filters made up of lumped components or by simulation of such filters. This means that not even the filter characteristics of a single point of the basilar membrane are adequately described by such models, and only models containing or simulating distributed constants are adequate.

Fortunately, it is not true that distributed systems can't be well modeled by systems of lumped components. Møller's last sentence points the way: "simulating distributed constants" is easily done within the space of rational transfer functions, using cascades of many approximate segment transfer functions as described in the previous section. Even the simple all-pole gammatone filter has been shown to produce reasonable glides (Lyon, 1998), and can be viewed as a simulation of a distributed system. Glides in the responses of pole–zero filter cascades are illustrated in Chapter 16.

12.6 Group Velocity and Group Delay

Besides the phase velocity ω/k_{Re}, waves have another important velocity, called *group velocity*, the speed at which energy propagates through the medium. Group velocity is usually analyzed, or explained, by consideration of the propagation of amplitude modulation. Given a slowly varying envelope amplitude $A(t)$, the wave under consideration at $x = 0$ is

$$W(t) = A(t) \exp(i\omega t)$$

When this wave propagates in the $+x$ direction, the envelope might propagate at a different rate from the

carrier; we emphasize that here by writing the x dependence in terms of velocities rather than wavenumber:

$$W(x,t) = A\left(t - \frac{x}{v_g}\right)\exp\left(i\omega\left(t - \frac{x}{v_\phi}\right)\right)$$

Derivations of group velocity are straightforward (Elmore and Heald, 1969), but we won't bother here. The result is that the group velocity depends on the dispersion relation via the derivative:

$$v_g = \frac{d\omega}{dk_{\mathrm{Re}}}$$

A nondispersive medium is one in which the wavenumber is proportional to frequency, so v_ϕ and v_g are equal, and independent of frequency. Nondispersive means that all frequencies and modulations travel at the same speed, and any wave holds its shape. Light waves propagate without dispersion in a vacuum, but not in glass. Sound propagates nearly without dispersion in air. But many media are highly dispersive. Short waves in deep water, and short waves in the cochlea, have k proportional to ω^2, in which case the group velocity is equal to half the phase velocity (Rubin and Atkinson, 2001)—the modulations move only half as fast as the wave peaks and troughs. In the cochlea, the 3D geometry leads to group velocity being even less than half of phase velocity in a region between long-wave and short-wave behaviors (Steele and Taber, 1979; van der Heijden, 2014).

The filter transfer function $H(\omega)$ that relates two places separated by a distance Δx in the medium has corresponding *phase delay* D_ϕ and *group delay* D_g, which can be expressed either in terms of the velocities or in terms of the phase of the transfer function, $\phi(\omega) = -k_{\mathrm{Re}}(\omega)\,\Delta x$, thereby providing a useful connection between filter delays and wave delays:

$$D_\phi = \frac{\Delta x}{v_\phi} = \frac{-\phi(\omega)}{\omega}$$

$$D_g = \frac{\Delta x}{v_g} = \frac{-d\phi(\omega)}{d\omega}$$

Part III

The Auditory Periphery

Part III Dedication: Georg von Békésy

This part is dedicated to the memory of Georg von Békésy (1899–1972), winner of the 1961 Nobel prize in medical science and physiology for his work on hearing. I never met Békésy, who died when I was still an undergraduate. I so admire his persistent work against tough odds in hearing research, and giving us the basic observations and model of how the cochlea works via traveling waves, that I felt this part of the book should be dedicated to him. My favorite quote is from his 1974 paper, published after his death (thanks to his punctual submission), in which he recalled his realization that "... dehydrated cats and the application of Fourier analysis to hearing problems became more and more a handicap for research in hearing" (Békésy, 1974). I think we have gotten beyond the cat problem, and it is my hope that, through this book, we make progress toward the day when Fourier analysis no longer handicaps hearing research.

In this part of the book, we survey the auditory periphery, and develop a machine model of its sound-processing function. We review auditory filter models, and the fitting of auditory filter models to human and animal data, based on some of the filter types developed in Part II.

We extend these filter models to develop the CARFAC description of the cochlea based on filter cascade models of wave propagation, and present a machine implementation in detail, through the final output step of the cochlea: the release of neurotransmitter by the inner hair cells to stimulate the auditory nerve.

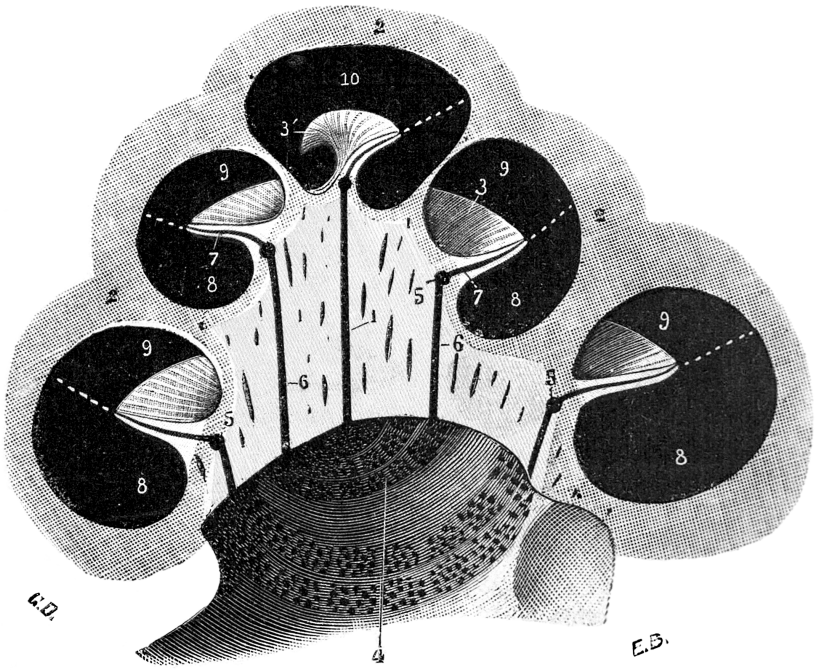


Fig. 890.

Coupe transversale du limaçon osseux : l'un des segments, vu par sa surface de coupe (*demi-schématique*).

A semischematic transverse cut of the "bony snail," the cochlea, published in color by Leo Testut (1897) (see color plates).

Chapter 13

Auditory Filter Models

> The original aim of this research was to obtain a mathematical expression for the amplitude characteristic of the hypothetical auditory filter that could be used to predict the power that a tone must have to be just audible in the presence of a given noise.
>
> — "Auditory filter shape," Patterson (1974)

> The auditory filter may be considered as a weighting function representing frequency selectivity at a particular centre frequency. Its shape can be derived using the power-spectrum model of masking which assumes: (1) in detecting a signal in a masker the observer uses the single auditory filter giving the highest signal-to-masker ratio; (2) threshold corresponds to a fixed signal-to-masker ratio at the output of that filter.
>
> — "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Moore and Glasberg (1987)

Over the last half century, many *auditory filter models* have been developed, analyzed, and applied to hearing-related problems. Linear filter models, along with more realistic quasi-linear level-dependent models, have been explored. We review several lines of development, and several criteria that filter models might try to satisfy, and show how the pole–zero filter cascade (PZFC) models achieve these desired properties.

Early attempts at describing auditory function by filters used three kinds of filter shapes: simple resonances, Gaussian filters, and rectangular filters. Most more modern auditory filter models can be seen as belonging to three main families (detailed in Section 13.4.1): the rounded exponential (roex) family, the gammatone family, and the filter cascade family. In many cases, independent efforts led to somewhat similar results, without necessarily sharing a name or any other relationship. I have discovered some of these relationships in retrospect, such as the early 1960s work by Jim Flanagan (1960, 1962) on gammatone, one-zero gammatone, and related pole–zero filter models of basilar membrane motion, long before the term gammatone was coined.

Transmission-line models of wave propagation on the basilar membrane go even further back, but the basis for approximating these systems as cascade-structured filter models was not made clear until after Zweig, Lipes, and Pierce (1976) showed how to apply the Wentzel–Kramers–Brillouin (WKB) approximation in their 1976 "cochlear compromise" paper. They ended up with a circuit model similar to the old transmission-line models of Wegel and Lane (1924), Peterson and Bogert (1950), and Ranke (1950), but the analysis that they explained led via the WKB method to a wider class of filter-cascade models of the cochlea, *cascade filterbanks*, as opposed to conventional parallel filterbanks (Lyon, 1982, 1998). Our current approach is based on such cascades that relate to the wave mechanics, but draws also on the gammatone line of development.

Nonlinear extensions of the basic filter models are more complicated. In consideration of phenomenological nonlinear models of peripheral auditory processing, Lopez-Poveda (2005) observed that "the user must generally compromise between the complexity of a model and its ability to account for a wide range of physiological phenomena," and "An optimum solution might be to combine the best of both approaches [level-dependent linear system and memoryless nonlinearity] in a way that is not yet clear." We can relieve these difficult choices through a simple model that incorporates both types of nonlinearity (memoryless distortion and level-dependent parametric variation) in a single mechanism that relates clearly to the underlying mechanics, starting with a linear pole–zero filter cascade, and adding feedback control of the pole damping, and a compressive cubic nonlinearity in the cascaded filter stages. This approach is incorporated in the CAR-FAC model (cascade of asymmetric resonators with fast-acting compression) developed in later chapters of this part of the book.

On Quasi-Linear Filters

When is linear not linear? What kind of nonlinear system lets us use linear systems descriptions effectively? How can a level-dependent filter with a strongly compressive input–output relationship be described as a linear filter, with a frequency response? These questions are addressed by the notion of a *quasi-linear filter*.

A quasi-linear filter is really a family of filters, with typically one parameter that chooses between them. In the case of auditory filters models considered here, the parameter is a signal level (an input level or an output level or some such parameter). Each filter in the family is an ordinary time-invariant linear system, described by a frequency response and other conventional descriptions, such as impulse response, poles and zeros if rational, transfer function, etc.

When the input to the auditory system is a broadband noise-like signal of a particular level, the behavior (whether observed physiologically or psychophysically) is often described fairly well in terms of linear filtering. But for different input levels, different filters are needed. Across a wide dynamic range of levels, significant changes of gain, bandwidth, and such are needed to fit the data, signifying a strongly nonlinear process. Yet at any particular level, a linear filter model fits well.

A description of the auditory system in terms of a quasi-linear filter consists of a combination of a family of linear filters and the relationship that controls the parameters of that family in response to a level measurement. Dynamic variation of level and parameters, as would occur in an AGC loop when the level is changing, is not part of the quasi-linear model.

13.1 What Is an Auditory Filter?

The auditory filters that we consider here include both those motivated by psychoacoustic experiments, such as detection of tones in noise maskers, and those motivated by reproducing the observed mechanical response of the basilar membrane or neural response of the auditory nerve. One thesis of this work is that a single model can do a good job for both of these, and thereby provide a good basis for a machine hearing system. Since there are several stages of neural processing between the cochlea and our psychoacoustic perceptions, it would not be surprising if the best parameters were different between these types of models, but it seems likely that the linear and nonlinear filtering due to the cochlea plays a sufficient role in perception that we may find one set of parameters is adequate, at least for a range of machine hearing applications.

Green (1958), in his Ph.D. dissertation, provided an early summary of the state of auditory filter models and the concept of the critical band. Measurements at that time were not adequate to determine much more than bandwidth, and his comparisons of different filter shapes (rectangular, simple resonance, and Gaussian) did not yet lead to an understanding of how to determine better-fitting shapes. He wrote that "Theorists

who advanced rival views of how the frequency analysis was accomplished marshalled physiological and anatomical data to support their positions, while psychophysical data were not used as crucial evidence," and suggested that progress could be made by more psychophysical studies to complement the other evidence. Since that time, psychophysical experiments, especially on detection of sinusoids in notched-noise maskers, have driven progress in fitting auditory filter shapes for human hearing, giving rise to the roex and gammatone families of filters, including level dependence of bandwidth and asymmetry (Rosen and Baker, 1994).

Besides the level dependence of filter shape, other nonlinearities generate distortion products, or combination tones, as was discussed with respect to the cochlear resonance theory by Barton (1908) over a hundred years ago. He wrote, "there does not exist in the air any clearly sensible pendular vibration corresponding to the combinational tone, and we must conclude that such tones, which are often powerfully audible, are really produced in the ear itself." The incorporation of such instantaneous nonlinearities, along with the level-dependent quasi-linear type of nonlinearity, has been an ongoing theme in the development of auditory filter models.

When we talk about filters, we imply a curve that represents relative response as a function of frequency, as if the auditory system were a linear time-invariant bandpass filter. But the limitations of that filter concept, and the related concept of the critical band, were well appreciated already by 1970, as expressed by Jeffress (1970),

> The idea of the critical band as a filter (resembling an electrical filter) is only an analogy that is imperfect in many respects. The bandwidth of an electrical filter is the same over a wide range of measured levels; the ear's is not. Also, the bandwidth of a filter is either its width at the half-power (3-dB down) points or its equivalent rectangular bandwidth; and both of these measures lose some of their meaning (especially their predictive value) when the filter response is unsymmetrical as the ear's is. An ear's filter skirt is considerably steeper on the high-frequency side than on the low—that is, low frequencies mask higher frequencies more effectively than the reverse.

Rhode (1971) showed that most of this nonlinearity and asymmetry is apparent very early in the auditory system, in the mechanical response of the basilar membrane, the presumed main substrate of the auditory filter.

Their limitations notwithstanding, the notions of auditory filters, equivalent rectangular bandwidths, and even symmetric filter models have served many useful roles in the intervening decades. It has become commonplace to include nonlinearities and asymmetry in auditory filter conceptions, but not always in sound processing systems inspired by them. It is less common to tie auditory filter models to computationally appropriate structures—appropriate both in relation to the underlying cochlear mechanical system and in relation to efficient digital application in the form of a "bank" of filters for sound analysis. The filter cascade is the only structure we know that efficiently grows from a single-channel model to a multichannel bank of filters. The filter cascade structure is motivated by the underlying mechanics; as Sarpeshkar (2000) points out, nature chose "a traveling-wave architecture that is well modeled by a filter cascade instead of a bank of bandpass filters."

Duifhuis (2004) recounts the history of cochlear models, and divides them into two classes: (1) the transmission-line class and (2) the filterbank class. More specifically, he says, "The major difference is that models in class 1 take physical coupling between system elements into account, whereas in class 2 the channels are independent, and coupling is completely determined by the common input." Filter cascades provide a natural model of coupling in the forward direction, and an automatic gain control (AGC) feedback network can model some coupling between channels in both directions, so these cascades can be viewed as a bridge between Duifhuis's two classes: they don't support backward traveling waves as transmission lines do, but do model the forward wave to efficiently implement filterbanks. The filter cascade is our strategy for abstracting

the transmission-line models into efficiently runnable filter models.

By auditory filter we mean to include the whole range from simple linear symmetric critical band concepts through models of nonlinear wave propagation in the cochlea, but especially those models that can be efficiently implemented and applied to problems in human and machine hearing. For the purpose of this chapter we won't treat transmission-line models as filter models unless they have been abstracted to filter-cascade representations.

The psychophysical tasks on which auditory filter models are fitted are usually the detection of tones in noise, using a power or energy estimate at the output of the filter. This energy-detection approach was an early success in predicting psychophysical results that were not predicted by an *ideal observer* theory (Pfafflin and Mathews, 1962) (an ideal observer is a hypothetical signal analyzer and decision process limited by noise but not by any imperfection of its own, which can be invoked as a bound on, and sometimes an approximation to, human observer performance). Much of this chapter is based on fitting auditory filter models to masked tone detection thresholds in human observers—that is, the models are primary optimized on psychophysical data, even when the form of the model is more physically motivated. To the extent that a physically-motivated model can do well at predicting psychophysical results, we feel that the model is good.

13.2 From Resonance to Gaussian Filters

Since experimental data on things like threshold for detection of tones in noise do not reveal the shapes of auditory filters directly, scientists generally rely on easily parameterized filter shapes as auditory filter models, and try to find parameters that lead to those models predicting or explaining the data. As experimental data have gotten better, they support not only choosing parameters of such a model, but also comparing different parameterized shape models across a range of experiments.

The simple resonance has been frequently tried, and usually rejected, as a model of auditory filtering. The "single-tuned filter" or "universal resonance curve" was applied to the critical band concept and masking data by a number groups (Schafer et al., 1950; Webster et al., 1952; Tanner et al., 1956; Mathews, 1963; Patterson, 1976). It was found to provide a better fit to psychophysical data than a rectangular filter in some cases, but still not very good. In particular, Patterson (1976) found the skirts of this filter shape to be not steep enough, and/or the peak too sharp. The sharp peak and broad skirts are the same features that to Mathews and Pfafflin (1965) made the resonance "better" than a rectangular filter, because for the same equivalent rectangular bandwidth it has less variance in its power output; but this is not a direction supported by psychophysical data.

Patterson (1974) had good results with his "symmetric filter" shape, effectively a cascade of two universal resonances, or an order-2 complex gammatone. His psychophysical data were good enough at that time to suggest this improvement, an intermediate shape between a resonance and a Gaussian, but not yet good enough to provide evidence of asymmetry.

With multiple resonances in cascade, the skirts get steeper faster than the peak gets sharper, so a structure generalizing from Patterson (1974) addresses the basic limitation of the simple resonance shape; 3 to 5 identical resonances in cascade provide a satisfactory auditory filter model (Patterson et al., 1992). A cascade of many simple resonances makes a filter that approaches a Gaussian transfer function shape—that is, the Gaussian is the limit of gammatone filters of high order. Tanner, Swets, and Green (1956) had introduced the Gaussian as a potential auditory filter, and Patterson (1976) observed that the skirts of the Gaussian fall much too fast to be realistic. This problem was the opposite of what he and others had found for the resonance, which can be considered as the low-order limit of the gammatone family.

Patterson had thereby defined the extremes between which a good auditory filter shape was to be sought; he and his colleagues later proposed first the roex family (Patterson and Nimmo-Smith, 1980; Patterson et al., 1982), and then the gammatone family (Schofield, 1985; Patterson et al., 1988), as improved auditory filters

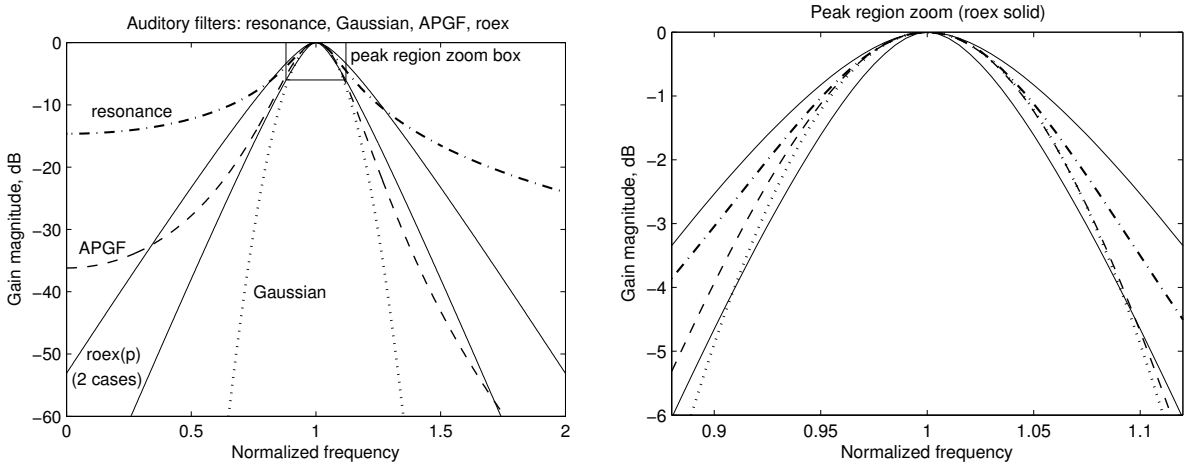


Figure 13.1: A few auditory filter model shapes, showing the large room for possibilities between the simple two-pole resonator and the Gaussian filter. The asymmetric 4th-order all-pole gammatone (APGF), and the symmetric roex(p) are included as examples of intermediate shapes. The illustrated filter shapes are matched for equal curvature at the peak, except for the sharper roex(p) case. The peak region zoom on the right shows how the matched roex peak curvature leads to rather wide skirts, and more reasonable skirts lead to a rather sharply curved peak, compared to the other filter shapes.

between these extremes. All of the filter models we address in this chapter fall into this middle ground, between a resonance (an $N = 1$ gammatone) and a Gaussian, as shown in Figure 13.1. This is also the range used by physicists to characterize shapes of diffraction peaks and other phenomena in distributed physical systems; they even sometimes interpolate between the limiting shapes using the Pearson type VII distribution, which has the shape of the general gammatone frequency response (Pecharsky and Zavalij, 2009).

13.3 Ten Good Properties for Auditory Filter Models

Auditory filter models will ideally fulfill a variety of roles, requiring a variety of good properties (Lyon et al., 2010a). In some cases we refer to how certain filter models, or families of filter models, meet or fail to meet these criteria. The particular models are described in more detail in the sections that follow.

1. Simplicity of description—there are various domains of description, and no filter model is likely to be simple in all domains. Gammatones and gammachirps, for example, are described by a simple time-domain impulse response, but have a complicated frequency-domain description. All-pole gammatones are very simple in terms of s-domain poles, are somewhat simple in the frequency domain, but rather complicated in the time domain. The roex filters have a simple formula for the (power gain) frequency response, but an equivalent time-domain response or implementation in terms of digital filters is hard to find. Filter cascades have very simple stages, and simple structural description, but overall complicated responses.

2. Bandwidth control—the first-order feature of an auditory filter is its bandwidth, which needs to be modeled as a function of at least the characteristic frequency (CF) of the cochlear place that the filter represents. In most models, the bandwidth is also controlled as a function of signal level, since it is well established in physiological and psychophysical data that bandwidths are lower at low signal levels, and increase at high signal levels.

3. Realistic and controllable relationship between peak shape and skirts—after bandwidth, the first-order shape feature of an auditory filter is how rapidly the response falls off near the band edges. All auditory filter

models provide a way to parameterize this important aspect of filter shape, and some may couple it to the level-dependent bandwidth parameter so that the filter shape changes realistically with signal level.

4. Filter shape asymmetry—data show that the filter skirt on the high-frequency side of CF is usually steeper than the one on the low-frequency side of CF (Patterson and Nimmo-Smith, 1980). Some models, such as the gammatone and the simpler roex forms, are inherently symmetric, or nearly so, and as a result do not model this asymmetry. Others, such as double-sided roex and double-sided gammatone shapes, simply provide two different models for the high side and the low side. But a two-sided modification of a gammatone completely destroys its time-domain simplicity and realizability as a rational transfer function. As an alternative, the gammachirp was developed to add a controllable asymmetry to the gammatone while retaining its other good properties; it can be skewed in either direction with one parameter.

5. Peak gain variation—psychophysical experiments don't provide much data on how the gains of the filters might vary with the signal level and bandwidth, but physiological data on cochlear mechanics do, if the auditory filters are associated with the mechanical waves. For some filter models, the filter's peak gain is fixed as the bandwidth varies, or can be independently controlled. A better property is that the peak gain should vary in a natural way along with the bandwidth, so that only one level dependence needs to be modeled. Models that can be described in terms of poles and zeros, in which the pole damping is level dependent, can provide such a natural coupling.

6. Stable low-frequency tail—when the parameters of an auditory filter are varied with signal level, the response of the low-frequency tail of the filter will ideally not change much. This constraint presents a problem for gammatones and gammachirps, though there are implementation approximations that can avoid the problem. The all-pole filters and filter cascades have an inherently stable low-frequency tail even as the pole dampings are varied to vary the peak gains and bandwidths by large amounts. A level dependent peak gain with a stable low-frequency tail corresponds to more asymmetry at high signal levels, which is what is observed.

7. Ease of implementation as digital filters—a simple description does not always mean it is easy to implement a bank of auditory filters and run sounds through it for machine hearing applications. In order to make a good digital filter, the model either needs to be in terms of poles and zeros, or convertible to such a description, or approximated by such a description. The roex filters are not used here, as their approximations by digital filters would be rather expensive. The gammachirp needs an ad-hoc approximation to be implementable, but is tolerably efficient. The real gammatone turns out to have a simple pole–zero description, though it was not known or exploited until the time of the Van Compernolle (1991) and Slaney (1993) analyses that also independently led to the simpler all-pole gammatone. The filter cascades were designed with digital implementation in mind from the beginning, so have simple two-pole–two-zero stages.

8. Connection to underlying traveling-wave hydrodynamics—other than the cascades, most filter models are just phenomenological, or descriptive of abstract filters. The filter-cascade family, on the other hand, was developed to connect with the mathematics of filtering by wave propagation, via the WKB method (Lyon, 1998), and to have at least the potential to be tied directly to wave-propagation parameters associated with the underlying cochlear mechanics, whether modeled or measured.

9. Realistic impulse-response timing and phase characteristics—if a filter's magnitude frequency response is reasonable, its delay and phase characteristics may be unimportant for many applications; but for comparison with physiological measurements, across a range of levels, these details can be diagnostic of whether the model is realistic, and hence can be relevant in knowing whether the model can contribute to an understanding or explanation of the mechanics.

10. Dynamic—in addition to being parameterized by level, the filter will ideally support being made dynamically variable, so that it can be used for processing sounds that vary in level over time. All of the filters in a filterbank need to vary sensibly, neither all alike nor quite independently, in response to signals of different power spectra.

	roex			gammatones				cascades		
	(p)	(p,r)	(pl,pu,r)	GTF	GCF	APGF	OZGF	APFC	PZFC	PZFC+
1. Simple	fd	fd	fd	td	td	ld	ld	ld/s	ld/s	ld/s
2. BW control	+	+	+	+	+	+	+	+	+	+
3. Peak/skirts	–	*	*	+	+	–	+	*	+	+
4. Asymmetry	–	–	+	–	+	+	+	+	+	+
5. Gain variation	–	–	–	–	*	+	+	+	+	+
6. Stable tail	–	+	+	–	*	+	+	+	+	+
7. Runnable	–	–	–	+	*	+	+	+	+	+
8. Waves	–	–	–	–	–	–	–	+	+	+
9. Impulse resp.	–	–	–	–	+	+	+	–	–	+
10. Dynamic	–	–	–	–	*	+	+	+	+	+

Table 13.1: Scoring various auditory filter models on the ten criteria. The domains of simple description are frequency domain (fd), time domain (td), Laplace pole–zero domain (ld), and Laplace per stage (ld/s). The * represents partial credit: for the roex and APFC peak/skirts shape criterion, some control but not a great fit; for the gammachirp filter (GCF), various criteria have been met by useful pole–zero filter approximations.

Many of these criteria have been introduced by others, sometimes implicitly, in the long history of development of auditory filter models. Comments on how certain filter models do on some of the criteria have been made as well (Irino and Patterson, 1997).

Particular models are described in the following sections; a summary of how well different models satisfy the above criteria is presented in Table 13.1.

13.4 Representative Auditory Filter Models

13.4.1 Three Lines of Auditory Filter Development

Auditory filter research has over time developed three widely-used families of filter models: first, the rounded exponential (roex) family; later, the gammatone family, including gammachirp and all-pole variants; and most recently, the filter cascades, both all-pole and pole–zero variants. Each family has good properties, and applications for which it provides a useful solution.

An important use of auditory filter models is to support applications in which full filterbanks efficiently process real sounds; these applications motivate and benefit from the filter-cascade family of models, since these models minimize the total computational complexity of a filterbank, as opposed to the complexity of a single auditory filter channel. Their structural efficiency is the reason that the filter cascades have been the basis for most work in analog VLSI hearing models (Sarpeshkar, 2000).

The roex family is useful mostly as a descriptive model, a way to parameterize and describe the shape of an auditory filter's magnitude transfer function; it has no corresponding phase, no time-domain equivalent, and no "runnable" implementation.

To get to an efficiently realizable filter, the gammatone was developed, especially after it was noticed that its impulse response was a good match to the physiologically derived *revcor functions* (de Boer and de Jongh, 1978), first-order Volterra kernels estimated by reverse correlation from action potentials on single fibers in the cat auditory nerve (Schofield, 1985), and that its shape can be fairly close to the roex shapes (Patterson et al., 1988, 1992; Rosen and Baker, 1994).

The gammatone-family filters bridge the other two (roex and filter-cascade) families, being useful as filter shape descriptions, but also implementable as real analog or digital filters to process sounds. The basic gammatone, while very popular, is still not sufficiently accurate or controllable in the ways we want. Its

variations (see Chapter 9), the gammachirp filter (GCF) and the all-pole and one-zero gammatone filters (APGF and OZGF), are much better in these respects.

There are of course other lines of development as well, sometimes easy to relate to these, and sometimes not. We will stop short of more complex models, such as cochlear models that are not cast as filters, and particularly exclude from this work any models that are not unidirectional (from input to outputs), even though models that include reverse cochlear traveling waves and *otoacoustic emissions* (sound emissions from the ear) have their own important roles.

Within the three families, we delineate several specific filter models to discuss in detail, to see how they rate on the measures introduced above. The results, summarized in Table 13.1, position the PZFC and its variants as promising filter models on which to build machine hearing applications.

13.4.2 Three Rounded Exponential Filters

The rounded exponential or roex filters can be seen as an effort to turn the "cartoon" of a triangular filter shape (Glasberg, 1982; Formby, 1990) into something quantitatively reasonable, by rounding the peak and specifying the skirt shape in terms of frequency deviation from the peak frequency.

The roex(p) filter has just one shape parameter, p, which can be thought of as either a bandwidth or a skirt steepness; shapes for two p values are illustrated in Figure 13.1. The roex(p, r) adds a parameter to control the skirt shape. To get asymmetry, different p parameters for the low side and high side allow control of asymmetry in the roex(pl, pu, r) filter. Other variations and combinations have also been tried; see Figure 13.2.

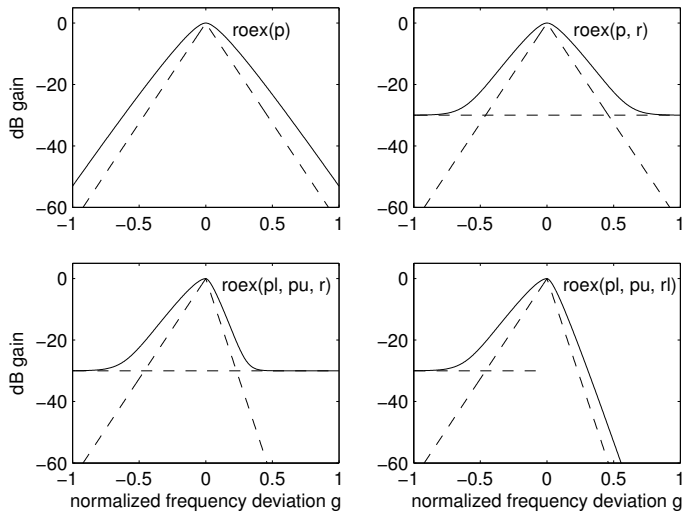


Figure 13.2: The roex family can take various parameters, including some illustrated here. In the roex(p) model (upper left), the factor $1 + p|g|$ rounds the filter's power gain above a symmetric triangular skeleton. In the roex(p, r) model (upper right), a floor at gain r is imposed (here at $r = 0.001$). In the roex(pl, pu, r) model (lower left), the upper and lower sides have different slopes, but the same floor. Allowing separate floor levels, or using no floor on the upper side, as Rosen and Baker (1994) did, allows more realistic asymmetric filter shapes (lower right) (we do not treat this as a separate model here, just a different parameterization of the asymmetric version). All of these are still too peaky (not rounded enough) near the peak, compared to more realistic models.

The roex power-gain shape, as a function of normalized offset frequency g (offset from center frequency, divided by center frequency), is formed as the product of an exponential skirt term and a factor $1 + p|g|$

that rounds the peak where the lower-side and upper-side skirts join. The p parameter can be interpreted as controlling the relative bandwidth, or the skirt steepness:

$$\text{roex}(p)(g) = (1 + p|g|)\exp(-p|g|)$$

Mixing this shape with a constant floor level r limits the dynamic range (Patterson et al., 1982):

$$\text{roex}(p, r)(g) = r + (1 - r)(1 + p|g|)\exp(-p|g|)$$

Treating the lower and upper sides independently, with width factors p_l and p_u, allows the introduction of asymmetry, for better fits to human psychophysical data; the r term has typically been kept the same on upper and lower sides (Glasberg and Moore, 1990), though it would probably work better to let it be zero on the high side and just use r to model the low-frequency tail.

The roex filters are expressed as power-gain functions, as opposed to amplitude-gain frequency responses, since they are used for weighting noise power and don't correspond to a real filter (Patterson et al., 1982). It would also be possible to use the roex shape as amplitude, and its square for power, which would give a more rounded and better-fitting peak. In the original conception of the roex with a more general polynomial factor, a more rounded peak was also possible (Patterson and Nimmo-Smith, 1980).

There are further roex variations that we won't discuss specifically, such as the roex(p, w, t), a weighted combination of roex shapes with different exponential slopes (Patterson et al., 1982), and a two-sided version of it with six parameters (Rosen et al., 1998). These variations can provide more shape control, but are not qualitatively different from the others in their properties. Rosen, Baker, and Darling (1998) have expressed the opinion that these parameterizations may be *too* flexible.

13.4.3 Four Gammatone-Family Filters

The gammatone filter (GTF) has been hugely popular, mostly due to its simple description in the time domain as a gamma-distribution envelope times a tone. It has been implemented (made runnable) through a number of methods and approximations, but usually not in the most straightforward way based on its Laplace-domain pole–zero decomposition, since that decomposition was not widely understood until at least the mid 1990s. The GTF has an inherently very nearly symmetric frequency response, which is not a great match to auditory data, but has better peak/skirt shape than the roex filters.

The equations for the gammatone-family filters have been detailed in Chapter 9. The resulting filter shapes are compared in Figure 13.3. They are generally not as simple as the roex equations, since these filters are specified in other domains—either as impulse responses or as poles and zeros.

The gammachirp filter (GCF) (Irino and Patterson, 1997) is a generalization of the GTF that allows a realistic and controllable frequency-domain asymmetry, and a corresponding realistic time-domain "chirping," or a "glide" in the instantaneous frequency of its impulse response. But it does not have a pole–zero decomposition, and needs other approximations in its implementation.

The all-pole gammatone filter (APGF) is another approach to providing a realistic asymmetry, while at the same time simplifying the Laplace-domain description and implementation of the GTF. The APGF has been used as an approximation for implementing the GTF (Van Compernolle, 1991; Slaney, 1993), but has advantages of its own in terms of asymmetry and impulse response (Lyon, 1997).

The one-zero gammatone filter (OZGF) is a slight generalization of the APGF that by adding a single real zero in the Laplace domain achieves good control of the low-frequency tail shape (Lyon, 1997; Katsiamis et al., 2009).

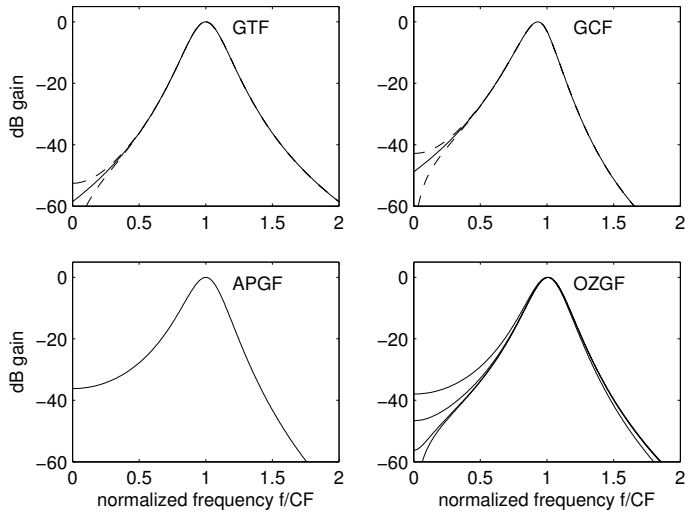


Figure 13.3: Auditory filter model shapes in the gammatone family include the real and complex gammatones (upper left) and gammachirps (upper right)—the real versions at several phases shown dashed—the all-pole gammatone filter (lower left), and the one-zero gammatone filter (lower right). The OZGF has explicit control of the low-frequency tail shape by the zero parameter, whereas the real GTF and GCF tail shapes vary as a side effect of other parameters.

13.4.4 Three Filter Cascades

The three filter-cascade models that we investigate are the APFC, PZFC, and PZFC+ models.

The all-pole filter cascade (APFC) (Lyon, 1997, 1998) is the basis for most silicon cochlea work (Lyon and Mead, 1988a; Watts et al., 1992; van Schaik et al., 1996; Sarpeshkar, 2000). This type of cascade typically has difficulty achieving a narrow enough bandwidth, and has an unrealistically long delay if tuned for reasonable frequency-domain shape.

The pole–zero filter cascade—PZFC, based on the "two-pole, two-zero sharper" filter stage of Lyon (1998)—has a much more realistic response in both time and frequency domains, due to its better approximation of the underlying wave mechanics, and is not much more complicated. The additional degrees of freedom from the placement of a zero pair near the pole pair allows the tailoring of response properties, such as the delay and high-side steepness of the filter, while retaining the other desirable features of the all-pole filter cascade.

For the purposes of our filter fitting experiments, the PZFC model is one in which the zeros are fixed while the poles move in reaction to the signal level. After exploring other pole–zero filter cascade variants, we settled on what we call the PZFC+ as an especially good filter model in terms of impulse response zero-crossing stability. In this model, the zeros in each stage move along with the poles, staying aligned at a proportional Q or damping factor.

A cascade built from filter D of Section 8.6 is a PZFC+, in that the zeros move with the poles. It makes a filter with less zero movement than the typical fitted PZFC+, in that the zeros move exactly as far as the poles do, rather than moving somewhat more as the PZFC+ typically prefers. We call this constrained PZFC+ a PZFC3 (the names are designators inherited from our experimental filter fitting software). An alternative constraint is to keep the zero damping equal to (not just proportional to) the pole damping; this constrained version is closer to the typical fitted PZFC+, and we call it the PZFC2 variant. In this variant, the zeros move more than the poles do, by the same factor that their frequency is higher (typically a factor of $\sqrt{2}$ or a fitted value near there). In the most general PZFC+, the damping proportionality factor is a fitted parameter of the

model.

The shapes and controllability of the PZFC and PZFC+ are not significantly different from those of the APGF and OZGF illustrated in Figure 13.3. These models are structurally simple, but do not have simple equations for their spectral shapes or for their impulse responses. Subsequent chapters, especially Chapter 16, relate these filter-cascade models to our CARFAC cochlear model.

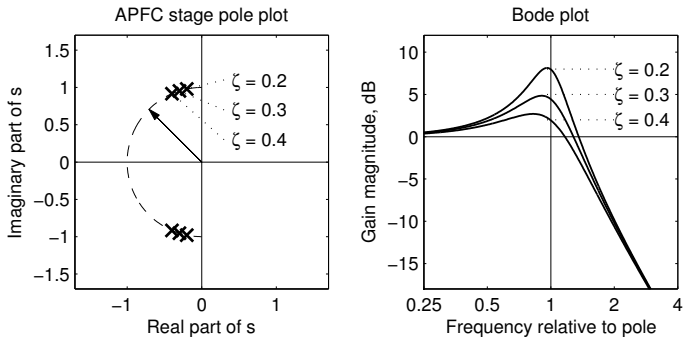


Figure 13.4: Diagram of the level-dependent motion of the poles of an APFC stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The positions indicated by crosses in the s-plane plot (left) correspond to pole damping ratios (ζ) of 0.2, 0.3, and 0.4. Corresponding transfer function gains (right) of this resonator stage do not change at low frequencies, but vary by several decibels near the pole frequency.

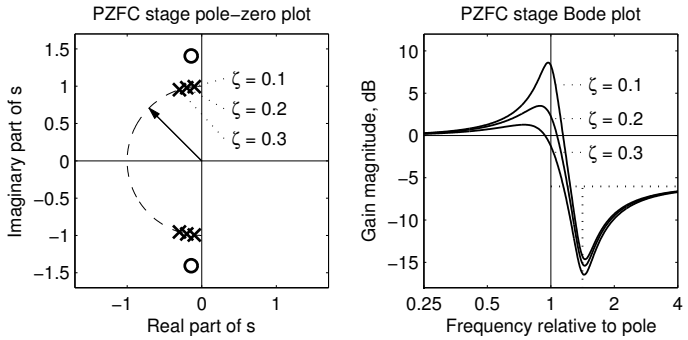


Figure 13.5: Diagram of the motion of the poles of a PZFC stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The positions indicated by crosses in the s-plane plot (left) correspond to pole damping ratios (ζ) of 0.1, 0.2, and 0.3, while the zero's damping ratio remains fixed at 0.1. Corresponding transfer function gains (right) of this asymmetric resonator stage do not change at low frequencies, but vary by several decibels near the pole frequency. The fact that the stage gain comes back up after the dip has little effect in the transfer function of a cascade of such stages.

13.5 Complications: Time-Varying and Nonlinear Auditory Filters

Linear time-invariant filters are the foundation for all frequency analysis including auditory modeling; they have the property that they can be completely described in terms of their gain (including phase) for sinusoidal inputs. They produce at the output only the same sinusoid frequencies present at their inputs. More complex wave shapes can be described as sum of sinusoids and the responses to each component can be simply

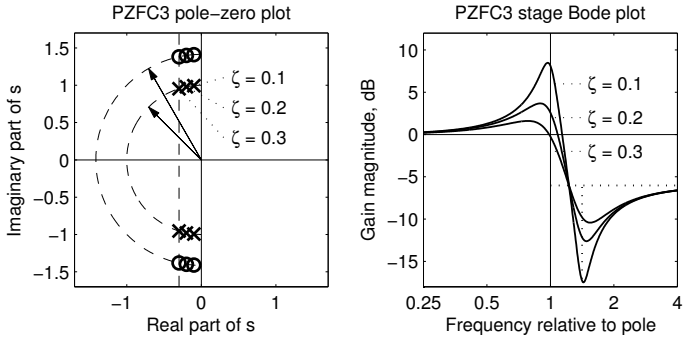


Figure 13.6: Diagram of the motion of the poles and zeros of a PZFC3 stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The zeros move along with the poles in the PZFC+; they are constrained to the same real-part value in this PZFC3 variant.

added to get the response to the complex wave. Yet, very early, Fletcher (1924) had commented on auditory nonlinearities, "the ear displays a nonlinear response to external applied forces. This nonlinearity produces subjective tones; all the summation, the difference, and the harmonic frequencies as well as the impressed frequencies produce nerve stimulation."

The auditory filter can be approximated as linear, but the fact that we can't explain hearing fully in terms of sine waves is also telling us that something involved in hearing is nonlinear. We know that the auditory system does not respond linearly: doubling the input does not double the response, even at the earliest levels of processing in cochlear mechanics. We find changes of gain and changes of bandwidth as a function of sound level; we hear, and find on the basilar membrane, frequencies not present in the input. Some of these nonlinear and level-dependent effects occur in the auditory periphery, in the mechanisms that the auditory filter is intended to model, and can be modeled on top of linear filters as a foundation.

Linear filters can be parameterized in many ways. If we allow some of these parameters to be not constant, but time-varying, then we have a nonlinear filter, or a linear time-varying filter. The *compressive gammachirp* (Irino and Patterson, 2001) is one such level-parameterized filter, an approximation to the gammachirp using dynamically movable poles and zeros. The all-pole gammatone, one-zero gammatone, all-pole filter cascade, and pole–zero filter cascade are more easily given a compressive nonlinear response via movement of their poles (Lyon, 1997).

If the parameters vary slowly, compared to the frequencies in the sound, in a way that depends on the sound itself, then the filters will act about like linear time-invariant filters, but level dependent; when we look at the models as level-dependent linear filters, we call them quasi-linear. The output will still not contain frequencies not present in the input, if the level-dependent parameters vary slowly enough. For many purposes, level dependence may be enough, though we know that other frequencies—distortion products—can sometimes be important as well, as reviewed in Chapter 4.

For the purpose of this chapter, we consider only quasi-linear filters as auditory filter models; but there are extensions possible. Nonlinear filters can be complicated, but a limited class is enough for many purposes: linear filters cascaded with memoryless nonlinearities. These can provide both an overall nonlinear response as a function of level, and a reasonable amount of nonlinear distortion products. This class of filters has been applied to auditory problems as bandpass nonlinearity (BPNL) models (Pfeiffer, 1970; Duifhuis, 1976) that sandwich a nonlinearity between a pair of linear filters, and as the dual-resonance nonlinear (DRNL) models of Goldstein (1990, 1995), Meddis et al. (2001), Lopez-Poveda and Meddis (2001), and Sumner et al. (2003b) that are built around several gammatone filters, incorporating a linear "tail" filter in parallel with a BPNL "tip" filter.

Kim, Molnar, and Pfeiffer (1973) introduced a model that incorporated ten stages of two-pole filters modified to have nonlinear damping terms in their differential equations. In the small-signal linear limit, their system is a 10th-order all-pole filter. It is close to an APGF, but the 10 stages have their natural frequencies decreasing at 3% per stage (over a total range of less than a half octave), so it is also a short piece of an all-pole filter cascade (APFC). The distributed nonlinearity was motivated by hydrodynamic wave propagation, so it resembles a nonlinear APFC in that respect, as well. At the time, with borrowed time on a PDP-12 minicomputer, ten stages with one output was all they could simulate. Motivated partly by Kim et al.'s model, Lyon and Mead (1988a) extended this system to a full multi-output APFC analog VLSI cochlea using nonlinear two-pole stages.

In a system with AGC, a feedback loop works to help keep the output level from varying too much. We apply the term AGC to auditory filters because the models fit best when the level dependent parameters are controlled by the filter output level.

13.6 Fitting Parameters of Filter Models

Auditory filter models can be fitted to experimental data in several ways. The filter's magnitude-gain frequency response can be optimized to predict human masking data, or the impulse response, including phase and delay, can be optimized to match certain properties of animal auditory nerve data. We call such optimizations *fitting*: adjusting the model parameters such that the filter model provides good matches to, or predictions of, the outcomes of the experiments.

13.6.1 Fitted Psychoacoustic Filter Shapes

A number of teams have repeated and extended experiments on human detection of tones in asymmetric notched-noise maskers (Patterson, 1976; Patterson and Nimmo-Smith, 1980; Patterson et al., 1982; Lutfi and Patterson, 1984; Glasberg et al., 1984; Moore et al., 1990; Rosen et al., 1998; Baker et al., 1998; Glasberg and Moore, 2000; Baker and Rosen, 2006). Others provided increasingly sophisticated analyses to derive auditory filter shapes that would predict the experimental data (Patterson and Moore, 1986; Moore and Glasberg, 1987; Glasberg and Moore, 1990; Rosen and Baker, 1994; Irino and Patterson, 2001; Patterson et al., 2003; Unoki et al., 2006).

We recently adapted their methods, and used their data, to assess how well our OZGF and filter-cascade models do at explaining the data, with very positive results (Lyon, 2011a). Each fitted model type was adapted to include AGC, that is, to have parameters controlled by its output level, as opposed to its input level, as suggested by Rosen, Baker, and Darling (1998), who contrasted models with parameters controlled by the masker level (essentially the input level to the filter) versus those controlled by probe level (the probe tone level being more nearly related to the output level of the filter):

> Models with filter parameters depending on probe level fit the data much better than masker-dependent models. Thus auditory filter shapes appear to be controlled by their output, not by their input. Notched-noise tests, if performed at a single level, should use a fixed probe level. Filter shapes derived in this way, and normalized to have equal tail gain, are highly reminiscent of measurements made directly on the basilar membrane, including the degree of compression evidenced in the input–output function.

Unoki et al. (2006) showed that the gammachirp variants can provide better fits with fewer parameters than the roex models. Lyon (2011a) found that in many cases the OZGF and PZFC can provide better fits with fewer parameters than the gammachirp models. In that work, we explored models with 3 to 14 fitted parameters, but also showed that many of them with more than a few parameters were overfitted, in the sense that they did not

Human Notched-Noise Masking Experiments

Human auditory filter shapes can be inferred from the results of an ingenious family of experiments on the detection of tones in noise, especially by using noise bands arranged asymmetrically above and below the tone frequency, as illustrated in Figure 13.7.

A *notched noise* consists of two frequency bands of noise with a quiet frequency band (the notch) between them. Such noises have been used as maskers in tone-detection experiments, to get at the filtering that the auditory system does, since the 1950s (Webster et al., 1952); the method became more important in the 1970s (Patterson, 1976; Patterson and Nimmo-Smith, 1980), after it became clear that listeners were employing an "off-frequency listening" strategy to detect masked tones. That is, the interpretation of experimental data was that in trying to detect tones, listeners were effectively paying attention to the filter channel with best signal-to-noise (SNR, or tone-to-masker) ratio, rather than to the channel with the filter's peak frequency matching the probe tone's frequency. Taking this effect into account, experiments with *asymmetric notched noise*, that is, using probe tones placed off-center in the notches, provided a way to better assess the effects of different parts of the auditory filter shape.

Detection thresholds in these tests are based on *two-alternative forced-choice* (2AFC) experiments, in which a probe tone is present in one of two stimuli, and the subject has to say which one. The tone detection threshold, for a given noise spectrum and level, is taken as the tone level for which subjects are 75% correct. Filter model parameters can then be fitted by optimizing, in a squared-error sense, the prediction of these experimental thresholds given the noise parameters.

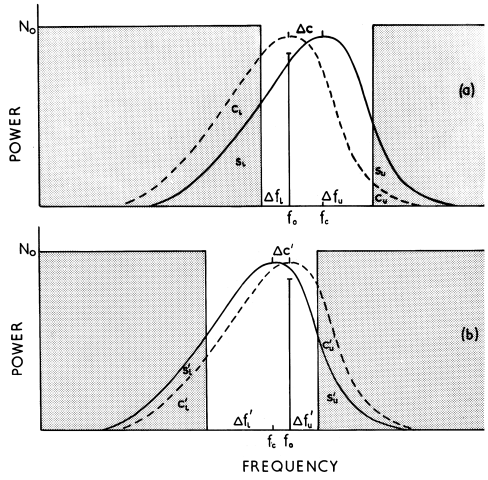


Figure 13.7: In the asymmetric notched-noise masking experiment, the presumption is that the auditory filters that are shifted to pick up the most favorable ratio of probe tone power to total noise power (the solid curves shown, as opposed to the dashed curves with their filter peaks at the probe tone frequency) are the filters that determine the tone detection threshold via a threshold signal-to-noise ratio. By using various different notch widths on the low and high sides of the probe tone frequency, as in the top and bottom plots here, the experiment's detection threshold data from human listeners provide indirect information about both sides of the filter shape. [Figure 1 (Patterson and Nimmo-Smith, 1980) reproduced with permission of AIP Publishing.]

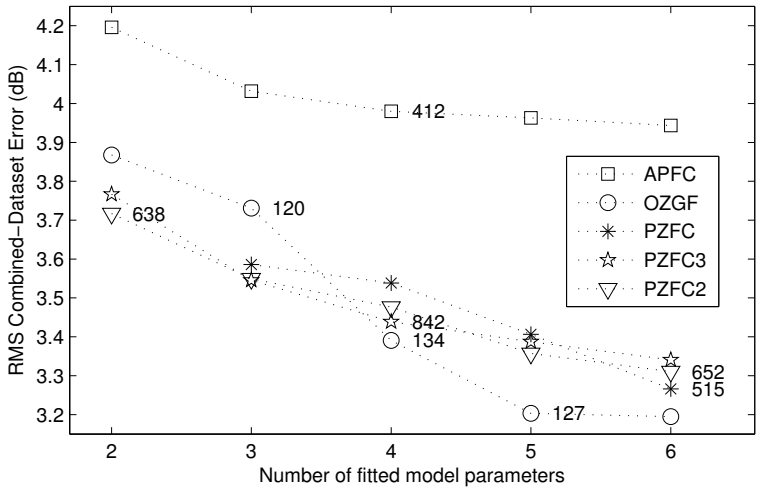


Figure 13.8: Threshold-prediction rms errors for several auditory filter models, versus number of fitted parameters, on the combined human masked-threshold datasets of Baker et al. (1998) and Glasberg and Moore (2000). The fit numbers are for reference only; different filter models are identified by different symbols, as shown in the legend. For each model type, only the fit with lowest error at each number of parameters is shown (at each number of parameters, several different parameterizations are possible within the model fitting framework). The errors are monotonically decreasing, since adding a free parameter never increases the error. The PZFC+ variants (PZFC3, star, and PZFC2, triangle) are the PZFC modified to have the zeros move with level, parallel with the poles, as opposed to the original PZFC (*) for which the zeros are fixed.

generalize well from a training dataset to an independent testing dataset. In this chapter, we therefore focus on models with relatively few parameters, from 2 to 6. We focus on the models that have corresponding efficient implementations as filterbanks, and therefore we do not include the roex and gammachirp families going forward. Since the 2011 work, we have found better combinations of few parameters; for example, with just four parameters, what works best for most models is one bandwidth parameter and a three-parameter quadratically frequency-dependent factor that controls how the filter output level changes the bandwidth.

In counting fitted parameters, we do not count structural choices (choice of model type and variant, such as PZFC3) nor parameters that are fixed at conventional values (such as $\sqrt{2}$ for the ratio of zero frequency to pole frequency in the PZFC family, or order 4 in the OZGF family). These conventions allow us to reduce the dimensionality of the parameter search space, but still allow us some flexibility in finding good models.

Our experiments confirmed that a filter architecture that gives a natural coupling of gain, bandwidth, and shape to level-dependent parameters provides a parsimonious model with no loss of realism (relative to the datasets used, at least). At the same time, the cascade architecture provides the stable low-frequency tail similar to that which had been added by developing compound structures (parallel or cascade) for the level-dependent roex and gammachirp models.

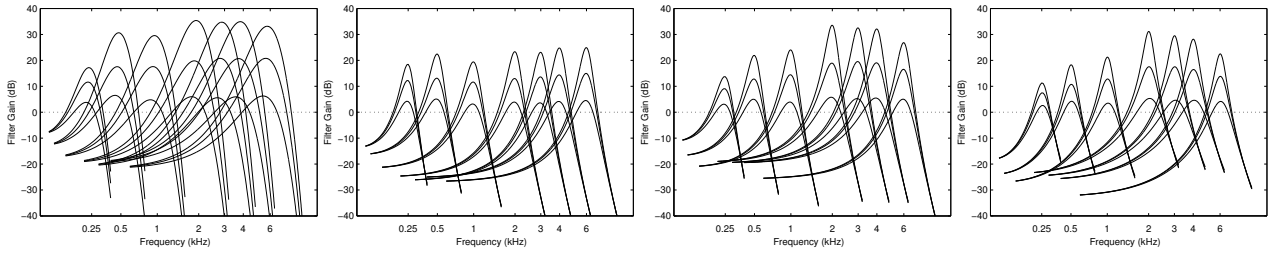
These experiments also confirm the usefulness of the AGC feedback configuration (Lyon, 1990; Carney, 1993), where the filter's own output is the signal whose level controls its parameters. The filter models based on feedback from the output always provided better fits with fewer parameters than the models with forward control from the input noise spectrum, the type of model that Unoki et al. (2006) analyzed.

In the typical alternative to using the filter's own output to control its parameters, others (Zhang et al., 2001; Unoki et al., 2006; Rosen and Baker, 1994; Tan and Carney, 2003) have used a linear (non-level-dependent) *control-path* filter whose output controls the parameters of the signal path. This approach can be easier to implement, as it is a feed-forward computation with no feedback loop, but the idea of a separate

control-path filter is hard to reconcile with the structure of the auditory system. Only feedback-based models are considered in this chapter.

13.6.2 OZGF and PZFC Variants Provide Good Fits with Few Parameters

We predicted that the "OZGF will provide a significant benefit in applications that need a better model of level dependence or a better low-frequency tail behavior" (Katsiamis et al., 2007). This prediction has been somewhat confirmed with respect to human masked-threshold data. As shown in Figure 13.8, with 4 to 6 parameters, the OZGF models provide the best fits to the experimental data. The other families can do as well or better (not shown) if we give them a one-zero parameter to explicitly control the low-frequency tail level. The 2–4 parameter OZGF fits do not have the zero, so they are the all-pole (APGF) variant, and don't fit as well as the ones with a finite real zero shaping the tail; with few parameters, the PZFC family does very well. In general, though, the differences between the models are not very significant; with the exception of the APFC, they all fit the data well, with an rms error of 3.2 to 3.9 dB. The resulting model filter shapes are illustrated in Figure 13.9 and Figure 13.10.



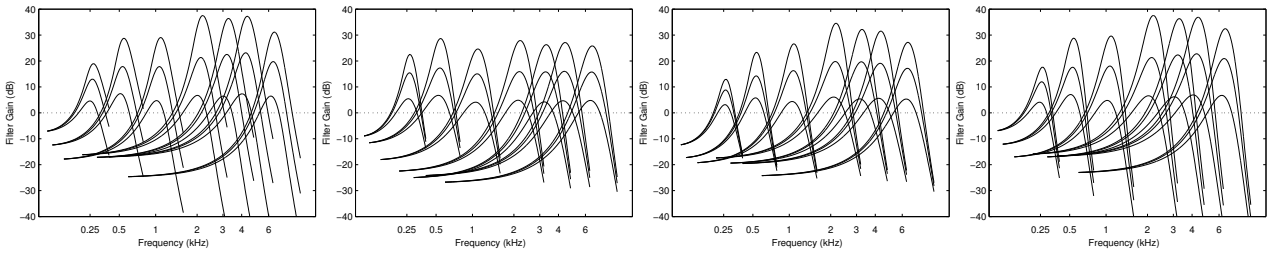
(a) APFC 412 (b) OZGF (APGF) 120 (c) OZGF (APGF) 134 (d) OZGF 127

Figure 13.9: Auditory filter gain plots for an all-pole filter cascade and for some OZGF model types, including APGF (the one zero moved out to infinity to make it an all-pole filter). The frequency axes are on the ERB-rate scale. In each case, the curves represent filter gain when the tone detection thresholds are 30 dB (highest curves), 50 dB, and 70 dB (lowest curves). The curve spacing is related to the input–output compression: curves close together, as at 250 Hz, correspond to a response that is only slightly compressive, while curve tips 15 dB apart represent a 4:1 compressive response. The APFC is not competitive in terms of prediction error, since its bandwidth is too large and its high-side rolloff too steep. Compare these shapes with the conceptual level-dependent filter description of Figure 10.1 D, and with the measured mechanical responses of Figure 10.2 D.

It should be noted that the OZGF filters are very close to what Flanagan (1960) proposed for modeling basilar membrane motion ("a function having a simple zero at the origin and third-order, complex-conjugate poles was considered"), though making the parameters level dependent was not foreseen at that time, and his filters were not recognized as being related to gammatones until much later (Lyon, 1997).

13.7 Suppression

A challenging problem for cochlea models is to explain suppression as observed in auditory physiology: a suppressor tone at one frequency can reduce the neural or mechanical response to a probe tone, near the best place for the probe, with the suppressor being either higher or lower frequency than the probe. AGC modifying a frequency response was proposed as a model of suppression by Allen (1981), to explain his two-tone data in cat auditory nerves:



(a) PZFC 515 (b) PZFC2 638 (c) PZFC2 652 (d) PZFC3 842

Figure 13.10: Auditory filter gain plots for several PZFC model types, including the PZFC2 and PZFC3 versions of the PZFC+ with different constraints on the zero positions relative to the pole positions.

> Apparently the effect of the second tone is to modify the frequency response of the neural signal. The source of this phenomenon has been widely studied and is called two tone suppression. It is related to the nonlinear basilar membrane response as seen by Rhode..., and is, I believe, characteristic of an automatic gain control. Understanding and modeling this nonlinear effect is at the forefront of modern cochlear research.

Neural synchrony suppression is easy to explain by local saturating nonlinearities, but rate suppression is a bit harder. Ruggero, Robles, and Rich (1992) have shown that two-tone neural rate suppression has its basis in the mechanics of the cochlea. Some nonlinear filter models, such as the MBPNL (multiple bandpass nonlinearity) model of Goldstein (1995), have been able to demonstrate such a two-tone suppression, by incorporating a low-frequency bandpass with an expansive nonlinearity combined with CF bandpass filters before and after with a compressive nonlinearity. But there is no physical analog of such an expanded-dynamic-range intermediate response in real cochlear mechanics.

An alternative is to model cross-place coupling, such that a strong response near the best place for the suppressor can reduce the gains remotely, significantly reducing the response to the probe near its own best place. The CARFAC structure provides such cross-place coupling, both via the filter cascade itself, in which reducing the gain of a stage affects responses at places more apical, but also through the coupled AGC feedback network that models the architecture and function of the efferent system that controls the activity of outer hair cells basal to the places that best excite the efferents (Kim, 1984; Lyon, 1990; Warr, 1992).

Presumably, suppression is an important contributor to the masking that is observed via psychophysical experiments, and ought to be explainable in the same modeling framework. The exact pattern and dynamics of coupled AGC feedback should therefore be fitted to data on suppression, but tone-in-noise threshold datasets are perhaps not rich enough to support such fitting. Irino and Patterson (2006a) found good fits to suppression data by taking the level control for their dynamic compressive gammachirp from a channel of higher CF—but this asymmetry may be less needed in the CARFAC due to the filter cascade's propagation of effects toward the apex. Suppression in an earlier cascade–parallel model was reported by Lyon and Dyer (1986); suppression above and below CF was demonstrated, but not calibrated or compared to data.

13.8 Impulse Responses from Physiological Data

In neural experiments, we can estimate impulse responses—really first-order Volterra kernels—by the process of *reverse correlation*: every time the neuron fires an action potential in response to a sound, a piece of the noise waveform that led up to it is added to a waveform accumulation buffer. If the sound is a white noise, the shape of the sum in the buffer approaches the effective time-reversed impulse response of the cochlea at

the point innervated by the neuron, as explained by de Boer and de Jongh (1978). These correlation-derived impulse responses are called *revcor functions*.

We want filter models whose impulse responses resemble the neural revcor data, or corresponding mechanical impulse-response data. Indeed, the gammatone model was introduced as a simple approximation to revcor functions measured in the cochlear nucleus of cats (Johannesma, 1972).

Data from mechanical and neural experiments (Carney et al., 1999; Robles and Ruggero, 2001b; Shera, 2001) show that the zero-crossing times, or local phases, of the filter's output in response to impulses are unequally spaced—unlike the equally spaced zero-crossings of the gammatone—and do not change much with signal level even as the filter gain and shape change. This observation puts an important constraint on how the auditory filter model should behave as its level-dependent parameters are varied. The unequal zero-crossing spacing, corresponding to a chirping or gliding instantaneous frequency of the impulse response, and the relative stability of the zero crossings with level, have been key properties that researchers have used in recent years to assess the realism of cochlear models (Tan and Carney, 2003; Temchin et al., 2011).

In the case of the gammatone, gammachirp, and APGF models, the zero-crossing times of the impulse responses remain exactly fixed as the exponential decay time parameter is varied; this variation corresponds to moving the poles of filters horizontally (varying real part) in the s plane. In the case of gammachirp (and its special case, the gammatone), this stability of zero crossings is apparent from the time-domain description in which a decay-time-dependent envelope multiplies a fixed oscillating term that determines the zero crossings, as has been pointed out by Irino and Patterson (2001) when they fitted gammachirp filters to both human masking data and cat auditory nerve impulse responses:

$$h_{\mathrm{gcf}}(t) = t^{N-1} \exp(-\gamma t) \cos(\omega_{\mathrm{R}} t + c \log(t))$$

In the case of the APGF, a similar relationship is apparent when the impulse response is written in a similar way, which involves a Bessel function in place of the sinusoid:

$$h_{\mathrm{apgf}}(t) = t^{N} \exp(-\gamma t)\, j_{N-1}(\omega_{\mathrm{R}} t)$$

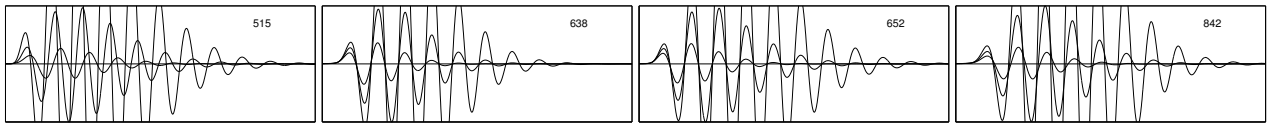
where j_{N-1} is a spherical Bessel function—see Abramowitz and Stegun (1972) transform 29.3.57 in combination with property 29.2.12.

Shera (2001) has also shown that this direction of pole motion in basilar-membrane-impedance models leads to nearly fixed zero-crossing locations.

For the gammatone, APFC, OZGF, PZFC, and other filters representable as rational transfer functions, the zero crossings are exactly fixed if the poles and zeros are all moved horizontally in the s plane by equal amounts. This observation follows from the shifting property of the Laplace transform, which says that shifting the Laplace transform by d corresponds to multiplying the impulse response by $\exp(dt)$. For real d, corresponding to horizontal movement, this change of envelope will not affect the zero crossings; it corresponds to adjusting the real γ in the factor $\exp(-\gamma t)$ in the above equations.

In some systems, it may be more natural to vary the damping, or pole Q, leaving the poles' natural frequencies fixed, in which case the poles move along a circle in the s plane, centered at the origin and of radius equal to the natural frequency ω_n. This is what we have done in our PZFC implementation; for auditory filter model parameter fitting, it makes no difference, since the optimal CF is selected for each data point. When damping is near zero, horizontal motion is nearly tangent to the circle, so these directions are not so different. But they may be different enough to make a testable difference in how well a model matches the observed zero-crossing stability. Moving the zeros by different amounts from the poles can approximately compensate for the effect of moving along nonhorizontal trajectories, at least in the early part of the impulse response. In the long-time limit, the decaying impulse response will ring at the ringing frequency of the pole with the longest time constant. That is, zero crossings will be determined by the imaginary part of the pole

with real part closest to zero, which in a cascade will be the poles of the last stage before the output tap.



(a) PZFC 515 (b) PZFC2 638 (c) PZFC2 652 (d) PZFC3 842

Figure 13.11: The impulse responses for the 1 kHz channel of three versions of the PZFC, at noise levels corresponding to the three tone threshold levels 30, 50, and 70 dB SPL. The large (off-scale) curves are for the noise level that leads to 30 dB SPL tone threshold, the medium (full-scale) curves for 50 dB, and the small curves for 70 dB. The base PZFC (left) has nonmovable zeros, and up to about 180 degrees of phase shift of the zero crossings between the high and low levels. The PZFC2 models allow the zeros to move more than the poles do, keeping the zeros' Q or relative damping the same as the poles, and achieves stable zero-crossing times. The PZFC3 variant (right) constrains the zeros to follow the poles horizontally, at the same real part of s-plane location, as in filter D of Section 8.6; this much zero movement limits the zero-crossing level dependence to about 45 degrees.

In the filter-cascade models, we assume that poles and zeros of the different stages move in a coordinated way, but in amounts proportional to their frequencies, so the shifting property does not exactly apply. Nevertheless, by careful choice of pole and zero motion directions and amounts we can achieve stable zero crossings, as illustrated in Figure 13.11. The first fitted PZFC model, in which the zeros are fixed and the poles move, does not achieve stable zero crossings—the zeros need to move about as much as the poles do. In the PZFC+ variant, the bandwidths of the zeros change in proportion to the bandwidths of the poles at each stage, with the constant of proportionality that can be 1 (for a PZFC3), or can be a fitted parameter that is optimized at about 1.4; the resulting fits to the masking data are not quite as good as the original PZFC is. In such a cascade, the zeros stay close to the poles of an earlier stage, approximately canceling out most of the effects of the cascade except for a few uncanceled poles in stages just basal to the place under consideration; the net filter is close to an all-pole model, and the fitting results are very close to the APGF or OZGF fitting results, as can be seen by comparing Figure 13.9 with Figure 13.10. Zero-crossing stability is not enforced, but the free parameter that determines how much the zeros move happens to give stable zero crossings in typical fits. Other ways of coupling the zero motion to the pole motion were not as good, in terms of zero-crossing stability.

Impulse responses and instantaneous frequency analysis from revcor and cochlear mechanics experiments also show a "glide" or "chirp" in the response of the cochlea, with an upward glide at high CF and a downward glide at low CF (Tan and Carney, 2003); this glide corresponds to the unequally spaced zero crossings mentioned above. In general, if the filters are minimum phase, the glide direction will be determined by the frequency response gain asymmetry; filters with a sharp high side will have an upward glide; that is, the initial cycles of the response will be at lower frequencies than the later cycles. To the extent that filter models get the asymmetry right, they will get the glide about right, as long as they are minimum phase, which most of the models we consider are (gammatone and gammachirp filters are not, but their complex versions, all-pole versions, and approximations are).

Another important aspect of impulse-response data is the group delay (see Section 12.6); again, delay is determined by the amplitude response in the case of minimum-phase models. The gammatone filters have their delay tied to filter shape and bandwidth, with higher orders providing more delay along with a somewhat different overall shape. The PZFC allows, by adjustment of the zeros, considerable room to tune the delay, to more or less than the delay of the typical order-4 gammatone-family filters. The PZFC has the property that as the cascaded segments are more finely divided (more stages per mm of cochlear place), the overall shape

and delay can be kept fixed by letting the zeros move closer to the poles. The all-pole filter cascade, on the other hand, will generally have too much delay and too much high-side slope at high numbers of stages per mm.

13.9 Summary and Application to Cochlear Models

Auditory filters—quasi-linear level-dependent models of cochlear filtering—provide the connection to linear system theory that makes many sorts of analysis and parameter-fitting possible. Some filter models are closer to full cochlear models than others. The early roex models provided a good advance over rectangular, resonance, and Gaussian filters; they provided a good spectral shape, but had no corresponding impulse response or pole–zero description. When an efficiently runnable filter was needed for machine hearing applications, the gammatone of order 3 to 5 was found to provide a shape fairly close to the roex, at the same time being efficiently realizable as a digital filter. When an asymmetric improvement was needed, the gammachirp and all-pole gammatone provided it. Finally, to take advantage of the forward wave-propagation structure of the cochlea and achieve a filter model that can be the basis for a full filterbank of appropriate shapes with a minimum of computation, the filter-cascade models were developed; the PZFC family was shown to provide most of the advantages of the other models as well. The filter cascades are not so easily described in either the frequency domain or the time domain, but have simple descriptions in terms of level-dependent pole and zero locations.

We have explained several criteria by which different auditory filter models may be compared. Our evaluation criteria may not be complete, but provide a starting framework in which the models presented, and new models to come, can be evaluated for various uses. By these criteria, the pole–zero filter cascade models provide an excellent base for work in machine hearing. In particular, these models correspond to efficient realizations that can be used as cochlear models to process arbitrary sounds, whereas some other models do not. To get to full cochlear models, they primarily need to be extended to include a good way to vary their level-dependent parameters dynamically in response to arbitrary sounds.

The better data-prediction fit of the PZFC and the approximately level-independent zero-crossing times of the PZFC+ are not achieved simultaneously, as they require different treatment of the positions of the zeros in the cascaded pole–zero filter stages. We have used the PZFC as the basis for a number of successful machine hearing applications, but are moving forward with a design with movable zeros. The more stable zero crossings will be important in binaural applications, where interaural phase should remain relatively independent of interaural level. In the CARFAC (see Chapter 16), the filter stage is based on PZFC3, a PZFC+ variant with the movement of zeros constrained by the structure of filter D; in this variant, the zero-crossings retain a small amount of level dependence, similar to what some recent studies find in real cochleae (Recio-Spinoso et al., 2009).

Chapter 14

Modeling the Cochlea

> ... the assumption of a 'passive' cochlea, where elements are brought into mechanical oscillation solely by means of the incident sound, is not tenable. The degree of resonance of the elements of the cochlea can be measured, and the results are not compatible with the very heavy damping which must arise from the viscosity of the liquid. For this reason the 'regeneration hypothesis' is put forward, and it is suggested that an electromechanical action takes place whereby a supply of electrical energy is employed to counteract the damping.
>
> — "The physical basis of the action of the cochlea," Thomas Gold (1948)

The cochlea is the ear's sound filtering and amplification system, consisting primarily of two fluid-filled channels, the membrane that separates them, and the tiny *organ of Corti* that sits along the edge of the membrane to amplify and detect motion.

How can we describe and replicate what the cochlea does? By models—especially by models that have a structure derived from the underlying physics and that have been validated against a range of measurements on real human and animal auditory systems. Models that are based on distributed linear system theory, plus appropriate nonlinearities, describe the cochlea well, and lead to efficient algorithms or machines that can replicate the cochlea's sound analysis function.

The system theory that we reviewed through Chapter 12 is all relevant—nonlinear, gain-controlled, distributed resonance-like filtering. We may conceptually look at the system response as an infinite number of outputs, at a continuum of locations, each representing a different nonlinear system—or more productively, we can focus on how the different locations build on each other, and make a more powerful and compact model through that approach.

In the cochlea, the membrane interacts with the fluid, constrained by the shape of the channel, to make a transmission line that supports mechanical *traveling waves*. Positions along this transmission line correspond to a large number of outputs, with a progression of different frequency responses, analogous to the old Helmholtz *resonance* view of cochlear function. As we emphasized before, the cochlea has important linear and nonlinear aspects; it even acts as a distributed amplifier, and adds energy to traveling waves to boost the response to weak sounds.

How we model this nonlinear sound preprocessor as a filterbank, by approximating the underlying physics as a nonlinear filter cascade, is the subject of this chapter.

The details of how the cochlea arranges to support a traveling wave, and how it amplifies and detects this wave, have been intensively studied for many decades. Aspects of that research are recounted and illustrated in this chapter, through our current modeling approach that is designed to capture enough of the cochlea's properties to support good machine hearing.

14.1 On the Structure of the Cochlea

Refer to the anatomical drawings of Figure 14.1.

Sound waves are coupled from the air in the outer ear canal into the fluid in the cochlea by the middle ear bones (*ossicles*): the hammer (*malleus*), the anvil (*incus*), and the stirrup (*stapes*). The footplate of the *stapes* pushes and pulls on the oval window (*fenestra vestibule* or *fenestra ovalis*), a flexible membrane that separates the fluid of the cochlea from the air space in the middle ear.

The cochlea consists of a cavity in the temporal bone, in the shape of a snail shell, divided down the middle by the *cochlear partition* into two main fluid-filled channels, the *scalae*. The scala that is directly driven by the middle-ear bones, which convey energy from the ear drum, or *tympanum*, is known as *scala vestibuli*. The other channel, which couples only to the air in the tympanic cavity of the middle ear, is known as *scala tympani*. The cochlear partition has two main regions: a stiff *bony shelf* (osseus spiral lamina) and a somewhat flexible, or springy, *basilar membrane* (BM).

Since the fluid in the cochlea is essentially incompressible, any fluid volume pushed in through the oval window must push out somewhere else; the round window (*fenestra tympani* or *fenestra rotunda*), connected to the nondriven channel, *scala tympani*, therefore bulges out when the oval window pushes on the fluid in *scala vestibuli*. The resulting fluid motion is approximately antisymmetric across the partition, with fluid motion up one channel matched by motion down the other. The forces, or pressures, involved in antisymmetric fluid accelerations create pressure differences across the cochlear partition, so its springy part, the BM, gets deflected, too, to the extent that its stiffness allows.

The BM is the mechanically important, somewhat stiff separator between the scalae, and supports the organ of Corti. The organ of Corti includes the assembly of outer hair cells that add energy to the traveling wave, and the inner hair cells that detect sound-induced motion. Associated with the organ of Corti is the *tectorial membrane* (TM), which overlies the hair bundles of the hair cells. Motion of the BM is the effective stimulus that the inner hair cells detect and convert to the neurotransmitter release that causes spiking of primary auditory neurons of the spiral ganglion, which send sound-evoked signals to the brain via the auditory nerve.

Scala vestibuli has an additional subregion, known as the *cochlear duct* or *scala media*, partitioned off by *Reissner's membrane*, a thin flexible membrane that has little mechanical effect, but that separates two types of watery fluid with important electrical properties: *endolymph* in scala media and *perilymph* in the other scalae. The ionic concentrations in these fluids are very different, and their electrical potentials (voltages) differ by a substantial 80 mV. This high-voltage region is sealed off from scala vestibuli by Reissner's membrane, from the bone by the *spiral ligament* and *stria vascularis*, and from most of the cochlear partition by the *reticular lamina*, a tough layer of *phalangeal cells*, the tough filamentous supporting cells that hold the hair cells.

Each hair cell sits with its hair-bundle end sticking through the reticular lamina, exposed to the endolymph with its positive potential and large concentration of potassium ions, and with their cell bodies surrounded by the more neutral potential of perilymph. These hair cells use their energy to pump positive potassium and calcium ions out, and they achieve a substantially negative internal potential of about -70 mV, or 150 mV different from the region immediately outside their end that holds the sensitive *cilia*, or *hairs*, that transduce motion. This 150 mV, known as the *endocochlear potential* or EP, is the largest potential difference found anywhere in the body, and it drives the sensitive and fast transduction that these cells achieve.

14.2 The Traveling Wave

In any medium, sound waves propagate through the continual exchange of energy between kinetic and potential forms. In air, sound travels as compression waves, with the motion of air mass carrying kinetic energy, and the compressibility of the air being the spring whose compression holds potential energy. In water (or

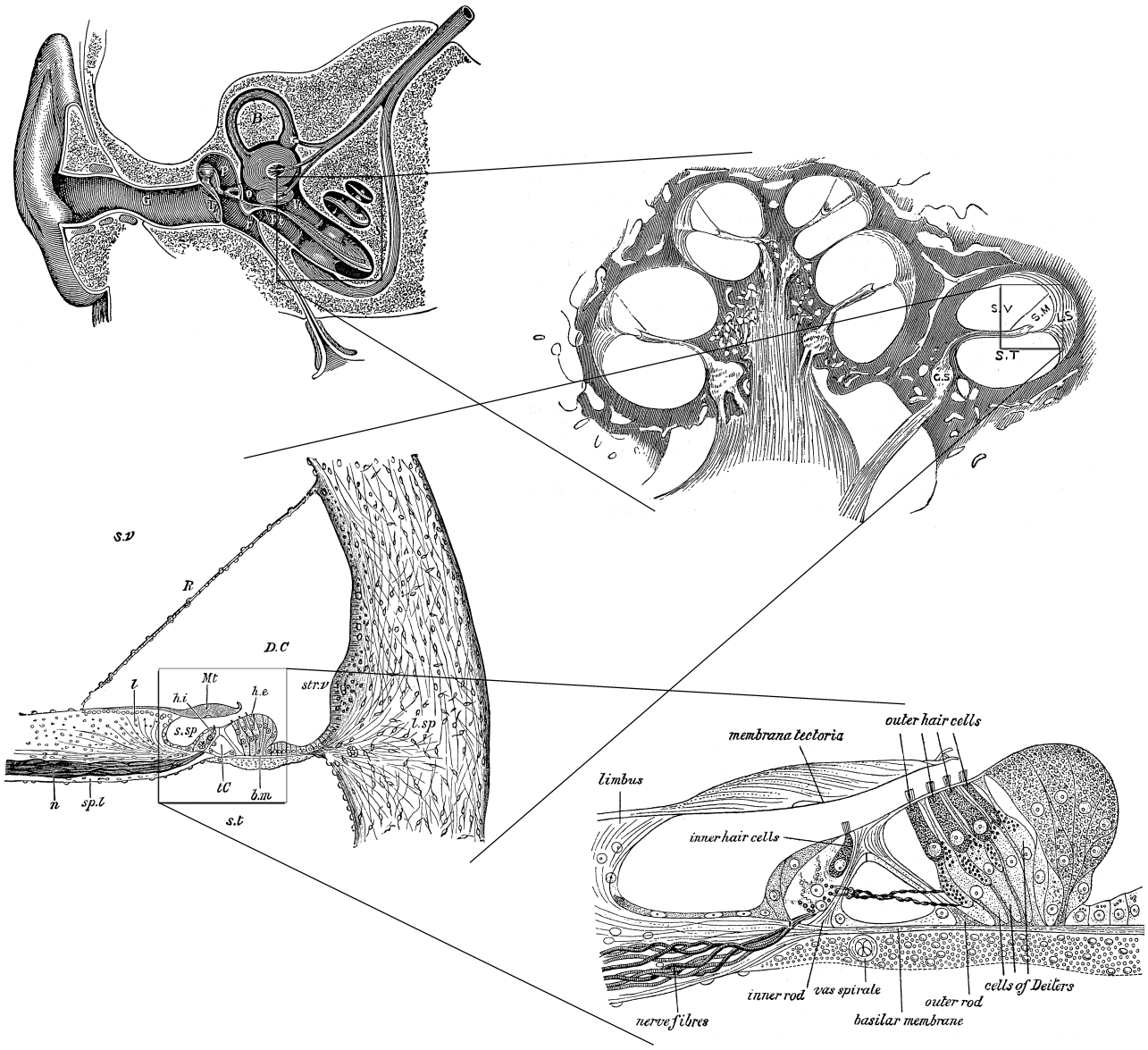


Figure 14.1: Four classic cross sections of the cochlea, from the macroscopic to the microscopic, with boxes and lines to show approximately how they relate.
Upper left: Leo Testut (1897) includes this drawing by Johann Czermak of the outer ear's sound path through the ear canal (G) to the eardrum, or tympanic membrane (T), and the middle ear bones that couple sound into the cochlea of the inner ear, via the oval window (O).
Upper right: *Gray's Anatomy* section through the cochlea. The structures that separate scala vestibuli (S. V.) from scala tympani (S. T.), in the region highlighted, are detailed in the next figure.
Lower left: This cross section through part of one turn of the mammalian cochlea, by Anders Retzius (1884), shows the cochlear duct (D.C, shown as scala media, S. M., in previous figure), scala vestibuli (s.v), scala tympani (s.t), basilar membrane (b.m), Reissner's membrane (R), tectorial membrane (Mt), nerve fibers (n), and the organ of Corti.
Lower right: This *Gray's Anatomy* drawing by Retzius shows a section through the organ of Corti, pointing out one inner hair cell and four outer hair cells.

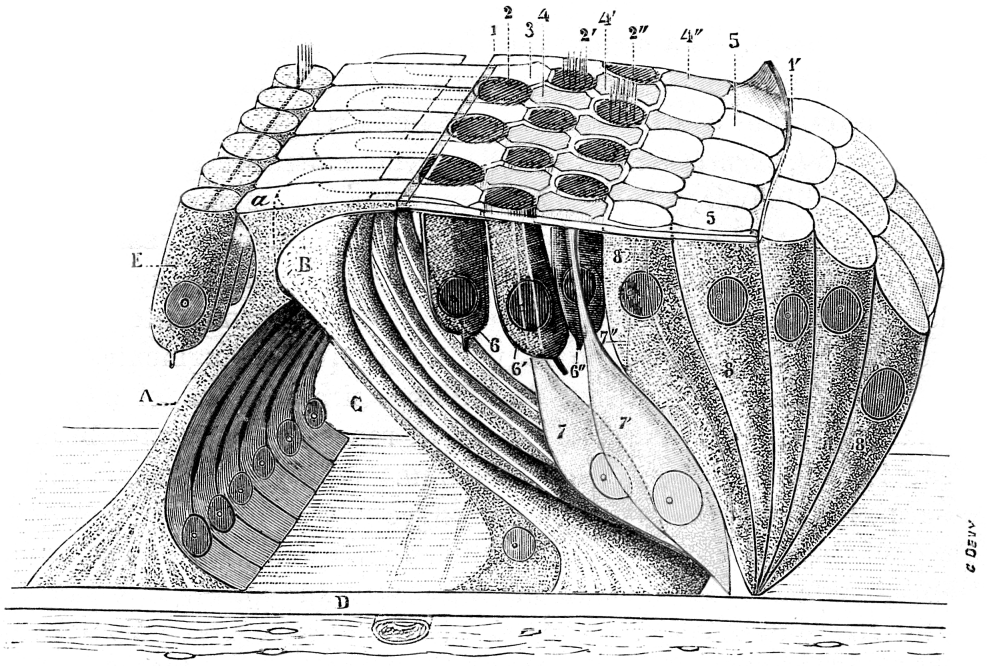


Fig. 918.

La même membrane, avec les cellules qui lui servent de substratum et dont l'empreinte lui donne son aspect réticulé (*schématique*).

Figure 14.2: The three rows of outer hair cells (dark; blue in the color plate) and one row of inner hair cells (E) sit with their upper ends and hair bundles exposed to endolymph in the scala media through the reticular lamina, but otherwise surrounded by a sealing barrier made up of the pillar cells (A and B), cells of Dieters (7), and other cells of the organ of Corti. The beautiful colored image was published more than 115 years ago (Testut, 1897).

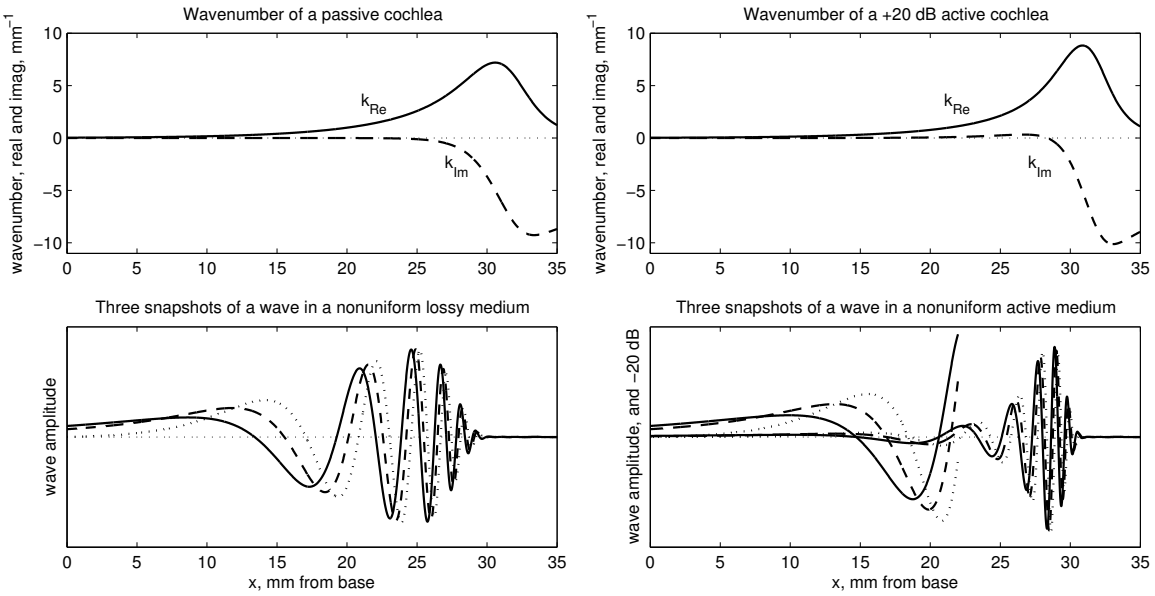


Figure 14.3: Three snapshots of a traveling wave in a passive cochlea (left), and in an active cochlea (right), responding to a sinusoid. The wavenumber, top, is estimated using the methods of Chapter 12, to correspond with our cascade of asymmetric resonators (CAR) model of Chapter 16. The slightly positive imaginary part of the one on the right corresponds to the active gain. The wave is calculated via the WKB approximation, at many more points than we would typically model in a filterbank. In the passive case, the amplitude peak is not very localized. To display the amplified signal in the active case, we cut its gain by a factor of 10 (–20 dB) after showing the part near the base that nearly matches the passive case. To get the large number of cycles from base to apex, we use 10 filter stages per mm, or 350 total, which is more than we would typically use in a machine hearing system (that is, the wavenumbers as plotted in mm^{-1} units are 10 times the natural log of the filter stage transfer functions).

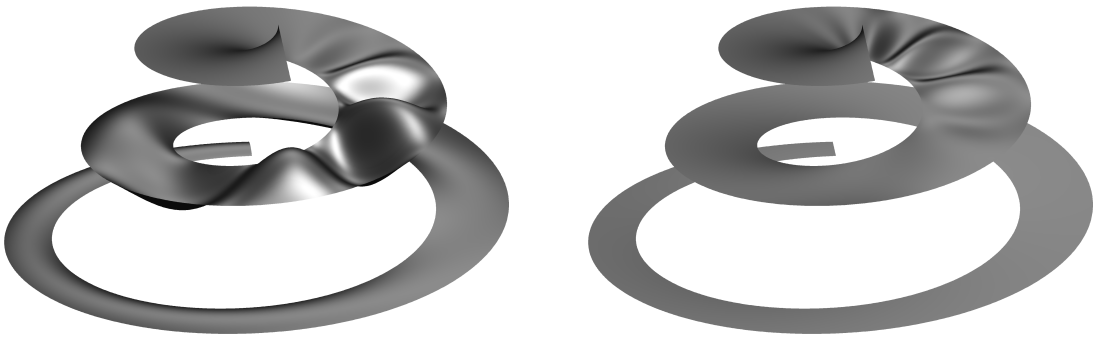


Figure 14.4: The traveling waves shown in Figure 14.3 are here mapped onto a 3D model of the basilar membrane, greatly exaggerated and stylized with colored lights (see color plates). The active case with 20 dB more gain (right) is rendered for a 30 dB lower input level, so it represents the response on the same scale with a factor of 1000 less input power, corresponding to a cube-root-compressive system (10 dB output level change for 30 dB input level change).

cochlear fluids), sound can also propagate similarly. Water is much less compressible and much heavier than air, so a fluid compression wave travels with higher pressures and lower displacements, and much faster, than in air.

Due to the very different displacements and velocities involved, sound in air couples to sound in water very inefficiently, even with the transformer action of the middle ear. In the cochlea, there is another wave mode that the middle ear couples to quite efficiently (Shera and Zweig, 1991): the fluid motion still makes the kinetic energy, but the potential energy is in the bending and displacement of the BM. It's much easier to displace fluid by bending the moderately stiff BM than to compress the fluid, so the motions are larger and pressures lower in this mode than in the fast compression wave. The wave speed is also much slower, even much slower than sound in air, so the wavelengths can be very short, even small compared to the dimensions of the cochlea. This short wavelength will show up as important in 2D and 3D modeling of the cochlea, as it has a lot to do with how the sound energy gets delivered efficiently to the organ of Corti.

When the stapes pushes on the oval window, fluid moves down the scala tympani, "through" the BM via deflection against its stiffness, and back down the scala vestibuli and out by stretching the *round window*, the relief connection to the air space of the middle ear. The BM doesn't simply bulge toward scala vestibuli, but propagates a wave, so that there are bulges up and down, traveling along the membrane. To the extent that the walls of the cochlea are rigid and the fluid nearly incompressible, the volume of fluid in each scala is held nearly constant, and the fluid displacements at the two windows are approximately equal and opposite.

For sounds of high frequencies, the traveling wave dies out before it travels very far through the cochlea. Lower frequencies travel further. For very low frequencies, the wave may get all the way to the end of the spiral, where there is a hole in the BM, known as the *helicotrema*, connecting the perilymph in the scala tympani with that in the scala vestibuli. This opening allows the average pressure in the scalae to be equal, so there's no average force bulging the BM. At very low frequencies, the volumes of fluid in the two chambers are not individually constant, though their sum is, as the BM moves up and down more as less as a whole, that is, with less than a half wavelength along it.

When Békésy observed the cochlear traveling wave, it was in a dead cochlea, so a necessarily passive system. Using the method of Chapter 12, and a plot in the style of Figure 12.8, we can estimate what such a wave might look like. See Figure 14.3 and Figure 14.4, which contrast the wave in a dead cochlea with the wave in a moderately active cochlea, using a pole–zero asymmetric resonator stage to define the wavenumber. The moderately active case corresponds to a typical speech sound level, around 70 dB SPL. In a healthy cochlea at low sound level, the localization of the wave response peak is sharper still, making it harder to illustrate the traveling nature of the wave in that case, since the amplitude near the peak is hundreds of times larger than it is in more basal regions that the wave travels through.

Traveling waves of motion at the cochlear partition move the hair cells relative to the tectorial membrane, mostly as a *shear* motion that bends the hair bundles back and forth, illustrated in Figure 14.5. Bending the hair bundles in one direction has little or no effect, but in the other direction opens channels that allow a rapid flow of positively charged ions into the cell. The resulting intracellular change in potential, known as the hair cell's *receptor potential*, is thought to be the driver of subsequent stages of reaction to sound, including mechanical length changes of outer hair cells and release of neurotransmitters by inner hair cells.

Fluids in the two scalae move with a symmetry of the sort that we see in a *differential* electrical transmission line such as shown in Figure 12.5, as opposed to a *single-ended* line. That is, fluid moves forward in one scala and backward in the other, by equal amounts. The fluid also has a component moving "up and down" with the membrane, in the same direction on both sides. Because of the symmetry, and when the frequency is not so low that we have flow through the helicotrema, we can typically ignore one scala, and treat the system as simply one fluid-filled chamber with a distensible membrane.

The propagation of different frequencies to different places is controlled by the changing mechanical properties of scalae and the BM. The cross sectional areas of the scalae start large and taper down, while the

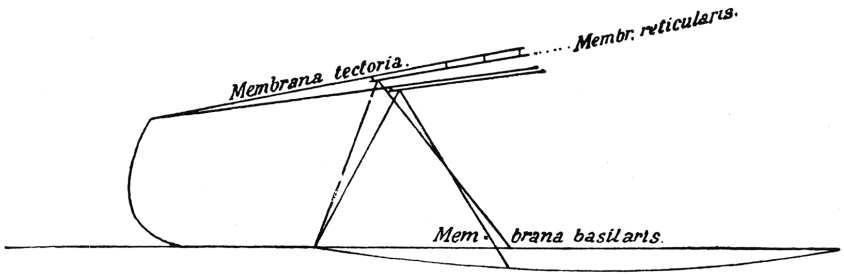


Figure 14.5: Diagram of how the hinged edge of the BM tilts the organ of Corti, causing a shear displacement between its top, the reticular lamina, where the hair cell cilia are, and the tectorial membrane (Kuile, 1900). The triangle represents the *pillar cells*, or *rods of Corti*, surrounding the tunnel of Corti.

BM starts narrow and stiff, and gets wider and more compliant (less stiff), for places moving from *base* (near the windows) to *apex* (near the helicotrema). As sinusoidal waves propagate from base toward apex, they slow down, and the BM displacement increases as the BM becomes more compliant. Beyond the frequency-dependent place where the wavelength gets quite short, the waves quickly damp out. As a result, there is a fairly well defined place of maximum response for each frequency, making a *frequency–place map*, also known as the *cochlear place map*.

In rough outline, this is how the structure of the cochlea leads to a filterbank-like function, with different places having best response to different frequencies. To understand it further, we investigate the mathematics of waves in nonuniform distributed systems using the techniques outlined in Chapter 12.

The behavior of the basilar membrane in response to sound is fairly well characterized by experimental data of many sorts (Robles and Ruggero, 2001a), but exactly how the wave on the basilar membrane interacts with the hair cells and the tectorial membrane, the problem known as cochlear *micromechanics*, remains somewhat open (Dallos, 2003; Cooper and Kemp, 2009).

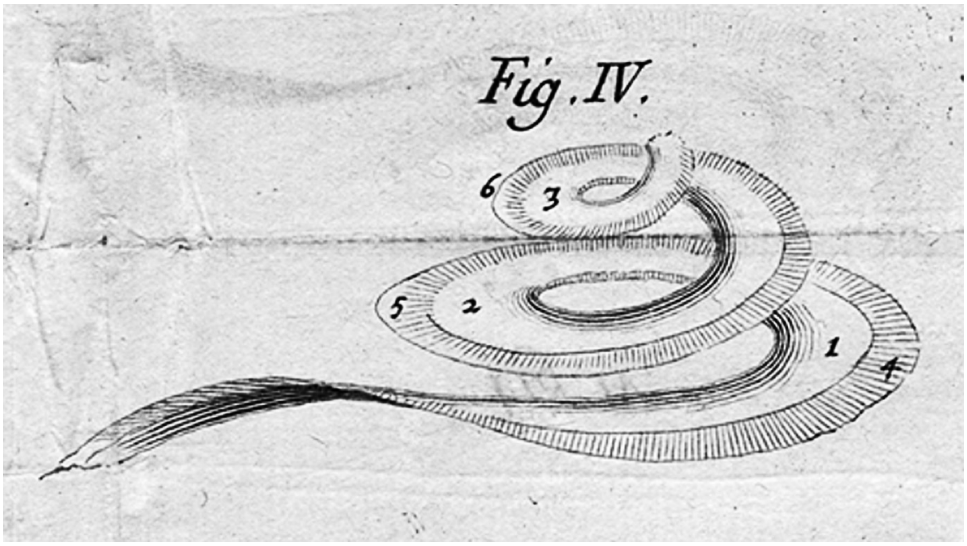


Figure 14.6: Duverney's 1683 drawing of the cochlea's spiral tuned structure. The inner lane (near the axis, his numbers 1–2–3) presumbly represents the bony shelf, the inflexible part of the cochlear partition that starts wide near the base and get narrower near the apex. The outer part (4–5–6), if it represents the basilar membrane, should start narrow and get wider, but he did not see it that way.

14.3 1D, 2D, and 3D Hydrodynamics

The analysis and modeling of cochlear hydrodynamics as electrical analogs, transmission lines, and such is a very mature endeavor, with many well-developed approaches, linear and nonlinear, of various dimensionalities. In this section, we'll take a cursory look at linear methods, emphasizing the analysis of uniform structures, leading to dispersion relations that we can use to design filters.

In the one-dimensional approach, pressure variations across the width and height of the scalae are ignored; fluid pressure and motion are functions of x only (that is, the dimension along the traveling wave's direction of motion, distance from the base along the BM), and fluid motion is in the direction of x only. Local volume changes due to fluid wave motion are accommodated by a distensibility of the scala cross-section area by a deflection of the BM. This 1D approach is accurate when the wavelength is long compared to the height and width of the scalae, so this approach is also known as the long-wave approximation.

The first such model (Wegel and Lane, 1924) was represented as an electrical analog with inductors as the series elements, to model the fluid mass, and resonators as the shunt elements—capacitors for the BM compliance in series with inductors for its mass.

Similar models were developed and analyzed over the intervening decades, usually with resistors also in the shunt impedance, to model energy absorption (Peterson and Bogert, 1950; Zwislocki, 1950; Caldwell et al., 1962; Zweig et al., 1976). It is typical in these 1D or long-wave models to use an electrical transmission line analogy, in the form illustrated in Figure 12.6.

For low enough frequencies relative to the characteristic frequency of a place (that is, far toward the base relative to the best place for a frequency), the fluid flow (or currents) are small and the pressures (voltages) are large; the membrane mass and loss are negligible and the membrane impedance is purely a compliance (that is, inductance and resistance in the shunt admittance, if any, have impedances that are negligible compared to that of the capacitance). In this region, the transmission-line model is essentially like the pure delay line of Figure 12.2. The long-wave approximation is accurate here, and gives a dispersion relation with no dispersion; that is, wavenumber k (radians per meter) is proportional to frequency ω (radians per second), and all frequencies propagate with the same velocity:

$$k^2 = K\omega^2$$

This dispersion relation derives from the partial differential equations describing the physical system (Lyon and Mead, 1988b); its relation between squares allows solutions propagating in both directions: $k = \sqrt{K}\omega$ and $k = -\sqrt{K}\omega$.

In order to get a big tuned response, the 1D models traditionally have a series-resonant circuit as the shunt admittance, with some membrane mass and loss, inherited from the Helmholtz concept of local resonators. But it has been shown (Shera and Zweig, 1991) that such a transmission line model leads to an unrealistic input impedance from the middle ear, in disagreement with experimental evidence, and would imply a low coupling efficiency there. The alternative is to have just a capacitor for shunt admittance, modeling a BM with stiffness but no significant mass. Direct measurements of wavenumber in the cochlea are best fit with a model in which membrane mass is small enough to have no significant effect except at the very highest frequencies, so membrane mass can be safely ignored except perhaps very near the cochlea's base (La Rochefoucauld and Olson, 2007). If there's also a small resistance associated with membrane motion, negative for low enough wavenumbers and positive for higher wavenumbers, it is still possible to get a resonance-like response, though not as sharp as with the traditional resonant-BM models.

Transmission-line models and analyses have mostly been passive, as opposed to having any active gain mechanism, though active *undamping* or negative resistance elements were proposed quite early (Gold, 1948). Such *active* models were later developed by many hearing researchers (Zwicker, 1979; Kim et al., 1980; Davis, 1983; Neely and Kim, 1983; Dallos, 1992). Negative damping (e.g. by negative resistance in a circuit

Early Cochlear Resonance and Wave Concepts

While the BM starts narrow and gets wider, the other part of the cochlear partition, the bony shelf (as well as the partition as a whole) starts wide and gets narrower. This bony structure misled the seventeenth-century French anatomist Joseph-Guichard Duverney (1683) to conclude that the cochlea was tuned to low frequencies near the base and high frequencies near the apex; see Figure 14.6. The eighteenth-century Italian Domenico Cotugno realized that the structure responsible for tuning was more likely to be the basilar membrane, and turned this around to the scheme that persists today; he also discovered that the cochlea was normally filled with fluid, not air as previously believed. Other eighteenth-century scientists who worked on the idea of frequency–place tuning include Valsalva, Boerhave, Zinn, Haller, and Goeffry (Shambaugh, 1910). In the nineteenth century, Hermann von Helmholtz tied up the local resonance theory with psychoacoustic and mathematical support.

Almost immediately, the Helmholtz concept of independent resonators, like the stretched strings of a harpsichord, came under attack from others who thought it seemed physically unlikely. It is a testament to Helmholtz's lucid description and analysis, and to his stature and authority, that the idea persisted as long as it did, and that even today it colors the thinking of many people about how the cochlea works. Alternative explanations of cochlear function had a hard time taking hold, with lots of early half-baked ideas, before Békésy observed the cochlear traveling wave in 1928. Even after that observation, there were continued difficulties, since models that fit Békésy's broadly-tuned wave observations could not explain sharp psychophysical and neural tuning.

Charles Herbert Hurst (1895) proposed a nonresonant traveling-wave theory, relying on coincidence of reflections to sort out different pitches. As it was described shortly thereafter (McKendrick, 1899; McKendrick and Gray, 1900),

> Hurst has suggested that with each movement inwards and outwards of the stapes, a peculiar wave is generated which travels up the scala vestibuli, through the helicotrema into the scala tympani, and down the basilar membrane to the fenestra rotunda. This wave is not a mere undulation of the basilar membrane, but it causes movements of fluid to and fro in each scala, and these produce a peculiar wave of pressure. As the one wave ascends while the other descends, a movement (or pressure) of the basilar membrane occurs at the point where they meet, and the movement is chiefly in the direction of the tectorial membrane, so that this membrane strikes suddenly on the hair cells and thus irritates the nerves. The point at which the waves meet will depend upon the pitch of the note, or, in other words, upon the time interval between the two waves. In this way, and without sympathetic resonance, the cochlea would, within limits, respond to tones of different pitch. The intensity of the movement of the tectorial membrane against the hair cells would, of course, correspond to intensity of tone.

This idea was a step toward wave theories, but not a realistic one.

Development of Cochlear Wave Concepts

Emile Kuile (1900) proposed an alternate nonresonant traveling-wave theory that, depending on frequency, sounds would set different lengths of BM in motion, with low frequencies affecting more length than high (Stewart, 1901; Fletcher, 1922).

Max Meyer (1907) published an account of the mechanics of the inner ear in which he rejected the local resonance hypothesis, based on his analysis of the properties of the basilar membrane. He argued that the BM was not under tension and thus would not behave elastically, and that any wave propagating by its displacement would have a wavelength long compared to the cochlea, such that the BM would move essentially as a whole. He appears to have not considered stiffness as alternative to tension as a way to get an elastic displacement; he treated the BM as having a nonlinear limit of displacement, such that larger portions would be displaced by louder sounds, but at all frequencies. He illustrated the antisymmetric motion of fluids in the scalae, as driven by the stapes, but didn't quite get to a wave response on the BM, so missed the opportunity to replace the resonance theory with a more physical wave theory.

About the same time, George Shambaugh (1910) and others developed a theory of the effective stimulus to the hair cells being a resonance of the overlying tectorial membrane (TM). Shambaugh felt that the TM was resonating "in response to the impulse of sound waves in the endolymph." His notion of a wave in the cochlea was a fast sound wave, like some others at that time. He supported the Helmholtz resonance theory while denying that the basilar membrane could be "a vibrating structure." Luciani (1917) supported this view in his eminent physiology textbook.

The move toward a more mathematical and physical model of traveling waves in the cochlea started with H. E. Roaf (1922), who wrote,

> Mass movement of the liquid can take place in one of two ways: Liquid may pass up the *scala vestibuli* through the *helicotrema* and down the *scala tympani*, or the *scala media* may be pushed towards the *scala tympani*. The resistance to these movements is in the former case the inertia of the mass of liquid to be moved, and the friction of the liquid against the walls of its containing tube, and in the latter case the tension of the basilar membrane (Reissner's membrane is usually represented as being flaccid).

This approach was further detailed using membrane stiffness (elasticity) instead of tension, and converted to an electrical analog, by Wegel and Lane (1924), and was given a good impetus when Békésy (1928) reported traveling waves that he observed on the BM.

Otto Ranke (1931) showed that Wegel and Lane's 1D or long-wave model would not be accurate at the point of maximum response, where the predicted wavelength was less than the duct height, and that a 2D model or a simplified short-wave model would work better.

Wever (1962) reviews the development of the traveling wave models of the cochlea, but never mentions the concepts of long and short waves, nor of linearity and nonlinearity. It wasn't until the later experimental observation of nonlinear active amplification (Kemp, 1979) that models of the cochlea began to be able to explain the subtle psychophysics of hearing. A range of historical overviews are available for the interested reader (Shambaugh, 1910; Luciani, 1917; Fletcher, 1922; Wever, 1949, 1962; Hawkins, 2001; Hachmeister, 2003).

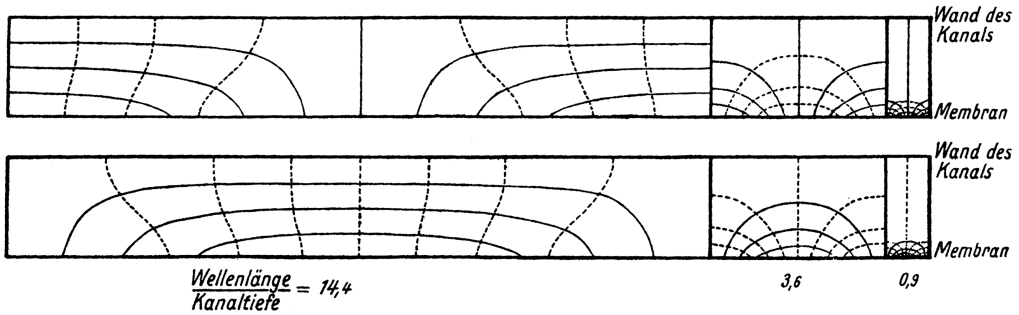


Abb. 2. Verteilung der Potentialströmung über die Kanaltiefe bei verschiedenen Wellenlängen.

Figure 14.7: Otto Ranke (1931) calculated these streamlines and iso-pressure lines for 2D waves in two narrow channels separated by an elastic membrane, at three wavelengths. He concludes, "Thus, while at long wavelengths almost all the pressure amplitude reaches the wall of the channel, at short wavelengths, the pressure at the channel wall remains nearly constant, and all the processes take place only in the immediate vicinity of the membrane." This is how the cochlear wave focuses sound energy into the vicinity of the organ of Corti. The left and middle conditions, with wavelength-to-channel-depth ratio (*Wellenlänge / Kanaltiefe*) of 14.4 and 3.6—corresponding to the wavenumber–height product $kh = 2\pi/14.4 = 0.44$ and $kh = 2\pi/3.6 = 1.75$—straddle the nominal $kh = 1$ boundary between long-wave and short-wave behavior.

model) with level saturation provides a good model of cochlear active gain at low levels, nonlinear growth of response at higher levels, and various two-tone suppression effects (Geisler, 1998), in qualitative agreement with observations in cochlear mechanics (Ruggero et al., 1992).

A general problem with one-dimensional models is that they predict wavelengths that get very short near resonance, putting them well outside their long-wave domain of accuracy. A two-dimensional approach can use a simple short-wave approximation instead, or can include 2D effects more accurately and can better model both the long-wave and short-wave regions, and the transition between.

14.4 Long Waves, Short Waves, and 2D Models

In a two-dimensional approach (Ranke, 1931, 1950; Siebert, 1974; Lighthill, 1981), the variation of fluid motion with distance y away from the BM is modeled, including components that move down the scala as in 1D as well as components that move with the BM, across the scala, requiring pressure gradients in both x and y. See Ranke's sketch in Figure 14.7, which shows the form of the pressure and fluid-flow waves on both the long- and short-wave sides of the transition region.

The characterization of such waves, in particular surface waves in water of finite depth, was worked out and published by Lord Rayleigh (Strutt, 1878), and by Sir Horace Lamb (1879, 1895). The wavelength as a function of frequency, the velocity, energy transport properties, wave patterns in a 2D slice of water, etc., were fully characterized in terms of hydromechanics and mathematics. The solutions included long waves, short waves, and the transition region between, just what we need for a complete 2D analysis of waves in a cochlea-like medium. It is a wonder that none of the early (before Ranke) proponents of traveling waves in the cochlea thought to apply that analogy, with the springiness of the basilar membrane substituted for the effect of gravity on the water surface. It appears that Ranke did not immediately apply Lamb's analysis to his model of the physics, but worked out an approximation instead.

In the case of the BM being modeled by just a compliance, like the capacitance we discuss in the LC delay

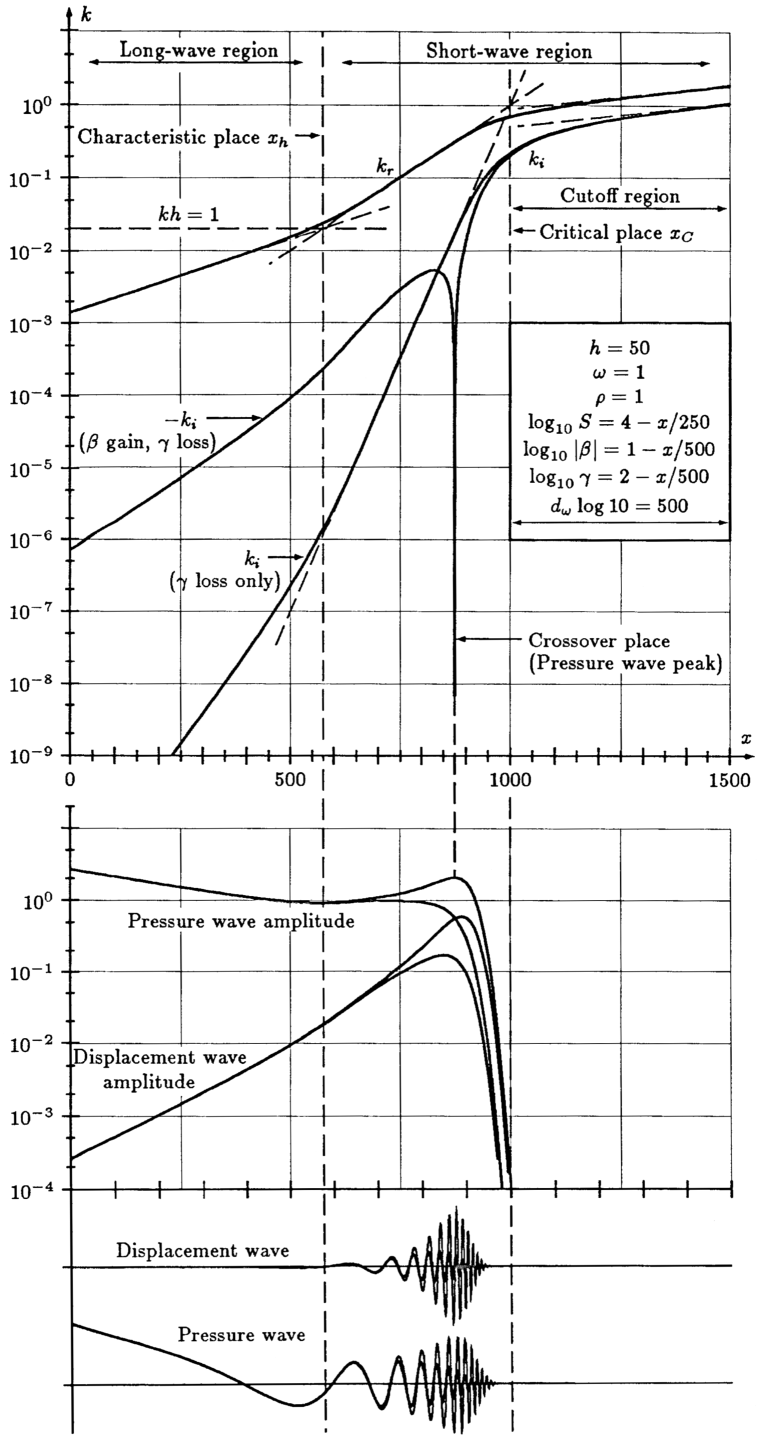


Figure 14.8: Wavenumber calculations from a 2D model as a function of place x, plotted as real and imaginary parts of k, with and without active gain, along with pressure and displacement waves with and without active gain; from Lyon and Mead (1988b). The difference between pressure waves and displacement waves is mostly in the base region, where the frequency is low compared to CF and the membrane is very stiff, so the energy propagates with relatively low displacement and high pressure. For a detailed view of active and passive cases, compare Figure 14.3, which is not based directly on a 2D model but on a filter cascade that gives a wavenumber that is comparable except in the cutoff region.

line, the 2D model's dispersion relation for the wavenumber at the BM is the same as Rayleigh and Lamb had for shallow-water waves:

$$k \tanh(kh) = K\omega^2$$

where h is the *height* of the scala (distance from BM to the rigid wall), for some parameter K that captures the other physical parameters such as fluid density and BM compliance (or gravity and water density in the case of water surface waves).

The hyperbolic tangent function is nearly equal to its argument when the argument is small, so for low frequencies and long wavelengths this relation is like that in the 1D or long-wave model, with k^2 proportional to ω^2, representing the region where fluid motion is along the axis of the scala and the transfer function is essentially a nondispersive delay. These long waves propagate like shallow-water gravity waves, where the water motion is mostly horizontal because the wavelengths are long compared to the depth.

At the other extreme, for high enough frequencies or wavenumbers, or short enough wavelengths, the tanh function saturates out at ± 1, and the dispersion relation approaches $|k|$ proportional to ω^2. This *short-wave region* is where the wavelength is shorter than the scala height h, and represents a fluid motion mode in which most of the motion is near the membrane, falling off exponentially like $\exp(-k|y|)$ with distance y away from the membrane. This behavior is the same as what are called deep-water waves, where the bottom of the water is so far away, compared to a wavelength, that nothing moves there. Lamb (1895) had already discussed these limiting regions of the tanh-based wave formulas, and their implications for wave velocity, along with the general case. Ranke's pictures of the wave patterns on both sides of the short-wave–long-wave divide are shown in Figure 14.7.

The dispersion relation shown above has real (lossless) solutions. When there are loss and/or active gain mechanisms in the system, they need to be incorporated into the dispersion relation to get the appropriate complex wavenumbers. Adding a viscosity loss term is straightforward. How to model the cochlea's active gain depends on the assumed microphysics, and is beyond the scope of the current discussion. What has become clear in recent years is that the active mechanism depends on a special *motor* protein: the *prestin* in the walls of the outer hair cells. Models for how the prestin-based electromotility of the outer hair cells couple to the traveling wave to add energy show good agreement with observations (Yoon, Steele, and Puria, 2011).

The solutions for k for a particular 2D model with hypothetical loss and gain functions (not based on any microphysics analysis), versus place x, for a fixed frequency, are shown in Figure 14.8. By comparing the pressure and displacement waves shown, it can be seen that at the basal end, the pressure wave propagates with slowly declining amplitude in the long-wave region, and nearly constant amplitude beyond that. A cascade of pole–zero filters with unity gain at DC and decreasing natural frequency will produce a similar wave pattern, versus the filter stage number.

The general solution for the fluid-flow pattern includes a hyperbolic-cosine y (depth from membrane) dependence, from two exponential terms constrained to make a zero flow velocity into the rigid wall across from the BM (or into the bottom below the water). Besides the complex 1D wavenumber at the BM, it is also possible to solve for a 2D wave vector in the fluid for these exponential-in-depth components. This wave vector shows not just the direction of wavefronts, but also the direction that energy is propagating. When there is an energy source in the BM boundary condition, the wavenumber of the larger component will point down the scala but slightly away from the BM, showing energy flowing from BM into the scala. Where the BM is lossy, the wave vector points the other way, showing the transport of energy into the BM, where it is dissipated—probably mostly in viscosity at a boundary layer while it is dragging the cilia of inner hair cells that detect the motion. The nature of the 2D wave solution is to *focus* the energy of the traveling wave into a smaller region near the basilar membrane as the wave slows down and the wavelength gets short, eventually delivering all the energy right into the basilar membrane itself, where it is detected by the inner hair cells of the organ of Corti.

In a three-dimensional approach, the extra width of the scala relative to the width of the BM has a signif-

icant effect in the transition region between the long-wave and short-wave regions. We don't have a simple dispersion relation formula for this type of model; rather, such models are generally solved numerically, for example as a way to fit the parameters of BM models to real physiological data, in the context of realistic dimensions (Steele and Taber, 1979; Lim and Steele, 2002). This approach is what led to the conclusion about BM mass being mostly negligible, except at very-high-CF places, mentioned above.

14.5 Active Micromechanics

The trick to making a good 2D or 3D model is to find a good model of the active and lossy behavior at the BM. There is plenty of experimental evidence that for frequencies somewhat below or near the characteristic frequency (CF) of the place under consideration, the cochlear partition is active, and adds energy to passing waves (Lukashkin et al., 2007). Conversely, for high enough frequencies, the partition is lossy and absorbs energy from waves. Turning this around, to the point of view of a wave of fixed frequency moving by places of decreasing CF, the wave is first amplified, up to some maximum amplitude near a place where the CF matches the wave frequency, and then is quickly attenuated beyond that place.

Although the active micromechanics of the hair cells and the surrounding structures in the organ of Corti is much studied and modeled, finding a good mechanical/mathematical model to explain the frequency or wavelength dependence of the loss or gain parameter has been slow in coming. It is easy to hypothesize a resistance-like term, giving a force proportional to fluid velocity at the membrane, corresponding to viscous loss at a surface layer. And it's easy to say that there's a corresponding negative "undamping" term; the trick is just to say how the negative loss is produced, and how it reduces with frequency. Recent studies of the micromechanics of the organ of Corti reveal directional asymmetry in the tilt of outer hair cells and phalangial processes that have been shown by analysis to provide the spatial phase shift needed to make the outer hair cells' activity provide an undamping effect (Yoon et al., 2007, 2011).

As modeling and measurement of cochlear waves progresses, we will have constantly improving data on the wavenumber as a function of frequency and place, which we can then use to constrain the design of our filterbank models.

14.6 Scaling Symmetry and the Cochlear Map

In some systems, the local properties of the medium—in particular the wavenumber $k(\omega)$—may be the same everywhere, except for a scaling of the frequency scale, such that there is a *prototype* or *mother* wavenumber shape k_{M} with

$$k(\omega) = k_{\mathrm{M}}(\omega/\omega_{\mathrm{r}}) \quad \text{for a place with local scale or resonant frequency } \omega_{\mathrm{r}}$$

If the scale changes exponentially with place x, then the overall transfer functions, starting from an infinitely distant base region at $x = -\infty$, will also have such a scaling symmetry. Such scaling is also known as *log scaling*, since the place coordinates map linearly to logarithms of frequency scaling factors.

In such systems, the transfer functions $H(\omega, x)$ to all places x can be described in terms of a mother transfer function of one variable:

$$H_{\mathrm{M}}(\omega/\omega_x) = H_{\mathrm{M}}(r) = A(r) \exp\left(-i \int_{-\infty}^{r} k_{\mathrm{M}}(u)\, du\right)$$

where $r = \omega/\omega_x$ is a normalized frequency parameter, ω_x is a reference frequency that depends on the place x as $\omega_x = \exp(x/d_\omega)$ for a characteristic length d_ω, $k_{\mathrm{M}}(u)$ is the mother function for $k(\omega, x)$, and $A(r)$ is the

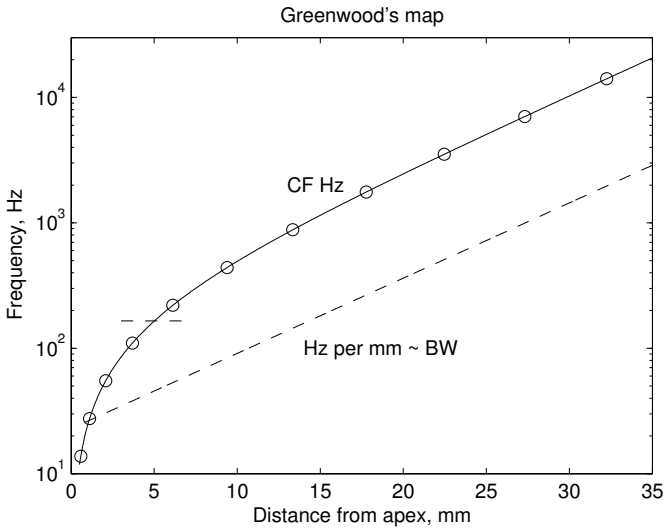


Figure 14.9: Greenwood's frequency–place map (Greenwood, 1990), showing the relation between places and their characteristic frequencies (CF). Points corresponding to frequencies of octaves (powers of 2 times A-440) are marked with circles. For most of the distance, the mapping is approximately geometric, or logarithmic. The dashed line shows the rate of change of frequency with place, in Hz per mm, which is proportional to the nominal bandwidth at each place. One way to get the Greenwood map is to integrate this exponential-in-place bandwidth, with distance from the apex, starting at zero center frequency but nonzero bandwidth.

amplitude correction factor that depends on conservation-of-energy considerations.

In real mammalian cochleas, this kind of scaling symmetry applies pretty accurately in the middle half of the cochlear place range, but there may be significant deviations near the base, due to the finite starting point and due to membrane mass that becomes relevant only at very high frequencies, and near the apex, due to the limited scala length and width that prevent continuing the log scaling indefinitely. Such a system is well modeled by a cascade of filters with geometric pole-frequency spacing, transitioning to linear spacing at low frequency, according to the Greenwood map or something like it; see Figure 14.9 and Figure 14.10.

14.7 Filter-Cascade Cochlear Models

We're now in a position to put together a cascade of filter stages to make a filterbank that models wave propagation in the cochlea. Each stage is like what Ren et al. (2011) call a "local transfer function," the transfer function from one point in the cochlea to another point further along in the direction of propagation.

In order for the stages to be implementable in circuits or in digital technology, they need to be expressed in terms of rational transfer functions, or poles and zeros. This cochlear modeling approach gives rise to the pole–zero filter cascade auditory filter models discussed in Chapter 13, and to a corresponding runnable dynamic digital version that we call the *cascade of asymmetric resonators with fast-acting compression* (CAR-FAC), discussed in Chapter 15.

The slightly more complicated *cascade–parallel* filter structure (Lyon, 1982) incorporated pairs of both poles and zeros as antiresonant notch filters in the filter cascade, motivated by the series-resonant circuits in the long-wave transmission-line model of Zweig, Lipes, and Pierce (1976). We later focused on cascades of simpler two-pole stages (Lyon and Mead, 1988a), motivated by an analysis of a 2D short-wave model with

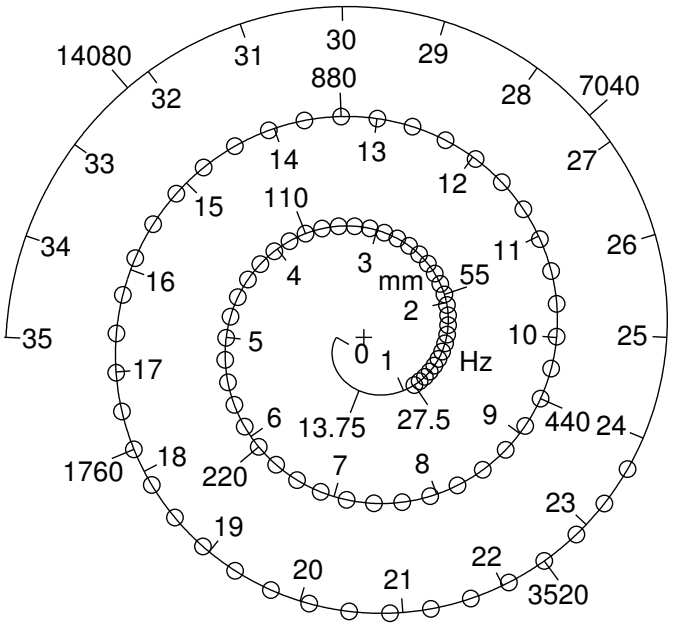


Figure 14.10: Greenwood's frequency–place map, illustrated on a spiral that approximates the shape of the human cochlea. Distances from the apex in mm are labeled inside the spiral, and frequencies of octaves on the outside. The fundamental frequencies, or pitches, of the notes of the 88 keys of a piano are marked by circles. Notice that geometrically spaced frequencies—octaves and notes—are about equally spaced, at nearly 5 mm per octave, in the basal and mid regions, but are bunched up near the apex, with only about 1 mm for the lowest octave of the piano. The human cochlea has about two and three-quarter turns; the final quarter turn shown in the center (the last 1 mm), which maps frequencies down to zero, should be interpreted as the helicotrema.

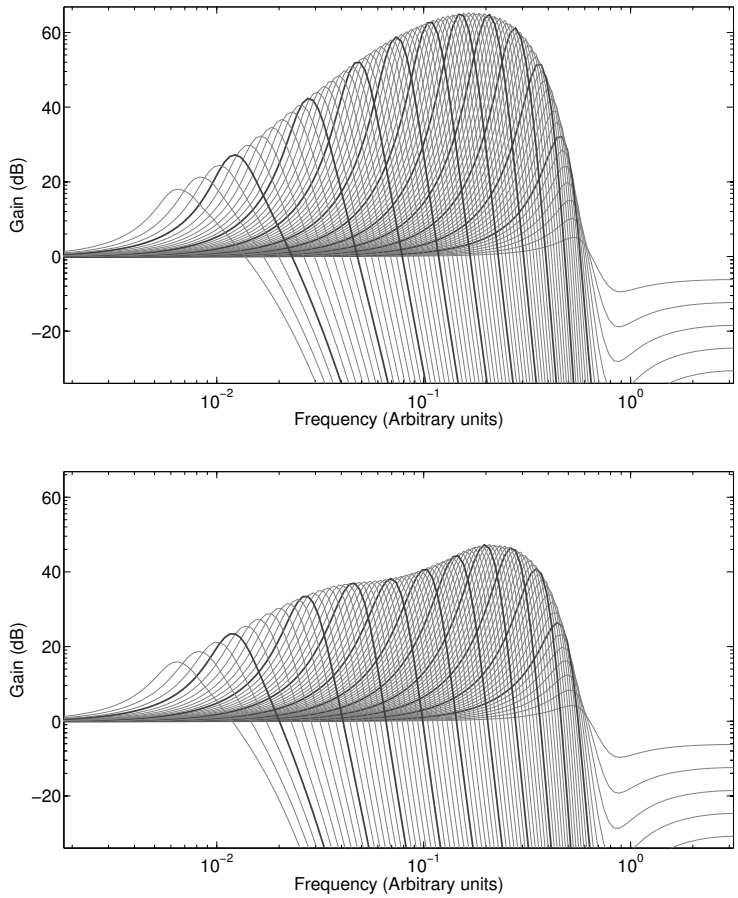


Figure 14.11: Adaptation of the overall filterbank response at each output tap, for the PZFC model of Lyon et al. (2010b). The upper plot shows the initial response of the filterbank before adaptation. The lower plot shows the response after adaptation to a human /a/ vowel of 0.6 sec duration. The plots show that the adaptation affects the peak gains (the upper envelope of the filter curves shown), while the tails, behaving linearly, remain fixed.

pseudoresonant behavior. But with these all-pole filter cascades (APFCs), it was hard to get a sharp enough high-side rolloff without excessive delay. Going back to the use of a zero pair at a frequency somewhat higher than the pole pair both gives a sharp cutoff and reduces the overall delay (Lyon, 1998). We call this a pole–zero filter cascade (PZFC), distinguishing it both from the APFC and from the early more complicated cascade–parallel pole–zero structure.

The cochlea has a substantial wave-propagation delay, on the order of several cycles of CF at each place. Filter cascades such as the APFC and the PZFC illustrate the fact that filters can exhibit a correspondingly substantial group delay, even though they are minimum-phase. This group delay is associated with the steep high-frequency rolloff of the gain response. The filter group delay is somewhat adjustable via the relative pole and zero positions.

Since we can get a cochlea-like response from individual stages as simple as second-order filters, each described by a complex-conjugate pair of poles and a complex-conjugate pair of zeros in the s plane, we limit ourselves to that level of complexity for the PZFC. If we later find better data from cochlear mechanics, we can revise the stage model, perhaps to higher order, as needed. Each pole is positioned slightly below the

corresponding zero in frequency, leading to a peak in gain near the pole frequency, followed by a sharp gain reduction at slightly higher frequencies, and then leveling off at a gain less than 1.

These filters will later be implemented digitally by mapping them to the z plane, as described in Chapter 16.

The filter cascades model the level dependence of auditory filters by having the poles and zeros move to positions of higher or lower damping; the resulting filter adaptation is illustrated in Figure 14.11. The initial low-damping small-signal positions of the poles and zeros are set for each stage, and the level-dependent nonlinearity is achieved by dynamically increasing the pole damping in each stage in response to the filterbank output. This modification of pole damping, or equivalently pole Q, corresponds to moving the poles either horizontally or along a circular trajectory in the s plane; we use horizontal motion, as shown in Figure 14.12. The peak frequency of the resonance shifts downward a little as the damping and bandwidth increase, but not as much as it would shift if we used motion along a constant-natural-frequency circle.

The initial pole frequencies are set to correspond to equal distances along a cochlear frequency–place map. Equivalently, each frequency is spaced from the next by a frequency difference proportional to the nominal local auditory filter bandwidth, for example as represented by the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990). For frequencies above about 200 Hz, the stage CFs form nearly a geometric sequence (equal frequency ratios), but for lower frequencies the sequence becomes more nearly arithmetic (equally spaced in frequency), potentially all the way down to nearly zero frequency. The zeros at each stage are placed at a frequency somewhat above the pole, say 40% higher. The resulting transfer function "bump" is a simple approximation to the gain bump in the cochlea due to active amplification and wavelength shortening before the strongly attenuating cutoff region.

Depending on the application, the pole frequencies or CFs may span a range of about 7 or 8 octaves, say from 25 Hz to 7 kHz, with usually 8 to 24 stages per octave of CF, for 56 to 192 total stages or channels. Each stage may have a maximum gain near 6 dB, relative to its low-frequency (or DC) gain, but with many cascaded stages with slightly varying CFs, the peak gain of the cascade may be 50 dB or more. Each stage will attenuate frequencies above its CF by just a few dB, but the cascade will yield a steep and deep cutoff of high frequencies.

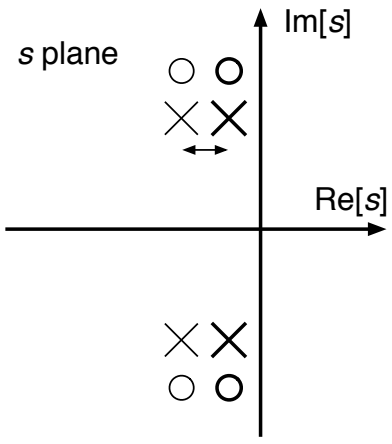


Figure 14.12: Diagram of the motion of the filter-stage poles and zeros in response to the CARFAC's gain-control parameter. The low-damping positions (heavy symbols) provide high gain near the pole frequency, compared to the high-damping positions (lighter symbols).

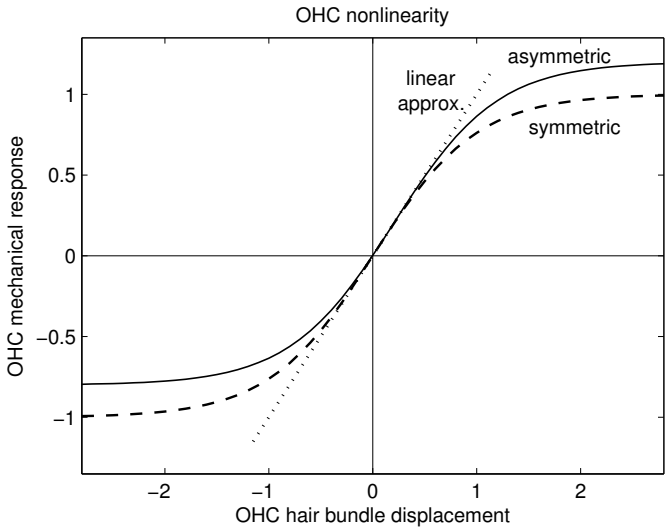


Figure 14.13: The transduction nonlinearity of the outer hair cells is a somewhat asymmetric *sigmoid* (solid), and is sometimes modeled as a symmetric sigmoid, such as a hyperbolic tangent (dashed). The slope of this curve is effectively a gain or active-undamping parameter, which is maximum near the rest (zero displacement) position.

14.8 Outer Hair Cells as Active Gain Elements

The outer hair cells are the magical elements that make hearing work. Damaging them by acoustic trauma or by ototoxic drugs such as the antibiotic *kanamycin* can be catastrophic to our auditory sensitivity. Yet what they do and how they do it are still far from being well understood.

What is clear is that when the outer hair cells are working, the mechanical response to a weak sound input, measured as basilar membrane displacement or velocity, is much greater than when they are not working. The net acoustic power delivered to the organ of Corti near the response maximum is much greater than the acoustic power that enters through the middle ear (Lukashkin et al., 2007). That is, the outer hair cells participate with the hydrodynamics in an active power amplification scheme, via a two-way transduction of motion to intracellular electrical signals and back. While many details of how they work have been uncovered in recent decades (Ashmore, 2008), we do not go into those details here—we just need to know that for small signals, the outer hair cells cause a sort of active undamping of wave motion, and that the energy available to do so is limited, so small signals are amplified but large signals are not.

The drug *carboplatin* selectively damages inner hair cells, rather than outer hair cells (Takeno et al., 1994), and also leads to hearing loss, but with a different pattern that suggests that the outer hair cells can still amplify traveling waves and generate otoacoustic emissions without the involvement of inner hair cells (Trautwein et al., 1996). The drug *furosemide* provides yet another insight into cochlear function. It reversibly reduces the endocochlear potential (EP), reducing the ability of both inner and outer hair cells to transduce signals. The result is a reduction in nerve sensitivity of 40 dB or more near frequencies where the neuron is most sensitive, presumably due to the failure of outer hair cells to amplify traveling waves, and also a reduction in sensitivity of about 10 to 15 dB at much lower frequencies, where active amplification has little role but the inner hair cells' sensitivity still matters (Sewell, 1984). Probing neural tuning curves, otoacoustic emissions, and other signals in reaction to such differing drug effects has been an important research method in clarifying the functions of different components of the cochlea.

The limited ability of the outer hair cells to provide a mechanical positive feedback proportional to BM

motion is typically modeled by some kind of a *sigmoid* nonlinearity, such a *logistic function* or *hyperbolic tangent* similar to that shown in Figure 14.13. For very small inputs, this function operates in a small-signal linear range, so the overall system looks like a linear system, with low distortion. For very large inputs, the function saturates and makes negligible contribution to the overall system response, so the system again approaches a linear limit with low distortion. Between these extremes, where the sigmoid function is operating substantially nonlinearly, and making a substantial contribution to the system gain, the output level will grow nonlinearly with the input level, and there will be a moderate amount of distortion as a result. Another source of nonlinearity, a gain-control loop that affects the outer hair cells' level of activity, helps to keep the system always away from the linear limits, so it is best to think of it as always being in that intermediate nonlinear regime.

14.9 Dispersion Relations from Mechanical Models and Experiments

The filterbanks that we can make by cascades of pole–zero stages correspond to dispersion relations in an underlying distributed wave propagation system that hopefully is a good model of the active cochlea. To see whether the dispersion relation corresponding to the filter cascade can be close to a good model, we need to compare it to the results from other methods. In particular, we need to look at dispersion relations derived from experimental data on the mechanical response of the cochlea, and at dispersion relations from cochlear models derived from hydromechanical modeling that includes the micromechanics of the active outer hair cells.

Unfortunately, good comparison data in the form of dispersion relations are relatively rare. Many models derived from physics are solved numerically without producing good formulae or curves for their dispersion relations. Most mechanical measurements are at single points, presented in a way that makes it hard to convert them to dispersion relations. But there are a few examples in the literature that we can compare to.

The dispersion relations from Lyon and Mead (1988b) that we illustrate in Figure 14.8 were derived from a 2D model with local activity, but the form of the positive and negative damping was hypothetical, based neither on micromechanical modeling nor on measured response data. It is useful for illustrating the concept of a locally active model, and for comparison with the filter-cascade model.

Zweig (1991) examines the mechanical data of Rhode (1971), in terms of λ^2, the squared reciprocal of the complex wavenumber. Like the other models, his fits show an active gain leading up to the point of the wave peak, followed by an energy absorbing region. But the fits are not regular enough to allow a more detailed comparison.

The data fits of Shera (2007) are not mechanical, but rather are based on first-order Wiener kernels of auditory nerve response at low levels in undamaged cochleas in chinchillas, and on similar signal analysis in cats (van der Heijden and Joris, 2003); in both cases they are thought to correlate well with the underlying mechanical response. Shera represents the dispersion relation by what he calls the *propagation function* and *gain function*, the real and imaginary parts of the complex wavenumber, respectively, as function of place and frequency, and concludes that

> ... at all locations examined, the gain functions reveal a region of positive power gain basal to the wave peak. The results establish the existence of traveling-wave amplification throughout the cochlea, including the apex.

Furthermore, Shera's smoothed data plots show that the gain bump before the wave peak is followed by a dip after the peak, resembling the gain function of our PZFC or CARFAC stage. His propagation function shows a maximum (a wavelength or wave-velocity minimum) right at the wave peak, in good agreement with the PZFC (see plots in Chapter 16). Shera's fitted average propagation constant has a maximum at about

Development of the Concept of AGC in Cochlear Mechanics

William Rhode (1971) observed a very nonlinear input–output relationship in cochlea mechanics, using his newly developed Mössbauer technique. In the same year, Rose et al. (1971) were among the first to suggest that observations on auditory nerve spike train patterns strongly suggested a mechanical "sensitivity control" in the cochlea:

> The capacity of a fiber to reflect the waveform of the stimulus when the latter greatly exceeds that sound pressure level which elicits a saturation discharge rate suggests the existence of a cochlear sensitivity control mechanism which may, but perhaps need not be, mechanical in nature. ... acceptance of nonlinearity drastically revises the classical, but nonetheless quite incredible, conclusion that at threshold the receptors are sensitive to displacements as small as a tiny fraction of the diameter of a hydrogen atom. It is also tempting to think that the receptors are not exposed to enormous variations in the amplitude of vibration as is postulated by the orthodox view. In fact, there is recent direct evidence [Rhode 1971] that the motion of the cochlear partition, in the region of maximal amplitude, is markedly nonlinear and therefore a very substantial error may be introduced in calculating the amplitude of the displacement at threshold by linear extrapolation of values observed at very high sound pressure levels.

By the end of the decade, modelers were taking note. Jont Allen (1979) made the case for AGC, in terms familiar to engineers, and began to connect it to efferent feedback:

> Given the opinions which we have so strongly expressed up to this point, the reader might reasonably ask what overall purpose cochlear nonlinearities serve. For those familiar with the data, one answer seems almost obvious: The nonlinear damping (as proposed in nonlinear cochlear models) acts to compress (attenuate) the frequency components of ... the neural excitation, near [CF] ... in order to increase the dynamic range of the filters. Thus the nonlinear damping acts as a mechanical automatic gain control.
>
> ...
>
> The outer hair cells are coupled to the efferent system, and COCB [crossed olivocochlear bundle] stimulation (stimulation of the outer hair cells through the efferent system) also gives rise to broadened tuning about CF in a manner very similar (as best we know) to the nonlinear level dependent mechanical damping. This experimental fact seems to be an important clue toward an understanding of the cochlear nonlinearity.

Allen (1981) continued to explain in the next of his sequence of papers on the state of cochlear modeling:

> A very significant feature of Rhode's data was that he found a compressive nonlinearity at frequencies neighboring the cutoff frequency. As a result, the output (BM displacement or velocity) varies much less than the stapes input displacement or velocity, for frequencies near the best frequency. The significance of this important finding will become clearer as we proceed, but, in my opinion, it is a precurser to an automatic gain control system which seems to be built into the cochlear filters. ... The automatic gain control nonlinearity also explains why the harmonic distortion is always below the primaries in intensity and does not grow large at large input levels as would be predicted from a power-law nonlinearity.
>
> ...
>
> It presently seems clear that this source of distortion is not the byproduct of some poorly engineered component. It is rather perhaps the negligible residual of a sophisticated local feedback mechanism in the mechanical motion of the properly operating cochlea, such as the automatic gain control system mentioned previously.

6 radians/mm, or a minimum wavelength of about 1 mm. With our PZFC model stages representing about 0.4 mm each (12 stages per octave of CF near the base), this corresponds to a maximum of about 150 degrees phase shift per stage, which is about twice what the PZFC stage provides. We have not attempted a more quantitative comparison or parameter fit, but it looks like our approach could be a better fit if we used twice as many stages.

The physical model of Liu and Neely (2009) uses a detailed model of the micromechanics of tilted outer hair cells to derive wave equations for a model with local active gain. They present an actual dispersion relation equation, which can be solved for k as a function of ω. But their model shows a wave peak where the wavelength is extremely short, about equal to the 0.07 mm tilt distance of the outer hair cells; this seems too short relative to most data—reported minimum-wavelength observations in cochlear mechanics are in the range 0.4 to 0.8 mm (Wilson, 1973, 1992; Ren et al., 2011). Too-short wavelengths correspond to rather long delays and too many wave cycles on the BM. Our PZFC approach would need more than 100 stages per octave to get enough delay to approach Liu's dispersion relation. Nevertheless, it is great that the model exists in this form, so that it is possible to do parameter adjustments and comparisons to see how different modeling approaches relate. Most other models in the literature with local active gain are not in a form that we can easily compare.

14.10 Inner Hair Cells as Detectors

So far, we've discussed how to construct a filterbank, to make many bandpass-filtered versions of the input sound, by cascading filters that model local wave propagation; see Chapter 16 for more details on the filters. To complete the forward signal-processing path of the cochlea model, we need to model the *detection* process at the inner hair cells, an adaptive rectifying nonlinearity whose output represents what the auditory nerve will send to the brain.

Though we sometimes use a simple (ideal) half-wave rectifier at the output of the filters to make an approximate *neural activity pattern*, that is not a very good model of what the hair cells do. The inner hair cell is adaptive, and reduces its response gain quickly after a signal onset. There are various good models for this behavior, which we examine in Chapter 18.

14.11 Adaptation to Sound via Efferent Control

The cochlea sends signals to the brain via the inner hair cells and the auditory nerve, as discussed in subsequent chapters; such signals from the periphery into the brain are known as *afferent* signals, and the neurons as afferent neurons. But the brain also sends *efferent* signals, via efferent neurons, out to the periphery. Something like 5% of auditory-nerve neurons are efferents, and most of these terminate on the outer hair cells, completing the control loop mentioned in Section 14.8. See Figure 14.14 and Figure 14.15, for relevant anatomical details.

Through many kinds of experiments, it is known that activity on the efferents (specifically, the medial olivo-cochlear efferents, or MOC neurons) causes the outer hair cells to provide less active gain to cochlear waves (Kim, 1984; Darrow et al., 2006; Guinan, 2010). That is, the brain can tell the cochlea to be less sensitive, or less responsive, to sound, by reducing the hydromechanical cochlear amplifier gain and thereby reducing the effective stimulus that is the input to the inner hair cells.

Another inhibitory effect, mediated by the lateral olivo-cochlear efferents, or LOC neurons, occurs where they make inhibitory synapses with the primary afferent auditory neurons, near where the inner hair cells make excitatory synapses (Kim, 1984; Guinan, 2010). This effect shows up in auditory-nerve response, but

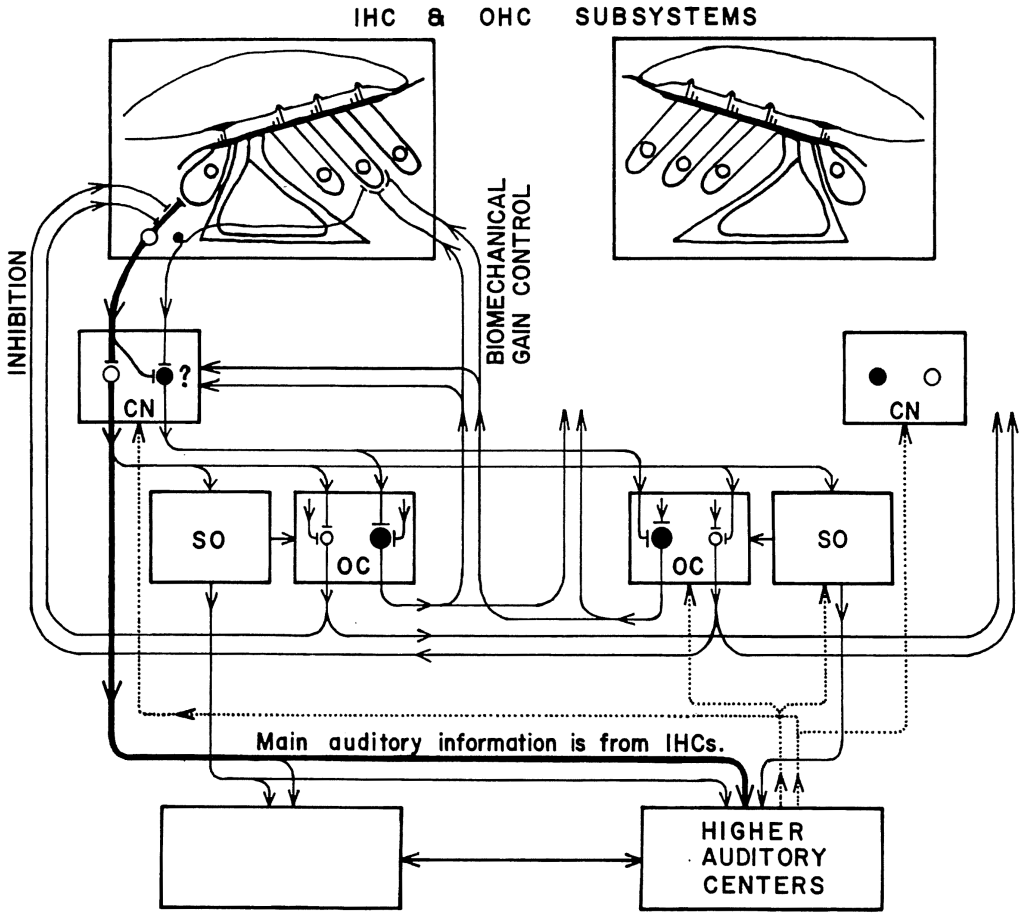


Figure 14.14: Duck Kim (1984) created this "block diagram for the hypothesized IHC and OHC subsystems in the cochlea and the brainstem up to the superior olivary complex and their connections to the remainder of the auditory system." The superior olivary complex (SO) drives the olivocochlear neurons (OC) that provide feedback from the brain to the cochlea, both to control the biomechanical gain and to inhibit the response of the primary auditory neurons that send auditory information from the inner hair cells (IHCs) to the cochlear nucleus (CN). Much of the feedback is crossing between left and right via the crossed olivocochlear bundle (COCB, not labeled), which is a convenient location for injecting electrical signals to directly control the cochlea's gains, both mechanical and neural. Filled circles represent neurons in the outer hair cell (OHC) subsystem. The hypothesized CN neuron with the question mark has since been identified in the marginal shell of the anteroventral cochlear nucleus (AVCN) (Ye et al., 2000). [Figure 7.3 of (Kim, 1984) reproduced by permission of Duck On Kim.]

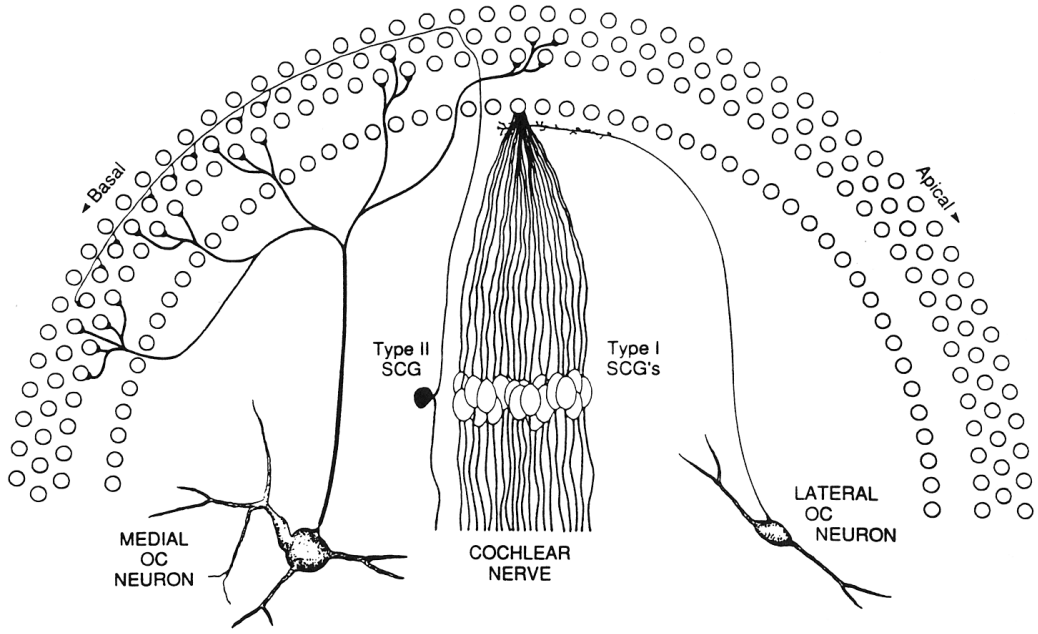


Figure 14.15: Bruce Warr's 1992 "hypothetical isofrequency unit of afferent and efferent innervation from the middle of the cochlea" shows the collection of different neurons types sharing a common CF, and how they relate to cochlear place (Warr, 1992). The small circles represent the one row of inner and three rows of outer hair cells in the spiral organ of Corti. For a given CF, the efferent feedback neurons from the medial olivary complex (OC) control outer hair cells over about a half-octave range of places toward the base from the place that drives the cochlear nerve afferents (Type I spiral ganglion cells), so that they can modulate the outer hair cell activity in amplifying traveling waves that are coming from the base (from the left in this drawing). [Figure 7.12 (Warr, 1992) reproduced with permission of Springer.]

not in mechanical response. For both types of efferent control, the default state of the system in quiet is to have its gain at a maximum; in response to sound, the efferents are activated and the gain is reduced.

The engineering model of this kind of system is *automatic gain control* (AGC), as described in Chapter 11 and elaborated for the cochlear model in Chapter 19. Though the system is nonlinear, the *forward* system can still be analyzed as linear when the parameters are fixed, as for a steady signal level; and the *feedback* system can be modeled with a linearized control loop that has the forward system's parameter dependence as part of it. The control loop might have time constants that are long compared to those of the forward system, making it easy to analyze. Or it might not, as we see adaptive effects in sensory systems on all time scales.

The auditory efferent fibers are more broadly tuned than afferent primary auditory nerve fibers (Kaiser and Manley, 1994), probably due to aggregation of afferent signals across a range of places. Efferents innervate outer hair cells across a range of places, mostly within a range of a few mm basal of the place innervated by afferents of comparable CFs (Warr, 1992). Any outer hair cell affected by an efferent is involved in the amplification of traveling waves to a range of more apical places, so that is one more mechanism of spreading of the gain-control effect across places. For all of these reasons, strong signals can cause the reduction of gain to signals at other frequencies, both higher and lower, affecting observed patterns of masking and suppression.

The efferents also respond to sound that enters either ear. As a result, a sound in one ear can affect the mechanics of the opposite ear! This mechanism has been called a "binaural gain control" (Brugge, 1992). Both of these, cross-frequency and between-ear AGC, are examples of *coupling*, or *coupled AGC* (Lyon, 1984): the gain at one place is reduced by feedback coupled in from other places, including the other ear.

The idea of AGC was applied to adaptation in the visual system by Albert Rose (1948) (not to be confused with Jerzy Rose who worked in the auditory system), and by others more recently. Cochlear AGC, or *biomechanical gain control*, has been explored since at least the 1970s (Rose et al., 1971; Allen, 1979; Evans, 1980; Allen, 1981; Lyon, 1982; Kim, 1984; Kick and Simmons, 1984; Lyon, 1984; Geisler and Greenberg, 1986; Weintraub, 1987; Lyon, 1990; Patuzzi, 1996; Zwislocki et al., 1997; Zhang et al., 2003; van der Heijden, 2005; Recio-Spinoso et al., 2009).

14.12 Summary and Further Reading

The filterbank-like functionality of the cochlea can be efficiently emulated by a cascade of simple filters. This structure follows from traveling-wave models, and has been used, in combination with automatic-gain-control concepts, to motivate digital filter implementations for machine hearing front ends since at least 1982 (Lyon, 1982). Incorporation of appropriate nonlinear mechanisms is straightforward—though the corresponding anatomy and physiology are complicated—and makes the resulting nonlinear filterbank more realistic.

For readers interested in more physical and mathematical depth, Reichenbach and Hudspeth (2014) have recently provided an excellent extended description of active cochlear mechanics.

The ideas of cochlear hydromechanical traveling waves and an AGC-like compressive nonlinearity are well established, though the micromechanical details remain unclear. The idea of outer hair cells adding energy to the traveling wave is widely, but not universally, accepted, and there are good reasons to remain uncommitted to this idea until it is more firmly proved. For example, Allen and Fahey (1992) and de Boer et al. (2005) discuss opposite viewpoints on the same set of experiments designed to measure the cochlear amplifier; and van der Heijden and Versteegh (2015) present a coherent technical refutation of the idea of active amplification. Fortunately, filter cascades as models of local dispersion relations should be effective no matter how this question is resolved.

Chapter 15

The CARFAC Digital Cochlear Model

> The modified transmission-line implementation, like the low-pass filter version, is an active system, with adjustments to the filter Q values changing the filter shapes and gains. ... This functional variation of Q with level gives a nearly uniform 2.5:1 compression ratio in the cochlear output for inputs ranging from 0 to 100 dB SPL.
>
> — "Accurate Tuning Curves In a Cochlear Model," James Kates (1993a)

The multiple-output *cascade of asymmetric resonators* with *fast-acting compression* (CARFAC) model of the auditory periphery combines concepts from many of the previous chapters, toward the primary goal of providing an efficient sound analyzer to support machine hearing applications.

An important secondary goal of CARFAC is to connect closely enough with known auditory physiology and psychophysics that it can be used to visualize and explain many interesting auditory phenomena. It is not a goal to be physically accurate, well calibrated, or in agreement with every detail of peripheral auditory processing, though it can be used as a starting point for those who have such goals.

Other than my own recent work (Lyon et al., 2010b; Lyon, 2010; Lyon et al., 2010a; Lyon, 2011b,a), the closest digital cochlear models in the literature to CARFAC are the variable-Q cascade–parallel models described by Kates (1993a), which also used cascades of asymmetric resonators specified by their poles and zeros; see the opening chapter quote above. The difference is that his stages include a second filter at each tap, and use a rather different pole–zero pattern, motivated by matching iso-rate neural tuning curves; ours is motivated by matching both psychophysical filters and physiological impulse responses, as discussed in Chapter 13.

15.1 Putting the Pieces Together

The CARFAC cochlear model pulls together the bits of knowledge that we surveyed in the previous nine chapters.

The coupled-form asymmetric resonators from Chapter 8 are combined into a filterbank with gammatone-like response as described in Chapter 9, using the cascade architecture motivated by Chapter 12. The linear system theory from Chapter 6, with its discrete-time version from Chapter 7, and especially the description of linear systems in terms of poles and zeros, lets us understand how the parameters of these cascades of simple filters determine the transfer functions that define the filterbank. Some of the nonlinearities that we learned about in Chapter 10 make the model more realistic than any linear filterbank can be; in particular, the use of the automatic-gain-control concepts of Chapter 11 make CARFAC dynamically level dependent. We previously analyzed a level-dependent but otherwise linear auditory filter model in Chapter 13, to show how varying a

filter's gain via the damping factors of its poles leads to a realistic quasi-linear model of auditory filtering—a good link between the linear filter theory and the nonlinear CARFAC model of the cochlea. In Chapter 14, we looked more carefully at how the physics of the cochlea relates to the traveling wave dispersion relations that our filter cascade emulates via the techniques of Chapter 12.

Putting these pieces together, CARFAC becomes a nonlinear cascade filterbank with realistic dynamic level-dependent filtering, realistic nonlinear distortion products, and realistic nonlinear detection producing a neural activity pattern output for further analysis in the auditory nervous system. And CARFAC is supported by open-source code that does all this with good computational efficiency.

15.2 The CARFAC Framework

The CARFAC's filtering is based on, but differs somewhat from, the Laplace-domain PZFC filter descriptions that we have analyzed in Chapter 13. The Laplace-domain PZFC description is a model of what a cochlear filter channel does, and is useful for many kinds of modeling. The PZFC auditory filter is based on a filter cascade, but is a model of one point in a cochlear-place continuum that we need to discretize, and is a statically level-dependent linear model. It needs to be made into a multiple-output filterbank, and made dynamically nonlinear (adaptively compressive) at all outputs to support general sound analysis by digital computing machines. The connection between the auditory filter approach and the machine hearing front end is straightforward, based on standard digital filter implementation methods and the cascade structure. But this connection has some minor complications and compromises that need to be faced, for example involving the choice of sample rates relative to the sound frequency range that we want our machines to hear, and the particular nonlinearities that are needed.

The CARFAC structure as developed in this chapter is a true multi-output runnable cascade filterbank. Its structure is not complicated, but has important details that need to be specified, among the filtering, the inner and outer hair-cell models, and the coupled automatic gain control (AGC) network that implements much of the compression part of the model.

15.3 Physiological Elements

The functional elements of the cochlear and auditory nervous system have been delineated in different ways by different investigators. The diagram of Dallos (1992), adapted in Figure 15.1, is probably closest to the structure of our CARFAC model (Dallos didn't label the loops, and labeled the outer hair cell (OHC) motor as "DC motor" in one version of the figure and "AC motor" in another, in contrasting different aspects of OHC action).

The asymmetric resonators in the cascade of asymmetric resonators (CAR) that models waves on the membranes are two-pole–two-zero filters resembling filter D of Section 8.2. Their bump-followed-by-dip frequency responses are a reasonable approximation to the cochlea's distributed wave dispersion relations with gain followed by loss. The height of the gain bump is set by damping of the resonator stage poles—low pole damping corresponds to high gain. A high stage gain corresponds to active amplification of the traveling wave, for frequencies in the gain bump region. In the real cochlea, the amplification is provided by OHCs, and the amount of gain is modulated by efferent connections. The digital model includes an OHC model that determines the amplification, or peak stage gain, via each stage's pole damping parameter; and it includes a feedback loop that controls how hard the OHC pushes toward lower damping, as part of an automatic gain control (AGC) loop. See Figure 15.2 for how these elements, and others, are interconnected. The BM (basilar membrane) response outputs shown there are useful in systems that reconstruct sound, such as hearing aids, while the NAP (neural activity pattern) outputs represent what the brain gets from the ear via the auditory

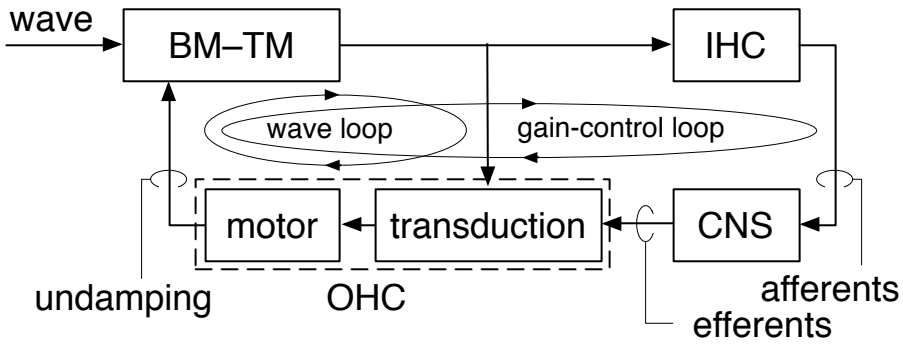


Figure 15.1: This diagram, adapted from Dallos (1992) and elaborated, shows the functional physiological elements of one location in the cochlea, which can be seen as a pair of feedback loops. The short loop, defining the hydrodynamic wave filtering system, involves the basilar and tectorial membranes (BM–TM) and active feedback from the outer hair cells (OHC), working at audio frequencies. The longer and slower loop, the afferent/efferent loop from the inner hair cells (IHC) through the brainstem of the auditory central nervous system (CNS) and back, controls the activity level of the OHCs, automatically adapting the system to the sound level. The instantaneous wave and the slower efferent feedback interact in the OHC, the nonlinear element whose "motor" action provides gain via active undamping of the wave mechanics.

nerve.

Figure 15.3 shows the approximate compressive input–output curve that CARFAC is intended to model; it is similar to an input–output curve shown by Kates (1993a). Studies in chinchillas have shown compressive-region slopes as low as 0.2, and more than 60 dB of gain at low levels, compared to postmortem (Ruggero et al., 1997).

Having fairly explicit relationships between physiological elements and model components is useful, but not crucial to our main goal. To some extent, we abstract multiple physiological elements into a single model element. The smoothing filters of the AGC loop, for example, model effects of the efferent system, but also model faster adaptive effects that are likely local to the organ of Corti itself.

15.4 Analog and Bidirectional Models

Our CARFAC model is designed for digital implementation in software or hardware. It propagates waves in only the "forward" direction. There are other good peripheral models that differ in these two points. Analog models were popular in the 1980s and 1990s, before Moore's law advanced the efficiency of digital computing by orders of magnitude, but are no longer as attractive compared to digital. Analog bidirectional VLSI models have been well explored (Watts et al., 1991; Hamilton et al., 2008; Wen and Boahen, 2012), but not much applied to practical problems. Bidirectional models are still important to those who want to investigate and understand otoacoustic emissions, but for machine hearing tasks only the forward waves are relevant.

The bidirectional analog model of Eberhard Zwicker (1986) is interesting mainly for how it incorporated explicit active undamping by feedback from a nonlinear OHC, like ours, including efferent feedback control of the degree of activity; see Figure 15.4. Using structures known as *wave digital filters*, digital bidirectional models have incorporated some of the same concepts, whether in linear approximation (Strube, 1985), or nonlinear (Friedman, 1990; Giguère and Woodland, 1993; Baumgarte, 1999). Some, as in the models of Zwicker and Peisl (1990) and Baumgarte (1999), incorporate smoothing across nearby channels in the efferent

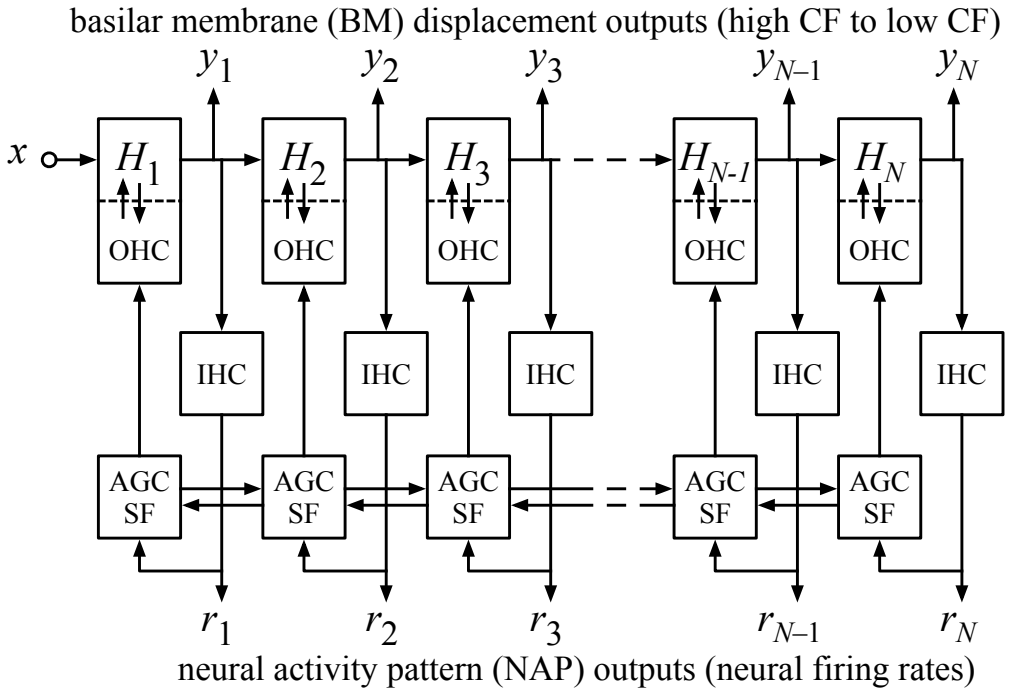


Figure 15.2: The *Cascade of Asymmetric Resonators* consists of the not-quite-linear transfer functions H_1 through H_N that model BM motion based on the cascade structure of Figure 12.9. *Fast-Acting Compression* is implemented by the other elements, including the OHC model that is tightly integrated with the filter stages and gives them their nonlinearity, and the coupled AGC smoothing filters (AGC SF) that modulate how the OHCs control the parameters of the filters. Between these main parts is a detection nonlinearity, such as an IHC model, which can have some compression and adaptive state of its own. The lateral interconnections of the smoothing filters allow a diffusion-like smoothing, or coupling, across both space and time. Outputs from the CARFAC include BM motion y_i (a set of compressed nearly-linear filterbank outputs) and an estimate of average instantaneous auditory nerve firing rate r_i, the NAP.

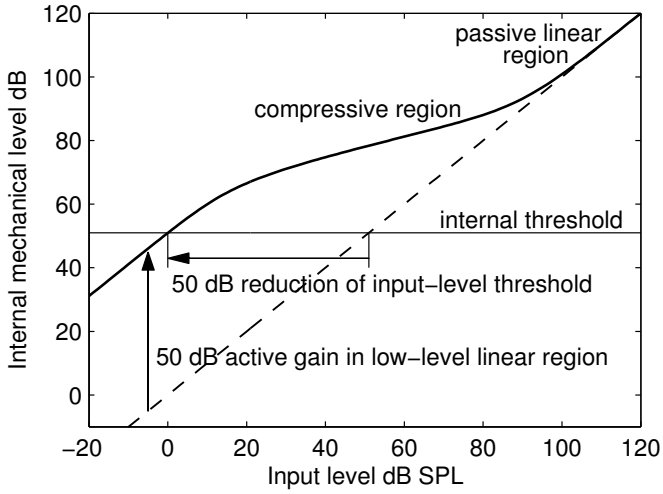


Figure 15.3: The compressive input–output curve exhibited by cochlear mechanics and emulated by the CARFAC model (solid) is compared with the passive linear or "dead" cochlear response (dashed), to show how extra gain at low levels reduces the input level needed to reach a threshold level of mechanical response. Here the mechanical threshold is chosen to correspond to 0 dB SPL with 50 dB of gain at low level. The curve is representative of the middle of the place or CF range, as opposed to very basal and apical regions that exhibit less active gain and less compression.

feedback, as we do.

There have been few other digital cascade models; Ambikairajah, Black, and Linggard (1989) presented a cascade of 3-pole–2-zero filter stages. There have been more digital cascade–parallel models—models that include a parallel bank of resonators or second filters connected to the taps of a cascade (Lyon, 1982; Zwicker and Peisl, 1990; Kates, 1991, 1993a). Among analog VLSI cochlear models, on the other hand, the cascade structures dominate (Sarpeshkar, 2000), probably because my first analog cochleas (Lyon and Mead, 1988a) were cascades.

15.5 Open-Source Software

The CARFAC model introduced here and detailed in subsequent chapters is implemented in open-source code, using simple low-level functions for the different parts (CAR/OHC, IHC, AGC) illustrated in Figure 15.2. The intentions are that the code should implement exactly what this book describes (though it may evolve slightly over time) and that the code should be an easily wrapped library, not a part of a particular environment. Functionally identical Matlab and C++ versions are supported (find CARFAC on GitHub, though it may move).

15.6 Detailing the CARFAC

In this short chapter, we introduced the structure of the CARFAC digital-filter model of peripheral auditory function, which we designed to be an efficient sound processor. The key pieces of the model are the CAR, the OHC model that dynamically controls the damping of the CAR, the IHC model that produces nonlinearly detected outputs from the CAR, and the AGC loop that adjusts the OHCs based on what the IHCs are doing. The overall CARFAC model is shown in Figure 15.2. All of these parts contribute to the fast-acting

compression—the FAC.

The CAR, OHC, IHC, and AGC are the subjects of the next four chapters.

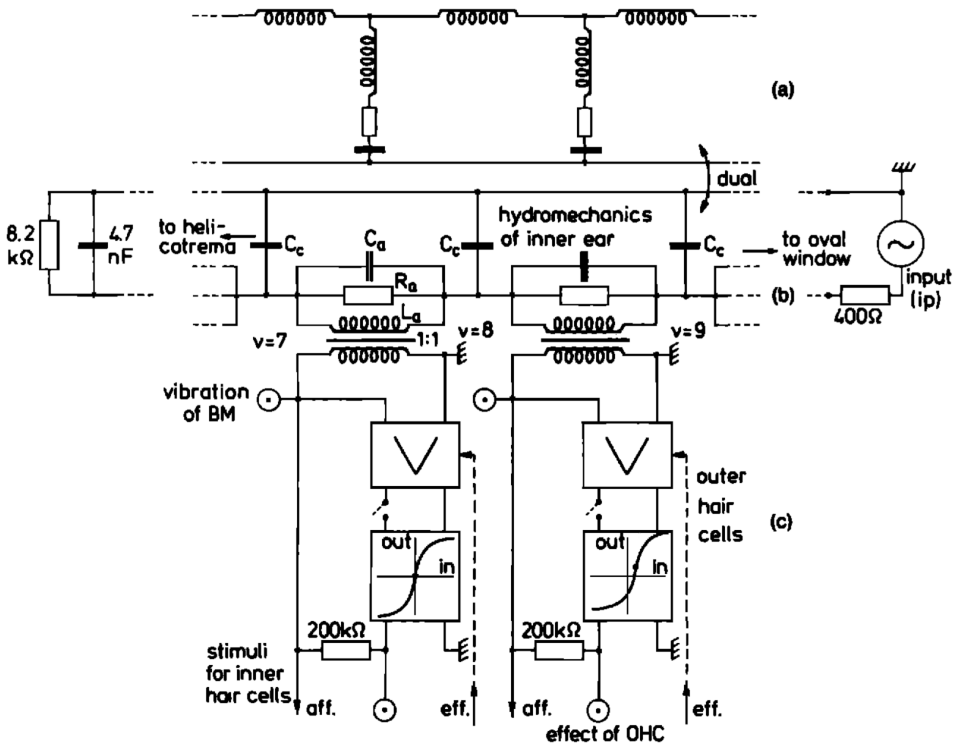


Figure 15.4: The analog bidirectional transmission-line model of Zwicker (1986), with saturating OHC nonlinearity and efferent feedback control, foreshadows the digital CARFAC functionality. Note the efferent ("eff.") control of the OHCs. [Figure 1 (Zwicker and Peisl, 1990) reproduced with permission of AIP Publishing.]

Chapter 16

The Cascade of Asymmetric Resonators

> Up to the threshold of the nervous system, the general outline of the process of frequency analysis is fairly clear. There is little room for doubting that the first main step of the process is essentially a matter of filtering. True, when one encounters electrical filters with one input and a number of outputs they are likely to consist of parallel selective networks fed from a common source. However, cascaded networks with taps at their junctions are not unfamiliar, and they provide a fairly exact analogue, insofar as general lay-out is concerned, of the cochlear analyser. The input is at the basal end of the cochlear partition, and the taps are the receptors or nerve terminals disposed along the length of the partition.
>
> — "Auditory frequency analysis," J. C. R. Licklider (1956)

16.1 The Linear Cochlear Model

The CARFAC model starts with a *cascade of asymmetric resonators* (CAR): linear two-pole–two-zero filters. We just need a recipe for how many sections to use, and how to set their coefficients to get a reasonable match to a linearized model of the human cochlea, at a range of levels.

The pole frequencies are chosen to correspond to equal spacing along the place dimension of the cochlea, by using the Greenwood map discussed in Chapter 14. The zeros are spaced a fraction of an octave above the pole frequencies.

Since we plan to extend this linear filter to incorporate nonlinearity through movement of the poles and zeros, we start with a filter form optimized to support that. A direct-form two-pole–two-zero filter stage with the input connected as shown in Figure 8.19 has a DC gain that does not vary as the coefficients are varied to move the poles. That implementation, referred to as PZFC in previous reports (Lyon et al., 2010b; Lyon, 2010), has been used in several applications of our cochlear modeling concept. But since the zeros don't move with the poles, it does not satisfy the physiologically observed condition that the impulse response zero crossings are very nearly unchanged with variation in level (Carney et al., 1999). For that, we need to move the zeros, too. The coupled form in Figure 16.1 makes it easy to move the zeros the same amount as the poles, which gives nearly the condition that we want. This configuration still allows a small level dependence of zero crossings (about 1/16 cycle per 10 dB, as shown below), since it doesn't move the zeros quite as much as the PZFC2 model of Section 13.8 does.

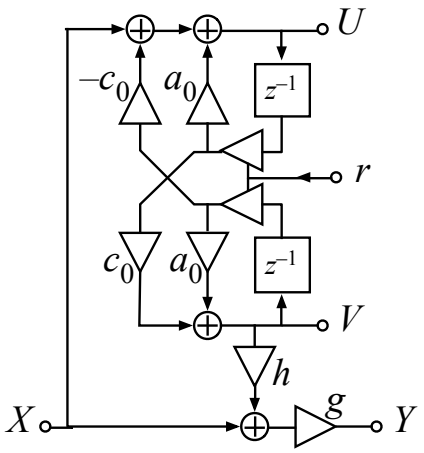


Figure 16.1: A pair of zeros is added to the coupled form by mixing the input with the filtered V output, as in filter D of Chapter 8. The resulting filter has zeros at the same radius in the z plane as the poles, which is a good place for them and gives the coordinated motion to keep the zero crossings of the impulse response from moving too much. The h coefficient controls the ratio of the zero frequency to the pole frequency, and the g coefficient is used to adjust the overall gain. In the arrangement shown, a factorization of coefficients to include an explicit pole radius parameter r is used to enable dynamic control of the damping. The pole radius is related to damping factor ζ by $r = \exp(-\zeta\omega_N T) = \exp(-\gamma T)$, as in Section 8.2 where the $\gamma = \zeta\omega_N$ is the negative real part of the s-plane pole position, mapped to the z plane via $z = \exp(sT)$. The a_0 and c_0 parameters, the cosine and sine of the pole angle, respectively, represent the pole positions $z = a_0 \pm ic_0$ in the zero-damping ($r = 1$) case.

16.2 Coupled-Form Filter Realization

The two-pole–two-zero filter stage can be realized in a conventional direct form, or a variety of other forms. The *coupled form* has the good property that it is easy to control the coefficients in such a way as to move the poles and zeros together, varying the filter's decay time constant while leaving the zero-crossing times nearly unchanged. And unlike some other configurations, the coupled form is "parametrically well behaved" and stable as its poles are dynamically moved (Mathews and Smith, 2003; Massie, 2012).

The coupled form is easily interpreted as a first-order filter with a complex state variable and a single complex pole. Calling the input X and the complex output $W = U + iV$, the transfer functions of the network in Figure 8.20 are found from W as described in Section 8.7:

$$\frac{W}{X} = \frac{z}{z - (a + ic)}$$

$$\frac{U}{X} = \frac{1}{2}\left(\frac{W(z)}{X(z)} + \frac{W^*(z^*)}{X^*(z^*)}\right) = \frac{z(z-a)}{z^2 - 2az + (a^2 + c^2)}$$

$$\frac{V}{X} = \frac{1}{2i}\left(\frac{W(z)}{X(z)} - \frac{W^*(z^*)}{X^*(z^*)}\right) = \frac{zc}{z^2 - 2az + (a^2 + c^2)}$$

When a pair of zeros is added to the transfer function of a two-pole resonator in the Laplace domain, making a transfer function of order two in both numerator and denominator as in filter D of Section 8.6, the high-frequency asymptote of the Bode plot will be flat (see Figure 8.14)—as opposed to the 12 dB/octave falloff of V/X (with its constant numerator and second-order denominator) or the 6 dB/octave falloff of U/X (with first-order, or one-zero, numerator). We can mix some of the input X with one or both of the outputs U and V to produce a flat high-frequency asymptote—which results in a pair of zeros. In the z domain, there is no high-frequency asymptote, but a pair of complex zeros has a corresponding effect, making a dip followed by a flat region in the frequency response.

Mixing the input with the resonator output V, with its one zero at $z = 0$, gives zeros at the same radius in the z plane as the poles, as explained in connection with Figure 8.21; taking the output one sample earlier, making the pole part of the filter be minimum phase with a z^2 factor in the numerator, or taking the output an extra sample later, removing the z from the numerator, would not put the zeros in a place analogous to those in filter D. Since this analogy to filter D is what we want, we settle on that special case, avoiding any potential extra coefficients that would be needed to get more explicit control of the zero positions. The resulting flow diagram is shown in Figure 16.1; its transfer function is:

$$\begin{aligned}\frac{Y}{X} &= g\left[1 + \frac{hcz}{z^2 - 2az + (a^2 + c^2)}\right] \\ &= g\left[\frac{z^2 + (-2a + hc)z + (a^2 + c^2)}{z^2 - 2az + (a^2 + c^2)}\right] \\ &= g\left[\frac{z^2 + (-2r\cos\theta_R + hr\sin\theta_R)z + r^2}{z^2 - 2r\cos\theta_R z + r^2}\right]\end{aligned}$$

in which we use θ_R as the normalized pole ringing frequency in radians per sample, or pole angle in the z plane; and r as the pole and zero radius in the z plane. Pole–zero constellations for a number of frequencies and dampings are illustrated in Figure 16.2.

For convenience, we will use the parameters $a_0 = \cos\theta_R = a/r$ and $c_0 = \sin\theta_R = c/r$, the values that the

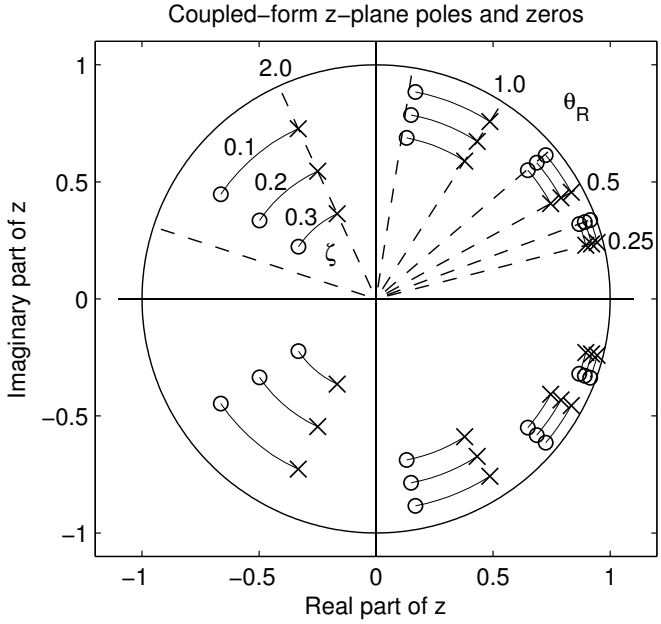


Figure 16.2: Pole–zero plot for the filter stage of Figure 16.1, illustrated for pole frequencies octave-spaced at $\theta_R = 0.25$, 0.5, 1.0, and 2.0 radians per sample, and damping factors $\zeta = 0.1$, 0.2, and 0.3, for the case $h = \sin\theta_R$; zeros are connected to their corresponding poles by solid thin arcs. This h value puts the zeros about a half octave above the poles, except at the highest pole frequencies, as shown by comparison with the radials (dashed) shown at $\sqrt{2}$ ratios (at the higher pole frequencies, the zeros squash closer to the poles, so they miss the dashed lines). Varying the a and c coefficients proportional to $r = \exp(-\zeta\omega_N T)$, or approximating that by $r = 1 - \zeta\omega_N T = 1 - \gamma T$ as we do here, moves the poles and zeros exactly along radial lines. We refer to γ/ω_R as the damping, even though it is not exactly.

a and c coefficients would have in the undamped ($r = 1$) case. Then we can change the damping, and move the poles and zeros together, by changing only r. Using these parameters, the transfer function is:

$$H(z) = \frac{Y}{X} = g\left[\frac{z^2 + (-2a_0 + hc_0)rz + r^2}{z^2 - 2a_0 rz + r^2}\right]$$

From the constant and quadratic coefficients in the numerator and denominator (r^2 and 1), it is apparent that, as long as the poles and zeros are complex, the zeros will be at the same radius r as the poles. The transfer function can be factored to make explicit the pole and zero positions at radius r and pole ringing angle θ_R and zero angle θ_Z:

$$H(z) = g\left[\frac{(z - z_{\text{zero}})(z - z^*_{\text{zero}})}{(z - z_{\text{pole}})(z - z^*_{\text{pole}})}\right]$$

$$z_{\text{pole}} = r\cos\theta_R + ir\sin\theta_R \quad \text{with } \cos\theta_R = a_0$$

$$z_{\text{zero}} = r\cos\theta_Z + ir\sin\theta_Z \quad \text{with } \cos\theta_Z = a_0 - hc_0/2$$

The condition for complex zeros becomes relevant for high-frequency channels, where $\cos\theta_R < 0$:

$$a_0 - hc_0/2 > -1$$

$$h < \frac{2 + 2a_0}{c_0}$$

For CARFAC, we use $h = c_0$, in which case the frequencies of the zeros will be about a half octave above the pole frequencies, as shown in Figure 16.2. For lower h, the zeros will be closer to the poles, making the response peak more asymmetric; conversely, higher h moves the zeros far away, leaving the response more like the response of the poles alone.

16.2.1 Stage DC Gain

The DC gain of the stage is only weakly dependent on r in the range we care about:

$$H_{DC} = g\left[1 + \frac{hc_0 r}{1 - 2a_0 r + r^2}\right]$$

To get unity gain at DC, we can solve for g:

$$g = \frac{1 - 2a_0 r + r^2}{1 - (2a_0 - hc_0)r + r^2}$$

To set g to get unity gain only in the undamped ($r = 1$) case, we can use the simpler formula:

$$g_0 = \frac{2 - 2a_0}{2 - 2a_0 + hc_0}$$

which is about 0.5 at low pole frequencies, where $2-2a_0$ is about θ_R^2, and hc_0 is also about θ_R^2 if $h = c_0 = \sin\theta_R$. At higher θ_R, g_0 will be closer to 1. If we fix g at this value g_0, and then increase the damping from zero, the DC gain will drop to be slightly less than 1, as shown in Figure 16.3. For realistic linear low-frequency tails, we might want to keep the DC gain from dropping as damping increases, so a more accurate g value could be used. Since many stages are cascaded, it doesn't take much change of DC gain per stage to have a significant

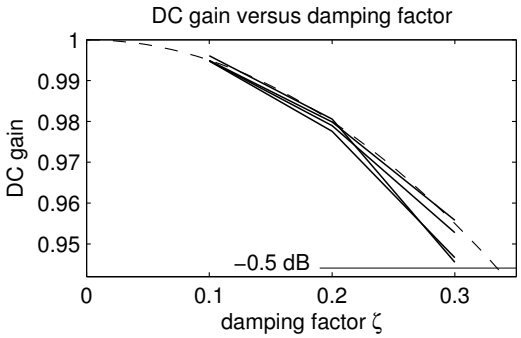


Figure 16.3: DC gains of the filter stages with pole and zero locations shown in Figure 16.2, when the gain coefficient g is fixed at the value that gives unity gain for the undamped case. The approximation $1 - \zeta^2/2$ (dashed) is most accurate at low θ_R. The thin line near the bottom indicates a loss of one-half dB.

effect on the gains of low-CF output channels of the filter cascade.

In practice in a cochlear model, the damping will not exceed 0.3 (where the gain bump goes away, roughly) except for very loud inputs, so the accumulated gain reduction in a cascade of about 100 stages can be about 30 dB or so (as the input level changes over a very wide range of levels), which may be acceptable, though it may give a more-than-realistic suppression of low frequencies by high frequencies. The g coefficient can be modified based on the computed damping value to keep the damping of high-CF channels from affecting the gains of low-CF channels, either by an exact calculation or by an approximate extra gain correction factor of $1 + \zeta^2/2$. Since it is unclear whether the tail gain variation would be a real problem for a machine hearing system, we do vary g with r to keep the CARFAC model more accurate.

16.2.2 Stage Response and Cascade Response

The gain and phase responses of typical two-pole–two-zero filter stages are shown in Figure 16.4 and Figure 16.5, for damping factors 0.1, 0.2, and 0.3.

The cascaded stages combine to provide a family of filters from the one input to the output taps between the stages. The resulting filters have peak gains that can be quite high, and can vary over a wide range depending on the stage damping parameters, as shown in Figure 16.6.

The cascade responses can also be characterized by the zero-crossing times and instantaneous frequencies of their impulse responses, as shown in Figure 16.7. Typical impulse responses at different dampings are shown in Figure 16.8, where the modest variation of zero-crossing times is apparent.

The total group delay of the cascade to its various outputs is shown in Figure 16.9, normalized to cycles of CF. At low damping, the delay can be up to 5 cycles.

The effects of varying the damping resemble the effects of suppression in cochlear mechanics, on gain, phase, and delay. The large reduction of group delay near CF, with increasing damping, is qualitatively consistent with the experimental findings of a *pivot of phase* by Versteegh and van der Heijden (2013), who note that for a variety of suppressors,

> . . . these three ways of changing the stimulus intensity had virtually the same nonlinear effects over a wide range of probe frequencies. For these three datasets, the largest nonlinear amplitude changes were found for probes just above CF. The corresponding phase changes were systematic but complex. Phase pivoted around a frequency near CF as intensity increased from 0 to 50 dB SPL. Below the pivoting frequency, phase acquired a lag; above the pivoting frequency, a lead.

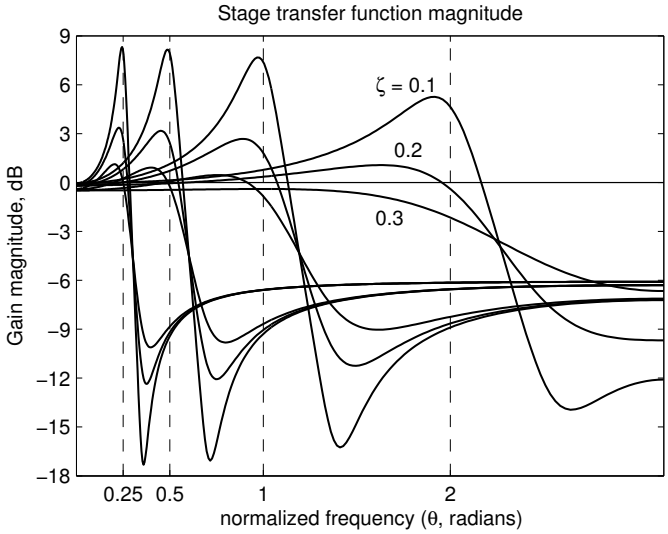


Figure 16.4: The CAR stage frequency-response gains for the four pole frequencies and three damping factors illustrated in Figure 16.2. For these plots, g is fixed, allowing the DC gain to deviate a bit from unity as damping increases. The deep "notch" behavior in the lower half of the plot leads to a very steep high-frequency slope in the cascade response; the fact that the gain comes back up some after the notch has little effect on the cascade filter shape, since the cascade gain is essentially zero in that region.

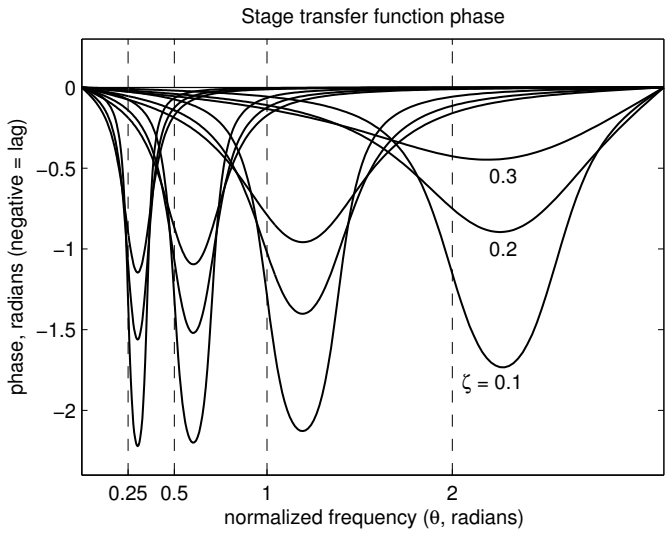


Figure 16.5: The phase responses of the stages with parameters illustrated in previous figures. As the damping changes, the phase stays approximately constant at a frequency just below the pole frequency, but goes through CF with a variable slope, indicating a variable group delay. Beyond CF, where the response is getting small, the phase lag is moving back toward zero, so the group delay there is negative.

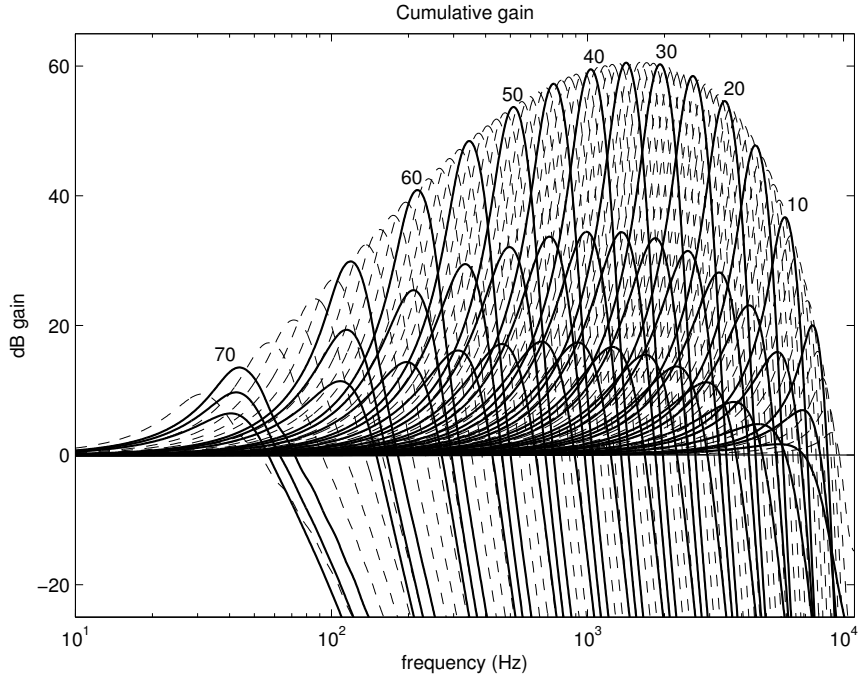


Figure 16.6: The cumulative frequency response (Bode plot) of a cascade of 71 pole–zero CAR stages, with 12 stages per octave at the high-frequency end. Every fifth output tap (or channel) is shown with heavy solid curves, for the same three damping factors as before; at the middle damping, all channels are plotted, with light dashed lines. The pole frequencies range from about 9900 Hz (2.818 radians per sample) down to about 30 Hz, based on equal spacing on a Greenwood map and a 22050 Hz sample rate. Peak locations of responses at the lowest damping define the characteristic frequency (CF) values used in subsequent plots.

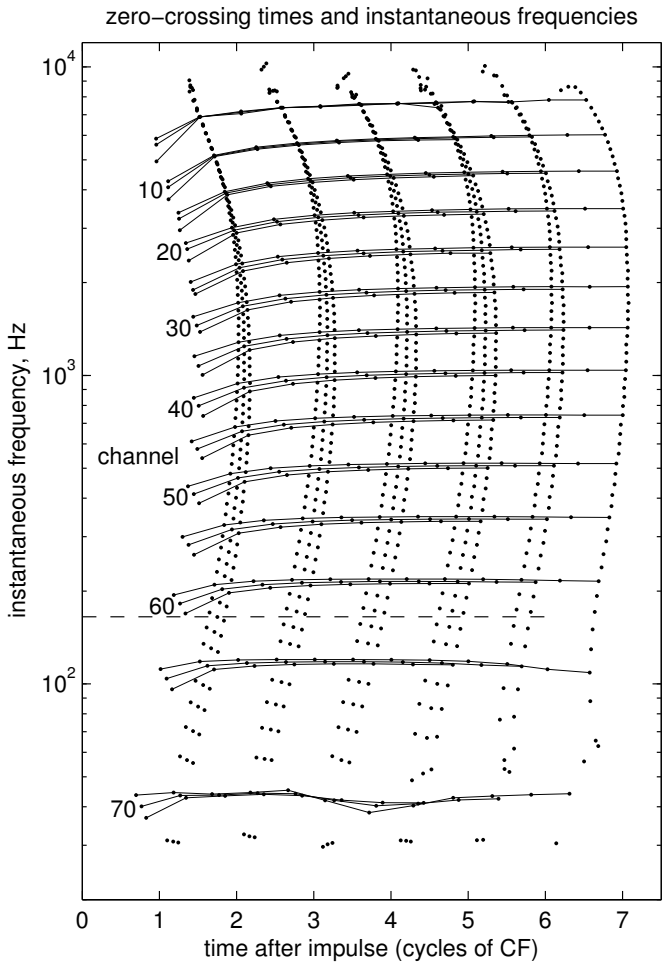


Figure 16.7: The instantaneous frequencies of the impulse responses of the 71 CAR channels, at the three damping levels, as a function of normalized time (cycles of CF after the impulse). Dots mark positive-going zero crossings of every channel, and negative-going zero crossings of every fifth channel. Fewer zero crossings are plotted for impulse responses with higher damping, since they decay sooner. Instantaneous frequencies are estimated near each zero crossing via Hilbert transforms of the impulse responses. Upward glides of about 20% are apparent for higher-CF channels. Lower-CF channels show less upward glide, but not much of the downward glide reported in the auditory nerve (Carney et al., 1999). The dashed horizontal line marks the break frequency in the Greenwood frequency map (see Figure 14.9), below which the channel spacing approaches linear instead of geometric. The zero-crossing times are seen to move through less than 1/4 cycle as the system adapts its gains through about a 40 dB range.

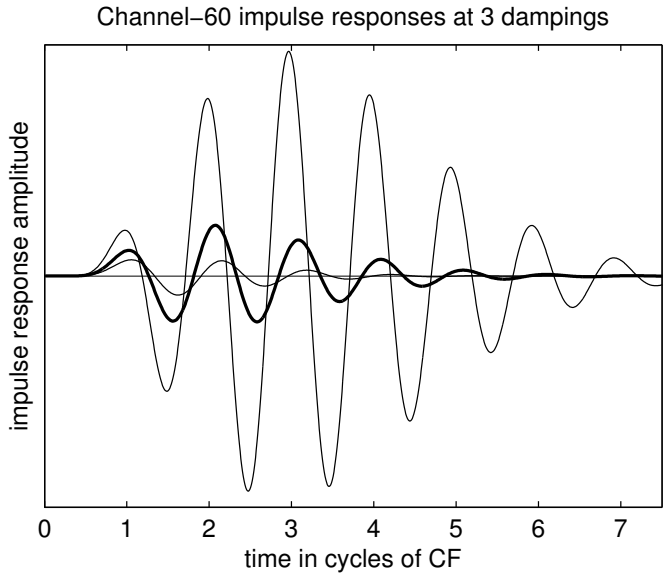


Figure 16.8: The impulse responses of channel 60 of the 71-channel linear CAR model at the three different dampings. The not-quite-aligned zero crossings are apparent. The smaller impulse responses correspond to higher dampings, as would be used at higher levels. The domain spans 7.5 cycles of CF, so the zero crossings align with those plotted in the previous figure. The group delays range from about 2 cycles at high damping (high level) to about 3.5 cycles at low damping (low level); see next figure.

Thus we see that the CAR is a linear system when its parameters (pole radii or dampings) are held constant, but in the context of fast-acting feedback that controls the damping in the stages, as discussed in subsequent chapters, it is a nonlinear system. This close integration of linear and nonlinear behaviors helps to make the cascade act as a useful cochlear model.

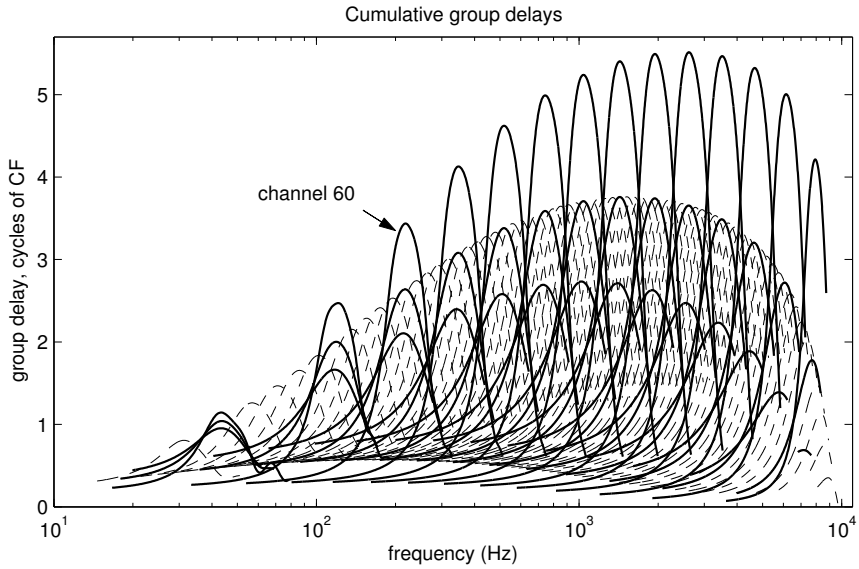


Figure 16.9: Group delays of the linear filter cascade, in units of cycles of the channel's CF, for damping factors 0.1, 0.2, and 0.3, plotted as in Figure 16.6. Channel 60, whose impulse response was plotted in the previous figure, is pointed out; it has a CF near 220 Hz. Delays peak near CF, as can be seen by comparison with Figure 16.6. To reduce clutter, the curves are cut off where the cascade gain is below 1 dB on the low-frequency side or below -3 dB on the high-frequency side. Near their peaks, the filters have about a cycle of delay per 10 dB of gain. The largest absolute time delay, near the apex, or low-frequency, end of the cochlea, is about 20 ms—one cycle of 50 Hz or two cycles of 100 Hz. As in Figure 16.6, every fifth channel is shown, except at the middle damping, where other channels are shown dashed.

Chapter 17

The Outer Hair Cell

> The CA (cochlear amplifier) model explains the detection of small differences in time as well as in frequency, the dual character of the electrocochleogram, recruitment of loudness in cochlear hearing impairment, the long latency of normal neural responses near threshold, acoustic emissions (both stimulated and spontaneous) and the locus of TTS (temporary threshold shift) in the frequency range above the exposure tone. Both the classical high-intensity system and the active low-level CA system are highly nonlinear and they combine to compress the great dynamic range of hearing into a much narrower range of mechanical movement of the cilia of the inner hair cells.
>
> — "An active process in cochlear mechanics," Hallowell Davis (1983)

To model the active and compressive amplifying wave propagation in the cochlea, the CARFAC stage incorporates dynamic nonlinearity or *fast acting compression* (FAC) through a structure we call the *digital outer hair cell* (DOHC), shown in Figure 17.1. This structure varies the positions of the poles and zeros, changing their radius r in the z plane by changing r in the coefficient formulas. The r value is increased from its passive value, where it would be for very loud sounds, to actively reduce damping for weaker sounds. Alternatively, we may say that the r value is decreased in response to sound, from its maximum value in quiet, down to a minimum or passive value.

Compression of a wide input dynamic range, by controlling the amplification of low-level sounds, is the primary function of this component, which is unique to the mammalian auditory system. The distortion that the OHC generates as an inevitable by-product of varying the gain is useful as a diagnostic, both for humans and for machine models. The distortion products are sometimes audible, contributing to perceptual differences between sounds.

The cell-membrane protein channel that makes outer hair cells work as active "motors" is known as *prestin*. Species with good high-frequency echolocation capabilities—bats and dolphins—exhibit an interesting convergent evolution of their prestin genes, presumably in support of good high-frequency amplification of feeble echos (Li et al., 2010).

17.1 Multiple Effects in One Mechanism

The DOHC in the CARFAC is the mechanism that integrates the instantaneous and parametric nonlinearities into one damping-control mechanism. The amount of the increase in r above the passive value is controlled by an instantaneous nonlinear function (NLF) of the filter's local state, scaled by feedback from a multi-time-constant AGC loop. The system is arranged to have a low-damping (high-gain) small-signal linear limit and a high-damping (low-gain) large-signal linear limit, with a realistic pattern of compression and distortion in the big range between these limits, where typical sound levels are.

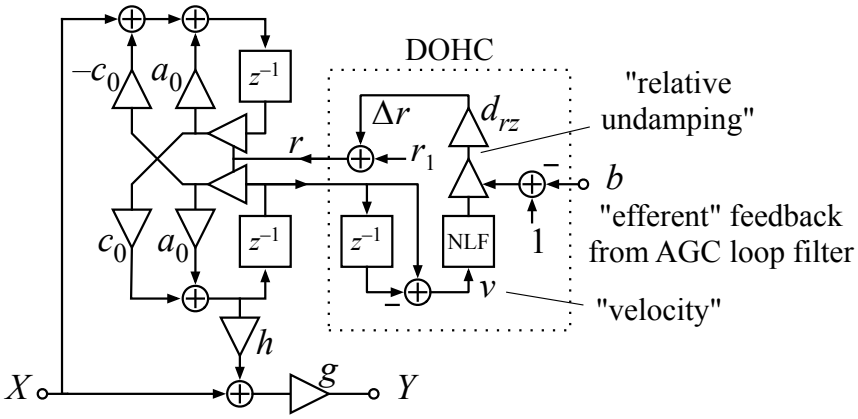


Figure 17.1: The linear filter stage of Figure 16.1 is here extended to incorporate nonlinearity via dynamic variation of the pole and zero radii (r) through functions localized into a digital outer hair cell (DOHC) block. The block computes a velocity (difference across a one-sample delay), then computes a damping (or a *relative undamping*, really), and finally computes and applies a corresponding r coefficient, incorporating both a local instantaneous nonlinearity based on the velocity, as well as "efferent" feedback from an AGC loop filter.

The compression due to the DOHC is *fast acting* in the sense that it incorporates not just a multi-time-scale dynamic gain-control (level adaptation) loop but also an instantaneous (suppression) effect—both via the same variable-damping mechanism. The instantaneous part, and perhaps also some of the faster feedback part, of the damping variation interacts with the propagating wave signal to make distortion products, or combination tones, that also propagate down the cascade and cause responses in the regions tuned to them. All of this is done with a modest number of arithmetic operations per sample per stage, and no per-sample evaluations of roots or transcendental functions.

In Figure 17.1, the maximum damping, for the high-level or passive limit, is set by the minimum-radius parameter, r_1, which depends on the stage CF. The "velocity" signal is the rate of change of an internal state variable of the coupled-form filter. The coefficient d_{rz} controls the rate at which the *relative undamping* affects the pole radius r; d_{rz} is set to about 70% of $1 - r_1$, to allow 70% of the damping to be cancelled. Thus the variable Δr ranges from 0 in the high-level passive case (b near 1, or high instantaneous v making low NLF output) to about $0.7(1 - r_1)$ times the relative undamping at low levels (b near 0); the relative undamping is a value between 0 and 1 since both the NLF output and the efferent feedback b are between 0 and 1. A 70% reduction of damping allows the stage gain to increase by about a factor of $1/(1 - 0.7)$, or about 10 dB; these parameters can be adjusted to give more or less compressive gain change.

17.2 The Nonlinear Function

A nonlinear function of velocity or displacement that increases damping is the typical source of cubic distortion tones (CDT) and compressive amplitude behavior in nonlinear models of the cochlea. For example, a damping increment proportional to the square of a velocity was used in each of the ten cascaded filter stages of the nonlinear system of Kim, Molnar, and Pfeiffer (1973); we looked at the low-level linear limit of this system in Figure 9.12, and discussed its equations as a distributed bandpass nonlinearity in Section 10.4. Each of their ten cascaded resonator stages is described by a nonlinear differential equation very similar to

one presented in a slightly simpler notation by Johannesma (1980):

$$\ddot{y} + (b_0 + b_2 y^2)\dot{y} + \omega_0^2 y = x$$

where b_0 is the low-level damping and b_2 controls the amount by which the square of the resonator's output increases the damping. This is the equation of a Van der Pol resonator (for $b_0 > 0$, which makes it stable) or a Van der Pol oscillator (for $b_0 < 0$, which gives it a periodic limit cycle) (van der Pol, 1926). The boundary between the stable and oscillatory regions, where the small-signal damping is zero, is known as a Hopf bifurcation (or Poincaré–Andronov–Hopf bifurcation after its independent originators).

Kim's resonator stage, made nonlinear via a sampling term proportional to $\dot{y}^2$, is similar, but using squared velocity instead of Johannesma's y^2. As Johannesma pointed out, it is known as the Rayleigh equation rather than the Van der Pol equation in that case; he notes that "Comparable phenomena occur if damping is dependent on both y and $\dot{y}$ ..." The equations are sometimes lumped together as Rayleigh–Van der Pol oscillators. The Hopf oscillator is similar, except that it uses the magnitude of a coupled pair of state variables (or a complex state variable), so it distorts less but has similar nonlinear compressive amplitude behavior when driven.

van Netten and Duifhuis (1983) analyzed the Van der Pol oscillator as a model of hair cell nonlinearity, and many others have since used this and other nonlinear systems that exhibit a Hopf bifurcation, in modeling cochlear nonlinearity. Duifhuis (2011) compares the various equations used for "a parabolic damping profile," and traces their history in auditory models back to Hall (1974), who we have discovered foreshadowed our current modeling approach, saying,

> Kim, Molnar, and Pfeiffer have already been able to account for a number of nonlinear auditory phenomena with a nonlinear model of the basilar membrane. Our model differs from Kim's in that it provides a representation of the entire length of the basilar membrane and enables us to observe DPs [distortion products] generated at one place on the membrane model and then propagated to another place. The model used in this paper is physically oriented (i.e., there is an attempt to relate elements of the model to physical elements of the cochlea), it represents the entire length of the basilar membrane, and we introduce an asymmetric as well as a symmetric nonlinearity.

Our DOHC realizes a related nonlinear damping control, with efferent modification, as detailed in Figure 17.1. We set its parameters to stay away from the bifurcation, that is, from the conditional instability where damping goes to zero; its form is also designed to keep the damping bounded, to yield a high-level linear behavior.

The idea of modifying the damping as a *saturating* function that is quadratic around small velocity was introduced by van den Raadt and Duifhuis (1990) (see also Duifhuis (1992)), who wrote

> Initially ... classical parabolic damping profile was used. Our major emphasis, however, shifted toward potentially more realistic biophysical models. For the damping term we now use a function in which two parts can be discerned. First, a passive (positive) part with exponential "tails," which provides response behavior with a log-like characteristic for external stimuli. Secondly, there is an active (negative) part that, if sufficiently strong, produces net active (negative damping = energy production) behavior.

Their negative-damping central region and positive-damping "tails" correspond to the central peak and tails in our NLF of Figure 17.2, though the functional forms we use are not quite the same.

CDTs come about because the damping term multiplies the signal velocity in the differential equation, so quadratic damping leads to cubic distortion. The nonlinear function can also generate a quadratic distortion

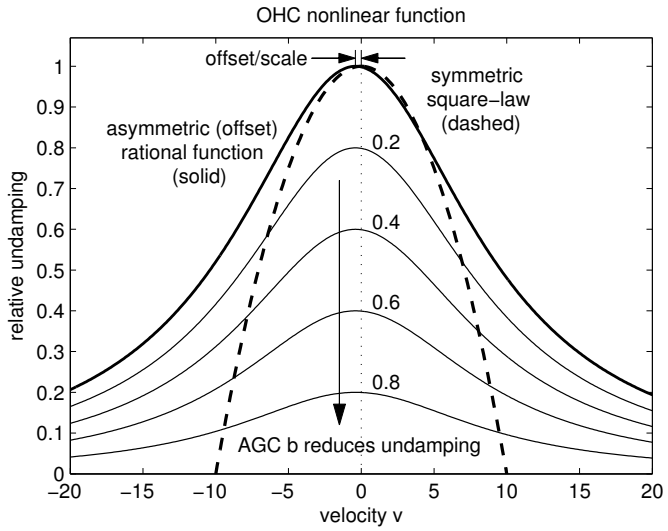


Figure 17.2: The NLF of the DOHC block in Figure 17.1 is shown here as the heavy solid curve. The dashed curve illustrates the sort of symmetric quadratic nonlinearity often used in Hopf oscillators, the Kim model, and various other cochlear models. The lighter solid curves show how feedback from the AGC loop filter multiplies the NLF output by $1 - b$, reducing the relative undamping that the DOHC supplies via this NLF.

tone (QDT) if it has an odd component (as opposed to only even-symmetric components such as $\dot{y}^2$). Keep in mind that, as discussed in Chapter 10, the characterization of nonlinearities by polynomials, as suggested by the terms *cubic* and *quadratic*, is perhaps not as useful as characterizing them in a way that avoids suggesting how the distortion grows with level, for example as simply *odd* and *even*. For reasoning about frequencies, it can be convenient to stick with low-order polynomial concepts, but Duifhuis (2012) points out that here the very terms QDT and CDT are misleading, since distortion products of orders higher than two and three are typically produced by the odd and even distortion components. Smoorenburg (1972) and Duifhuis (1989) propose alternative forms for compressive nonlinearities in the cochlea, using power-law functions with exponents between 0 and 1; these can be adjusted to avoid infinite slope at 0, but we avoid them because the computation of powers is expensive compared to ordinary arithmetic operations.

With an even-symmetric (e.g., squaring) nonlinearity, fluctuations of the filter coefficients (rc_0 and ra_0 in Figure 17.1) will include a double-frequency term that interacts with the BM wave to make a CDT, and thereby generate two-tone interactions such as $2f_1 - f_2$, which (for $f_1 < f_2$) will propagate through the cascade of subsequent filters and be amplified before arriving at its own lower-CF place of localization. For example, 800 Hz and 1000 Hz will generate a CDT at 600 Hz, which will propagate and be localized in a later filter stage (more apically) than the primary tones.

Offsetting the center of the nonlinearity away from the zero velocity point is a good way to give it an odd component, which will result in the damping control having components at the primary frequencies. First-order damping coefficient fluctuations will interact with the primary frequencies in the wave and produce quadratic distortion terms such as $f_2 - f_1$, or envelope components in general, which will propagate and be localized near the low-CF end of the cascade. For example, 800 Hz and 1000 Hz will generate a 200 Hz QDT component, which will propagate and be localized much more apically than the primaries.

A velocity-squared effect grows too rapidly at high velocities, so our formula for the nonlinear function (NLF of Figure 17.1, plotted in Figure 17.2) is a rational function that saturates toward zero instead, making

the damping saturate toward a high-level limit:

$$\mathrm{NLF}(v) = \frac{1}{1 + (v \cdot \mathrm{scale} + \mathrm{offset})^2}$$

with parameters scale = 0.1 and offset = 0.04.

The NLF in the DOHC can be thought of as the slope of an OHC's sigmoidal transduction nonlinearity; the slope approaches zero as the sigmoid saturates in either the positive or the negative direction. According to Geisler et al. (1990), this saturation provides an adequate explanation of two-tone suppression in the cochlea. The lower curves in Figure 17.2 reflect a reduction in the coupling from the transduction to the motor, as controlled by efferent feedback from the AGC loop filter (recall Figure 15.1).

17.3 AGC Effect of DOHC

As shown in Figure 17.1, using the NLF to control active undamping integrates easily with the AGC loop-filter feedback signal b: multiplying the NLF output by $1 - b$ turns down the gain by scaling back the undamping effect. Keep track of the multiple inversions here: high input level → high b → low $1 - b$ → low undamping → high damping (low r) → low gain → reduced (compressed) output level.

The maximum effect of undamping is adjusted via the d_{rz} parameter as shown in Figure 17.1 to reduce the damping to a minimum small-signal damping value. If the small-signal damping is zero or slightly negative, then the small-signal gain is infinite—but then the output level reduces the gain to a finite value by making the damping positive. Such a system is said to be *poised at a bifurcation* between stable and unstable regions (Moreau et al., 2003). If undamping is actively controlled by adding energy, then it seems likely that the damping may sometimes be driven to zero or negative, generating spontaneous output. Such bifurcation may be possible in the cochlea; whether this is normal, or a case of objective tinnitus, may depend on just how loud the result is. For machine hearing purposes, keeping the damping positive and the gain finite seems like a better idea.

The gain of the CAR filter channel is approximately a power of reciprocal damping—typically about fourth power, depending on how many stage resonant peaks overlap at a frequency. Damping grows in proportion to the AGC feedback level b, so the AGC gain is approximately like the $K = -4$ AGC case that we analyzed in Chapter 10. The b value will be constrained to not exceed 1 by a saturating level detector in the inner hair cell; see Chapter 18.

The NLF and the AGC feedback b affect the pole radius via the formula:

$$r = r_1 + d_{rz}(1 - b)\mathrm{NLF}(v)$$

for v being the local BM velocity. At a high velocity in either direction, the NLF approaches zero, so the gain due to active undamping is suppressed during two phases (positive and negative velocity regions) of a strong low-frequency signal passing through, as has been seen in physiological data (Ruggero et al., 1992).

This suppression is asymmetric (as modeled by our offset parameter), so the resulting distortion is not pure odd-order or even-order, but a mix. It only takes a little offset to make a QDT that propagates, with active gain, to a lower-CF place and provokes a significant response. Whether these distortion tones are useful in machine hearing applications is not yet clear, but they do seem to have a role in normal sound perception (Pressnitzer and Patterson, 2001).

An early description of the integration of the outer hair cell nonlinearity into a cochlear AGC concept, and its effect on the relative level of distortion products, was given by Jont Allen (1981):

> The automatic gain control nonlinearity also explains why the harmonic distortion is always

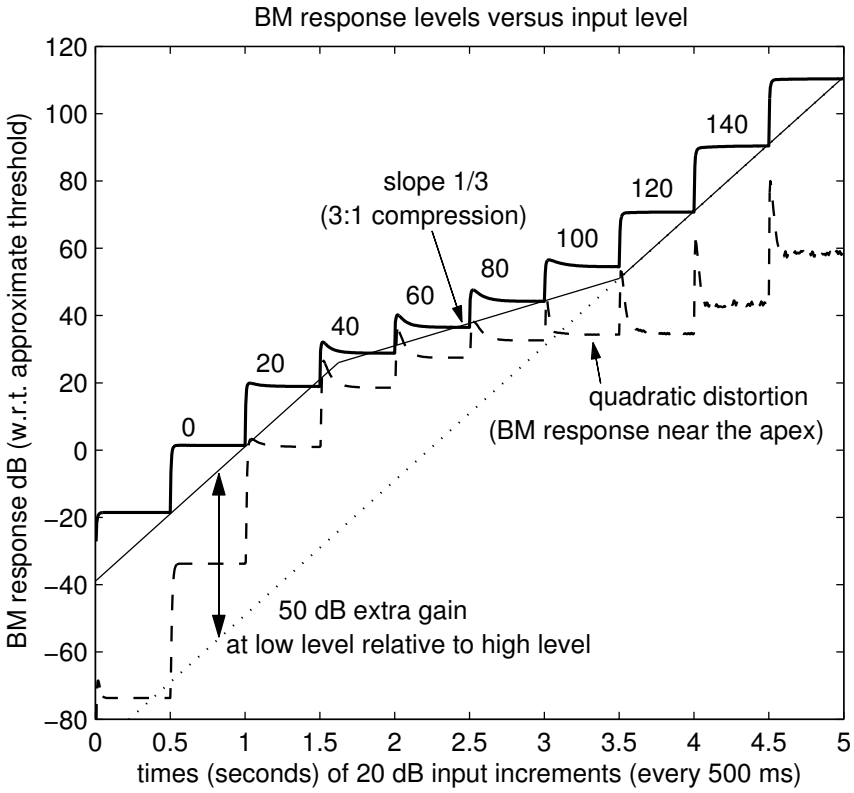


Figure 17.3: CARFAC responses versus stepped input level, at two places (solid and dashed), for a 4-tone input (1.6, 1.8, 2.0, and 2.2 kHz). The response for a place with CF near 1.7 kHz (solid) is compressive, but approaches linear at both very high and very low levels; thin lines approximate the steady-state response levels at which each input level step settles. The low-level linear region has 50 dB of gain compared to the high-level linear region; compare Figure 15.3. The DOHC scheme with offset asymmetry leads to a fairly high level of response to quadratic distortion products at a place with CF near 200 Hz (dashed), approximately tracking the level of response to the primary tones through most of the normal compressive range of hearing. The low-level and high-level linear regions generate relatively less distortion, as indicated by the relative response level at the QDT place. In a real ear, other parts of the system would likely distort strongly at high levels.

> below the primaries in intensity and does not grow large at large input levels as would be predicted from a power law nonlinearity. In fact, the intermodulation distortion never seems to be greater than –15 dB equivalent ear canal sound pressure level relative to the primary signals.
>
> The source of the nonlinearity remains unknown, although there are some good reasons to believe that its generation is related to motions of the stereocilia of the outer hair cells. It presently seems clear that this source of distortion is not the byproduct of some poorly engineered component. It is rather perhaps the negligible residual of a sophisticated local feedback mechanism in the mechanical motion of the properly operating cochlea, such as the automatic gain control system mentioned previously.

Allen's comparison of "–15 dB equivalent ear canal sound pressure level relative to the primary signals" is reasonably consistent with response level differences less than 15 dB as shown in Figure 17.3, due to the compression.

17.4 Typical Distortion Response Patterns

It is hard to evaluate or characterize an outer hair cell model in isolation, so we illustrate its response patterns in the context of the complete CARFAC, including the feedback loop of components to be detailed in subsequent chapters.

Figure 17.3 shows the input–output level response at the BM (that is, not as viewed through the IHC or neural response, but in the wave mechanics). This example includes the output level of the QDTs as measured in channels near the apex, as well as the nearly linear response of channels near the base. The stimulus is a 4-tone complex (1.6, 1.8, 2.0, and 2.2 kHz) in cosine phase, which has a peaky envelope and elicits a strong buzzy 200 Hz pitch sensation.

The three-region linear–compressive–linear input–output response is the cooperative behavior of the CAR, IHC, AGC, and OHC components. At low levels, where the AGC feedback is negligible, the OHC provides maximum undamping, and at high levels, the OHC provides negligible undamping, so both of these regions yield linear filterbanks. The middle region has about a 3:1 (cube-root) compression characteristic, changing the gain to the 4-tone complex by about 50 dB over a 75 dB range of input levels.

The level dependence of QDTs is complicated (Cooper and Rhode, 1997). It is inherently quadratic (expansive) in the small-signal region. In the CARFAC with default parameters, the response level of the 200 Hz QDT is roughly in proportion to that of the primary tones, about 8–10 dB down, through the mid levels of hearing (40–80 dB SPL at least), as shown in Figure 17.3. This level may be unrealistically high, according to the observations of Pressnitzer and Patterson (2001), though we have not yet tried to quantify it with their cancellation paradigm.

As a function of place, the response levels are as shown in Figure 17.4. The very compressive and nonmonotonic nature of the QDT response described by Cooper and Rhode (1997) is apparent (near channel 60), though he saw nonmonotonicity at a much lower level.

Figure 17.5 shows a scatter plot of samples of BM velocity and the resulting instantaneous relative undamping that the OHC provides, in the context of the same 4-tone complex at different levels. At very low levels, the relative undamping stays close to 1.0, which reduces the CAR stage damping from 0.35 to 0.1. At each level, the range of variation of the damping remains moderate, though it could drop from near 1 to near 0 instantly at a strong onset. For the scatter plot, only the channels most responsive to the primary tones are sampled, and only after the response level has settled. The range of BM velocities sampled roughly doubles (6 dB increase in response) for each 20 dB increase in input level—close to cube-root compression.

A 20 ms segment of the BM response of all channels is shown in Figure 17.6, along with another plot of what it would look like without the offset asymmetry in the OHC's NLF, for the 4-tone complex at 60 dB

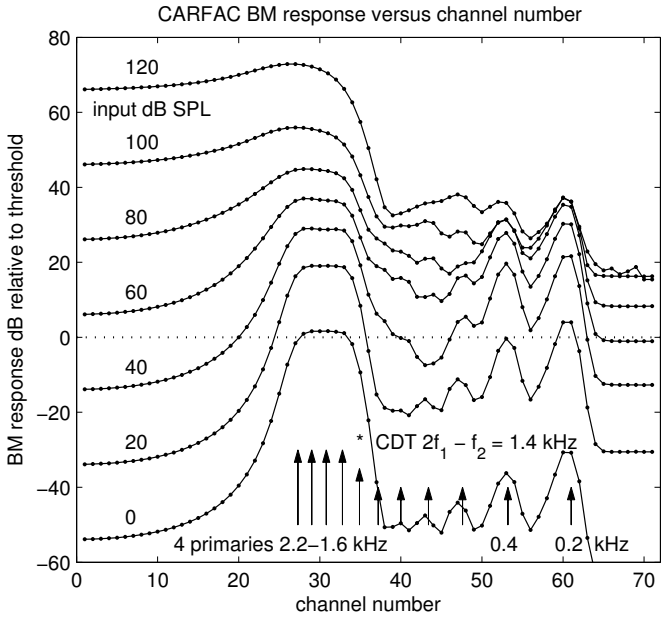


Figure 17.4: The CARFAC's steady-state BM response level at all places (channels), for some of the input levels used in Figure 17.3. The input is a four-tone complex at 1.6, 1.8, 2.0, 2.2 kHz (longer arrows mark places with CFs corresponding to these primaries). Places with CFs at lower multiples of 200 Hz (shorter arrows) also respond, especially to quadratic distortion at the 200 Hz and 400 Hz places. The first low-side odd-order distortion frequency, whose 1400 Hz place is marked with an asterisk, can be interpreted as the $2f_1 - f_2$ CDT of the two lowest primaries. Though it is not spatially resolved, the response at this place is dominated by a 1400 Hz component. The horizontal dotted line at 0 dB response level represents an approximate detection threshold, suggesting that quadratic distortion may be audible even for an input level as low as 20 dB SPL.

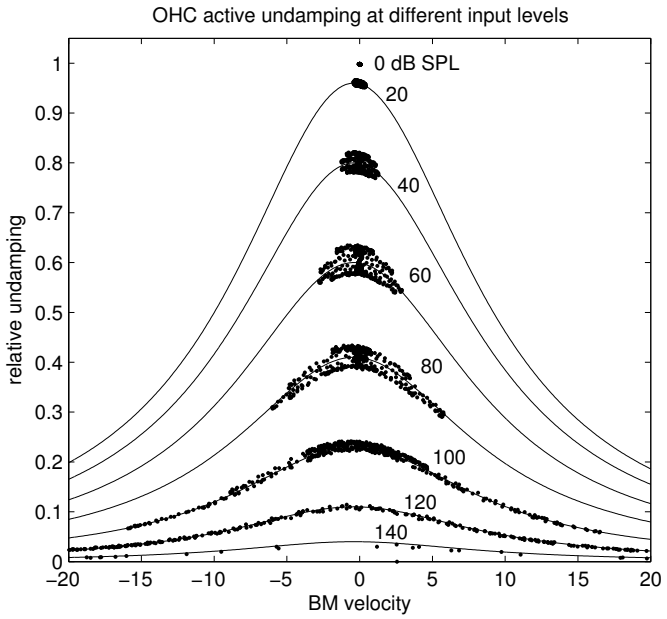


Figure 17.5: The OHC active undamping effect at various input levels, with 0–140 dB SPL input levels labeled. The input is a four-tone complex at 1.6–2.2 kHz, and the OHC effect is sampled near the most responsive place. Thin solid curves are scaled copies of the NLF that the points approximately fall on; that is, where the points might be for steady values of b, as in Figure 17.2. At the highest and lowest levels, the damping is nearly constant throughout the period of the stimulus, so relatively little distortion is generated (at 140 dB SPL, the BM velocity extends far outside the domain plotted, so the small bump in the middle has little effect).

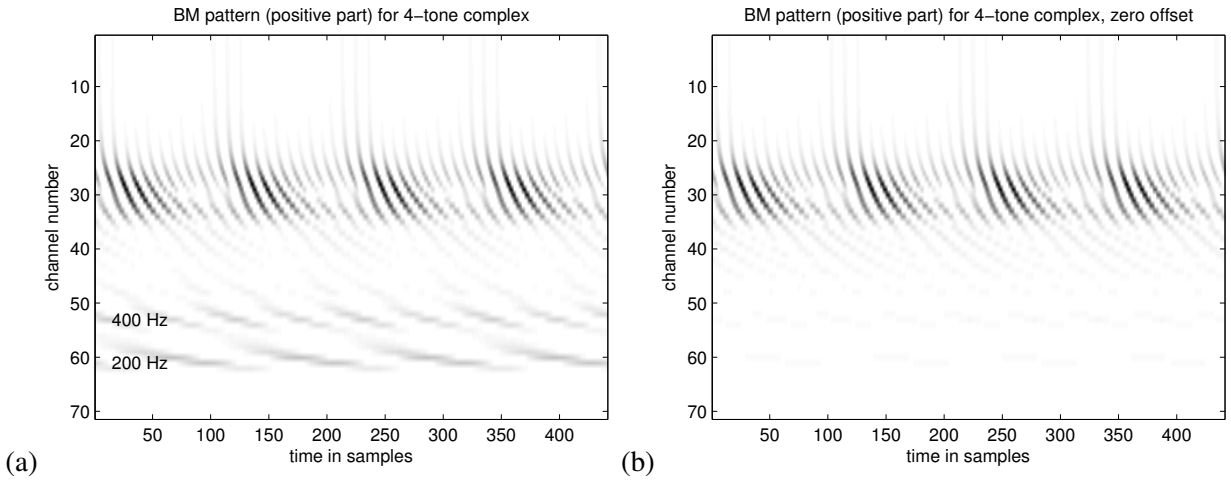


Figure 17.6: These cochleagrams show the positive part of the BM motion (filter outputs) for the 4-tone complex stimulus at 60 dB SPL, the level at which the relative amounts of 200 Hz and 400 Hz quadratic distortion tones is highest with the default NLF offset parameter. The left (a) image is for the default CARFAC, and the right (b) is with zero offset in the NLF. The 20 ms segment encompasses four cycles of the 200 Hz missing fundamental. The relatively small offset asymmetry shown in Figure 17.2 and Figure 17.5 is enough to cause relatively large QDTs.

SPL, where the relative level of the QDT is greatest. The difference in the QDTs with and without the offset is obvious. What is less obvious is that there is still a small QDT response even in the zero-offset case (too small to see in the figure), since the small primary-frequency ripple in the output of the AGC smoothing filter interacts with the signal just as the odd component of the NLF output does. This was the only mechanism for quadratic distortion in our previous-generation models, and has been discussed by Patterson et al. (2013).

CDTs are usually assessed by a spectral analysis, since they typically don't give rise to resolved peaks in the place dimension. Figure 17.7 shows the spectra of the BM response for all channels, obtained by Fourier transforming the segment shown in Figure 17.6 (the segment on the left there, with the OHC offset). At this level (60 dB SPL), all orders of distortion tones are apparent. At lower levels (40 dB SPL), there is very little response at 800 and 1000 Hz, as the QDTs are creeping up from 200, 400, and 600, while the CDTs are creeping down from 1400 and 1200.

17.5 Completing the Loop

In this chapter, we have focused on the outer hair cell model, but the responses that we show depend on the rest of the CARFAC components, including the inner hair cell model and the AGC loop filter that we cover in coming chapters, to complete the feedback control loop. Here we focus mostly on the steady-state level-dependent behavior, when the loop has settled so the AGC loop dynamics are not important, though we do see a little bit of dynamic response to steps in Figure 17.3 and some jitter due to imperfect AGC smoothing in Figure 17.5.

In the next chapter, on the inner hair cell, we'll see how saturation of the output of that detector element is an important part of achieving a high-level passive linear region. Then we will finish Part III with a chapter on the CARFAC's fancy coupled AGC loop filter that controls of the dynamics of the *b* input to the DOHC in response to level changes.

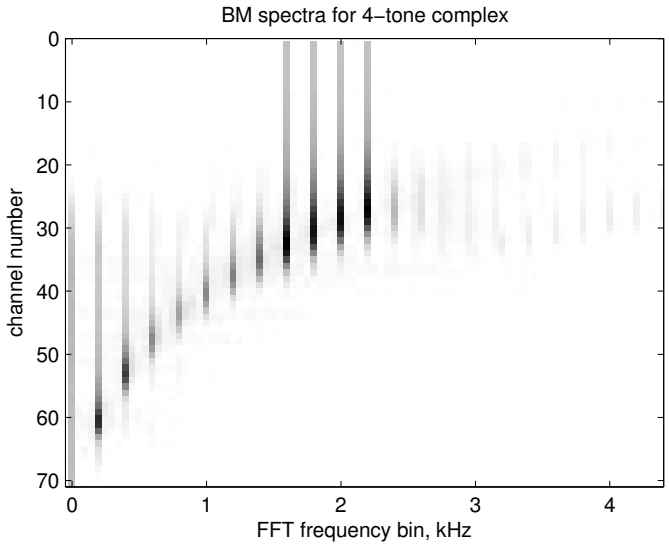


Figure 17.7: The CARFAC BM spectra for all channels, in response to the 4-tone complex at 60 dB SPL; that is, FFT magnitudes of the 20 ms segment shown in Figure 17.6(a), with each channel's spectrum plotted as a row. Visible distortion components include QDTs, including a DC response, and CDTs of all orders. For example, channels 35–40 show strong CDT $2f_1 - f_2$ (1400 Hz) and $3f_1 - 2f_2$ (1200 Hz) components (relative to the two lowest-frequency primaries). The tails above the peaks show where each DT component propagates from: primaries from the base, and distortion products from the region of strong response to the primaries. High-side distortion tones are very weak, as they have no chance to propagate through a region that amplifies them. The amplitude scale is cube-root compressed (sixth root of power) to make weak components visible in this plot.

Chapter 18

The Inner Hair Cell

> The sensitivity of the hair cells is extraordinary: the slope of the input–output curve can reach 20 mV per micrometer of displacement. If hair cells, like photoreceptors, can synaptically transmit statistically significant signals corresponding to 10 μV receptor potentials, the threshold sensitivity of the amphibian sacculus would approximate 500 pm (5 Å).
>
> — "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," Hudspeth and Corey (1977)

The inner hair cells (IHCs) are the transducers that sense sound-generated motion of the cochlear partition and deliver the result as afferent signals to the nervous system. Their position in the organ of Corti is shown in Figure 18.1.

Oscillatory waves in the fluid–membrane system of the cochlea, represented in the CARFAC model as the BM outputs y_i of Figure 15.2, couple to the IHCs' stereocilia (hairs), and thereby control receptor currents that change the hair cells' endo-cellular potentials and ion concentrations. The IHCs respond to these receptor effects in much the same way that other sensory cells do—for example, much as the retina's cone cells respond to light—by releasing neurotransmitter, which stimulates connected neurons to *fire* (create action potentials). The IHCs themselves don't fire as neurons typically do, but the connected primary auditory neurons do.

Acting as detectors, or rectifiers, the IHCs convert zero-mean bandpass signals into firing rates and neural fine time structure. Besides wave detection, IHCs have other signal-processing functions, including emphasizing onsets, responding to temporal structure over a range of time scales, and realizing further dynamic-range compression via adaptive nonlinear mechanisms. The IHC's short-time-mean output plays a key role in the AGC feedback loop that controls the adaptive gain and distortion in the mechanics. In the CARFAC model, the IHC block needs to be designed and evaluated in the context of the other elements; the OHC model relies on the IHC's saturating detection characteristic to impose a bound on the feedback signal produced by the AGC smoothing filter. The IHC's internal AGC relies on this saturation, too.

18.1 Rectification with a Sigmoid

As mentioned briefly in Chapter 14, the detection nonlinearity is not an ideal half-wave rectifier. One direction of cilia motion increases the receptor current, a positive ion current into the cell, and other direction shuts it off, but the curve does not have a sharp corner at zero displacement. The inward ion current causes a *depolarization* of the cell—that is, a reduction in the electrical potential difference between the inside and outside of the hair cell. Depolarization is the direction that corresponds to an *excitatory* response in neurons and hair cells. On the other hand, closing the ion channels leads to a small *hyperpolarization*, the direction corresponding to inhibition of response.

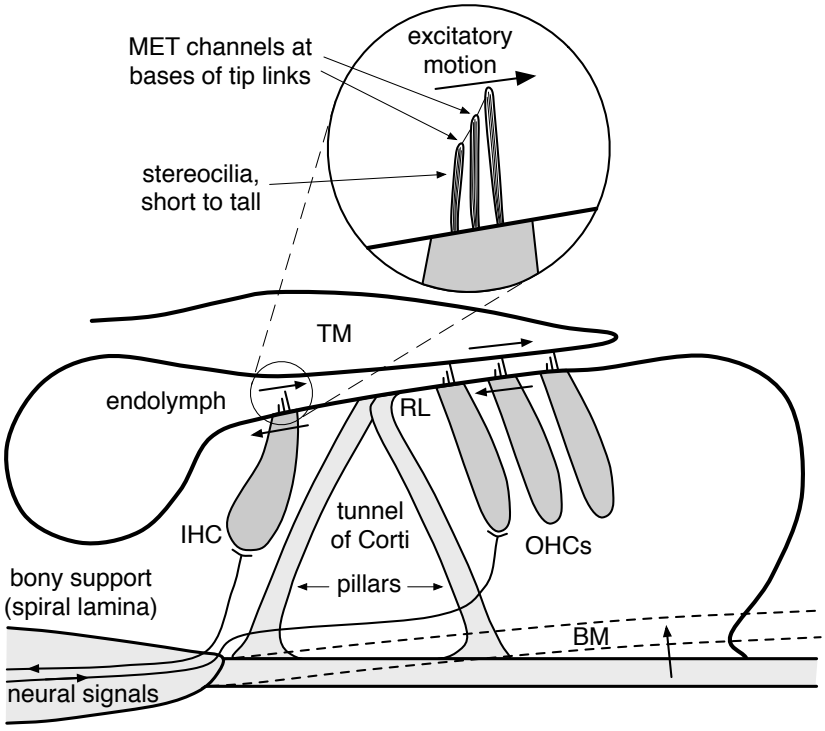


Figure 18.1: In the organ of Corti, the hair bundles of the inner and outer hair cells (IHC and OHC) are displaced by a shearing motion between the reticular lamina (RL) and the tectorial membrane (TM) when the organ of Corti pivots about the inside corner of the tunnel of Corti due to displacement of the basilar membrane (BM; exaggerated displaced position shown dashed). When the motion is in the direction of the arrows, the tip links between adjacent stereocilia (hairs) pull the mechano-electrical transducer (MET) channels open, allowing a positive-ion current to flow from the endolymph into the hair cells at the tips of the shorter cilia. The OHCs feed energy back into the hydromechanical wave, to a degree controlled by neural signals from the brain, while the IHCs send neural signals toward the brain.

Biophysics Connection: How Hair Cells Work

Hudspeth and Corey (1977) found that deflection of hair bundles causes a change in the ionic currents into hair-cells, following a sigmoidal curve of the sort described in the text. It was later conjectured (Hudspeth, 1982), and eventually accepted, that these currents are primarily through the stereocilia, via *mechano-electrical transducer* (MET) channels at their tips (Jaramillo and Hudspeth, 1991; Lumpkin and Hudspeth, 1995). In particular, current flows into each stereocilium where it is tied to the next longer one by a *tip link*, a thin chain of proteins that is tensioned when the cilia bundles are bent in one direction, and loosened when they are bent in the other direction. See Figure 18.1.

The tip link mechanically opens and closes the MET channel, through which positive ions (potassium and calcium, mostly) in the endolymph enter the hair cell in this first step of the mechanical to neural transduction. Many details of how this transduction works, including tip-link molecular mechanisms, have been worked out (Gillespie and Müller, 2009; Sakaguchi et al., 2009), though details remain elusive (Fettiplace and Kim, 2014; Zhao and Müller, 2015).

At rest, the MET channels rapidly *flicker* between open and closed, under thermal agitation, with a probability of about 0.1 or more of being open. For high displacements, at some point most of the channels are open and more displacement will not further increase the current; for displacements in the other direction, most are closed and the current will approach zero. Though there are only about two channels per stereocilium, each can pass a large ion current. The statistics of these conductance fluctuations give rise to a sigmoidal (*s*-shaped) detection nonlinearity.

Receptor potentials are often displayed with depolarization positive and hyperpolarization negative, with respect to a resting potential at zero. We take a different approach here, showing zero conductance, or the limit of hyperpolarization, as the zero line, and the conductance in quiet as a small positive response level. This nonzero conductance response in quiet is conceptually related to the small spontaneous firing rate of auditory neurons in quiet. The relations between AC and DC conductance, receptor potential, neurotransmitter release, and neuron firing rates are determined by the hair cell and neuron models, and are not necessarily simple.

For the rational-function sigmoidal detection nonlinearity of Figure 18.2, which is the one we use in the CARFAC model, a displacement peak amplitude of 2.0 represents a strong mechanical response, mostly saturated, as shown in Figure 18.3.

If the detection nonlinearity has no sharp corner at zero, the response for very small signals is close to linear (nonrectifying), plus a little bit of quadratic distortion (square-law rectification, or power detection). In quiet enough conditions, there is a nonzero mean output that is relatively insensitive to signal level—a flat low-level region on a plot of mean response versus log level. Even when there is enough signal to make the response fluctuate, the mean will not be much affected until the level is high enough to create a significant second-order distortion, as illustrated in Figure 18.4. The firing rates of subsequent auditory neurons will reflect the hair cell's response, in terms of both mean firing rate and instantaneous firing rate.

The level at which the mean response increases significantly from the response in quiet is known as the *mean-response threshold*, or in the case of neural firings of the attached spiral ganglion afferents (the primary auditory neurons), the *mean-rate threshold*. There is no actual threshold mechanism at work here, in the sense of a level that when crossed triggers an event or closes a switch. In fact, psychophysical sound detection thresholds are lower than mean-rate thresholds. We are able to hear weak sounds through patterns of nearly random neuron firings, via instantaneous rates being modulated symmetrically above and below their mean rate in quiet. The neurons use their own noise—random firings—to enhance the detectability of otherwise too-weak signals, in an application of the concepts known as *dithering* and *stochastic resonance* (McDonnell and Abbott, 2009).

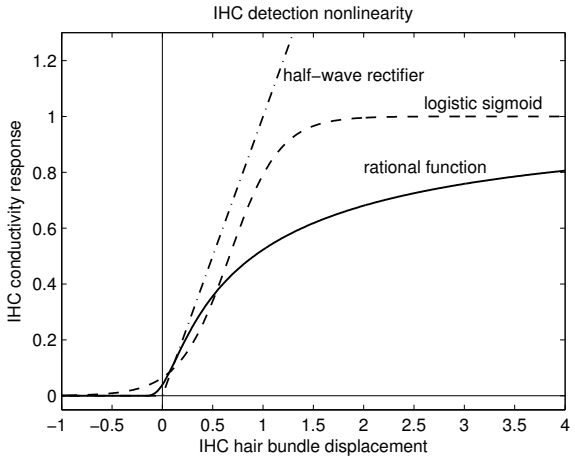


Figure 18.2: The transduction nonlinearity of the IHCs is some kind of a *sigmoid*, such as a displaced logistic function (dashed), and is sometimes modeled as simply a half-wave rectifier (dash-dot line). Other functional forms can also be used; for example, a constant segment at zero response, connected to a rational function (ratio of cubic polynomials) for the rest (solid curve), giving a cubic foot shape, a nearly linear middle region, and a slowly saturating shoulder.

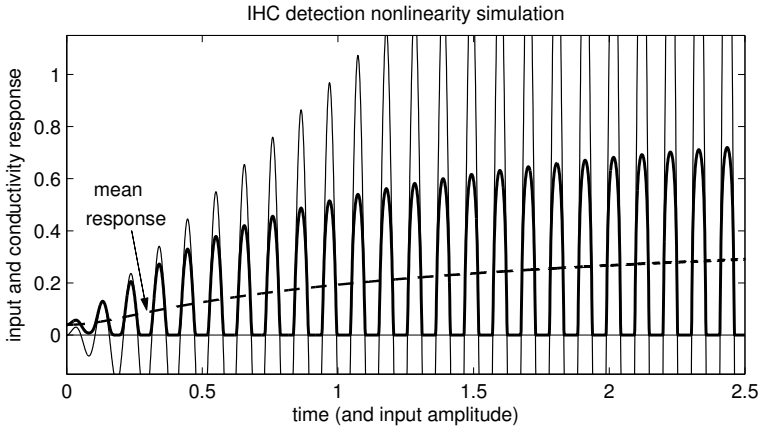


Figure 18.3: A simulation of the output of the rational-function detection nonlinearity of Figure 18.2 (heavy solid curve), when it is driven by an increasing-amplitude sinusoidal input (thin solid curve); this output function of time is the conductance $g(t)$ used in the digital IHC model of Section 18.3. The mean response (dashed curve) is also shown; for the purpose of this illustration, the mean is taken over many phases of the input sinusoid.

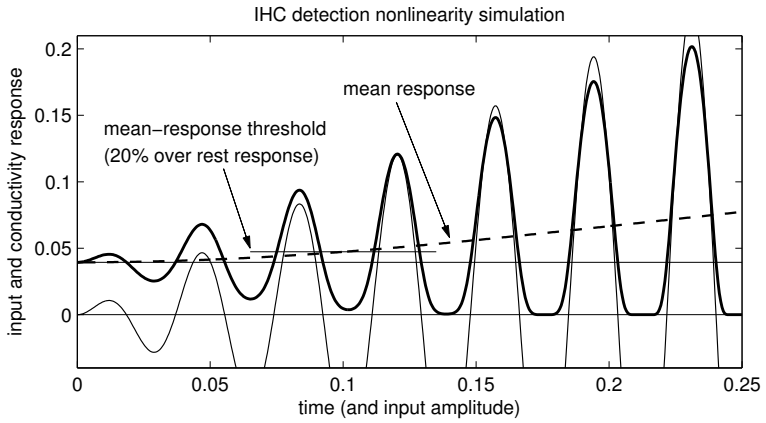


Figure 18.4: A simulation as in Figure 18.3, but for a more limited time and amplitude range, and using a higher frequency, to better illustrate the transition from a nearly linear response at low amplitude to a rectifying response at higher amplitudes. The mean response (dashed curve) increases very slowly (initially quadratically) where the response is nearly linear. It increases from the rest level to 20% higher as the input amplitude increases to about 0.1, where the response is distorted enough to start to look rectified.

The mean-response threshold is about 0.1 (on the arbitrary normalized displacement scale of the rectification nonlinearity), though there is good nearly-linear synchrony response below this level, as shown in Figure 18.4. The actual dynamic range of sound levels from threshold to saturation is much greater than this 26 dB difference at the hair cell, due to compression in the cochlear mechanics preceding the detection function. In the healthy functioning cochlea, the hair cell's response dynamic range is more than 70 dB at mid frequencies (Cheatham and Dallos, 2000), corresponding to roughly a 3:1 compression in the mechanics before the waves reach the IHCs.

18.2 Adaptive Hair-Cell Models

Davis (1957, 1965) proposed models of hair-cell transduction in which hair bundle displacement modulates the conductance, or resistance, of the hair cell's interface with the endolymph in the cochlear duct, using rectifying functions such as the sigmoids described above. His models have been refined and elaborated over the years; this section reviews some of these more modern models.

The IHC is much more than a detection nonlinearity. It is adaptive, and thereby compressive, reducing its response gain quickly after a signal onset, producing an output of reduced dynamic range. There are various good models for this IHC behavior, generally involving one or more *reservoirs* of neurotransmitter or potential or other key ingredient that can be quickly used up and only slowly replenished. In the simpler models, the receptor current into the cell and the neurotransmitter current out are the same signal; the current is proportional to the conductance, the rectified input signal, and to the adaptive state variable of the model, which can be interpreted as an intracellular potential, or an amount of neurotransmitter available.

Modern studies (Gillespie and Müller, 2009) suggest at least two adaptation time constants associated with the transduction channels themselves, of around 1 and 10 ms. We are not yet including those effects in the models.

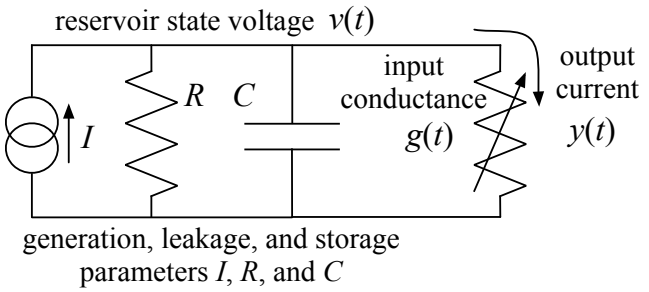


Figure 18.5: The Schroeder–Hall hair-cell model is described by this circuit schematic. The state variable is the voltage $v(t)$ across the capacitor, which is charged up by the current source I and discharged by currents through the fixed resistance R and through the input-controlled variable resistor with conductance (reciprocal of resistance) $g(t)$. The current through the variable resistor (the resistor symbol with an arrow through it) is the output signal. The same schematic can describe the Allen model, but there a saturating nonlinearity is used instead of the HWR for the $g(t)$ detection nonlinearity, and an output smoothing filter is added to reduce synchrony to high frequencies.

18.2.1 The Schroeder–Hall Model

In the Schroeder–Hall model (Schroeder and Hall, 1974), a soft half-wave rectifier (soft HWR) function of displacement determines a conductance $g(t)$ (a function of displacement, which is a function of time) that allows neurotransmitter to flow from reservoir to output, as illustrated in Figure 18.5. The soft HWR follows the straight HWR at the high end, but has a smooth curve toward zero rather than a sharp corner at the low end. The reservoir (modeled by the charge on a capacitor) is replenished at a limited rate in a first-order RC circuit, as shown in Figure 18.5.

$$C\frac{dv}{dt} = I - \frac{v}{R} - y \quad \text{where} \quad y = gv \quad \text{is the output current.}$$

The average high-signal output is upper bounded by the limit of the replenishment rate, so the average output saturates even if the detection nonlinearity doesn't. The instantaneous output, however, can be very large if the input, the BM displacement controlling $g(t)$, is very large. A too-big conductance leads to unrealistic too-strong emphasis of strong onsets; so a saturating sigmoid may still be a better detection nonlinearity for this model.

The recovery time constant of the model is the RC product; it determines the exponential replenishment of the reservoir potential after $g(t)$ goes to zero; we typically use about 10 ms. The attack or onset time, however, or how quickly $v(t)$ can fall, reducing the gain at a strong onset, can be much shorter. In terms of automatic gain control concepts as analyzed in Chapter 11, this faster attack time corresponds to the speedup factor in a $K = 1$ AGC loop.

18.2.2 The Allen Model

The Allen (1983) hair-cell model is similar to the Schroeder–Hall model, but uses a sigmoidal nonlinearity, rather than the soft HWR, to model the effect of hair-bundle displacement. Though Allen drew it and characterized it somewhat differently, the same circuit as shown in Figure 18.5 can apply. Allen identifies the capacitor state with an endocellular receptor potential, as opposed to a neurotransmitter reservoir. The response of the model is more realistic in that its sigmoidal conductance nonlinearity avoids the big overreaction at strong onsets.

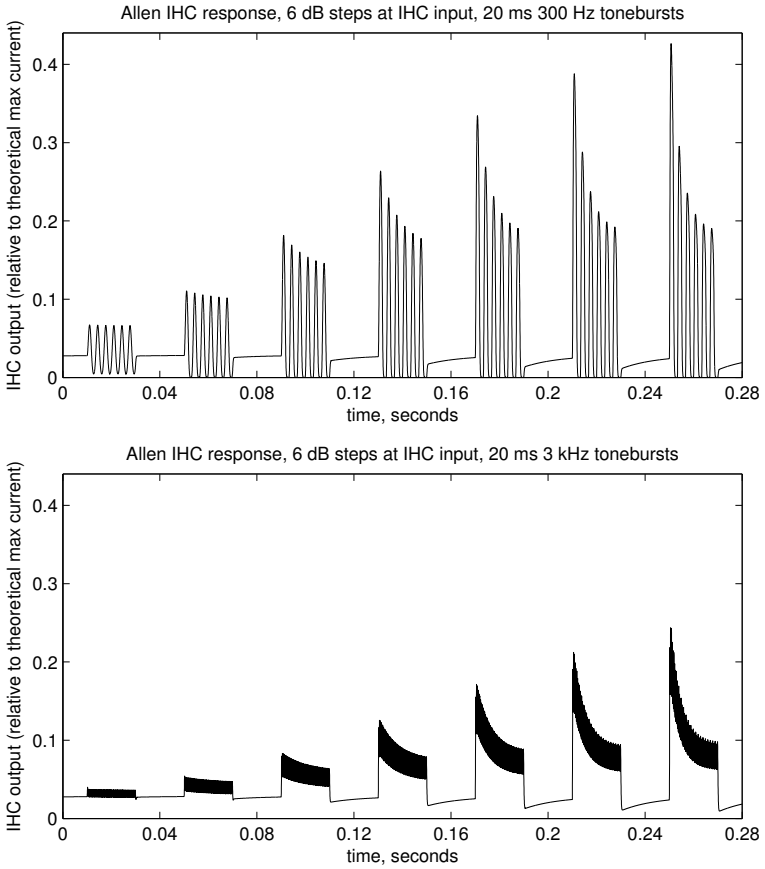


Figure 18.6: The response of the Allen IHC model to 20 ms 300 Hz (top) and 3 kHz (bottom) tone bursts in 6 dB increments, starting 6 dB below the mean-response threshold (with no cochlear filtering). A fairly strong onset emphasis at high levels, such as that exhibited here, is a key property of inner hair cells and their models. The 3 kHz synchrony is attenuated by the lowpass filter. There remains very good synchrony to onsets.

In addition, Allen includes higher-order lowpass dynamics of the neurotransmitter effect—that is, smoothing of the output current signal—reducing the synchrony to waveform details for frequencies above about a kilohertz. The lowpass filter that Allen suggests is a model of diffusion, realized as a multistage RC circuit, corresponding to several real poles. The exact details of such a filter aren't critical; the impulse response will be close to a gamma distribution (Papoulis, 1962).

Simulations of our rational-function adaptation of the Allen model for low and high frequencies are shown in Figure 18.6.

18.2.3 The Meddis Model

The Meddis (1986, 1988) model adds more state variables, representing other places that could have neurotransmitter in them. In particular, used neurotransmitter in the synaptic cleft can be recycled by reuptake mechanisms, increasing the available neurotransmitter in the free transmitter pool more quickly than new neurotransmitter can be brought in. Reuptake also reduces the forward effect of the neurotransmitter release on the postsynaptic cell (the primary auditory neuron), since it reduces the concentration in the synaptic cleft.

Meddis also incorporates a local store between the neurotransmitter factor and the free transmitter pool—an extra RC filter stage. And he uses a soft saturating nonlinearity, much like our curve in Figure 18.2.

18.2.4 Choosing a Model

Van Compernolle (1991) presented an analysis and comparison of the Schroeder–Hall and Meddis models. Similarly, Hewitt and Meddis (1991) have done studies and parameter tuning of various models, paying attention to synchrony, firing rate, and onset emphasis for a wide range of stimulus levels and steps. But both of these works were in the context of an otherwise linear basilar membrane model; that is, the entire burden of nonlinear compression was on the hair-cell model. In later papers, Meddis and his colleagues presented IHC models that work with a filterbank that models compressive cochlear mechanics (Sumner et al., 2002, 2003a). The hair-cell and neural mechanisms in these models are very detailed, and may be useful for making the models more accurate on a range of phenomena, but are perhaps beyond what we need for machine hearing applications.

In the above-mentioned tests with linear sound input to hair-cell models, the soft-HWR nonlinearity of the Schroeder–Hall model gave way too much onset firing rate. Both a limited sigmoidal nonlinearity and a nonlinear filterbank contribute to fixing that defect. Such models are still not as realistic as a model with more state variables, such as the various generations of Meddis model. As a compromise, we use an adaptation of the Allen model, with a saturating rational-function sigmoid detection nonlinearity, in the CARFAC implementation, as it captures most relevant effects and is simpler than the Meddis model.

18.3 A Digital IHC Model

A block diagram of our digital inner hair cell (DIHC) is shown in Figure 18.7. Besides the linear filters at its input and output, the update equations for the DIHC can be deduced from the circuit of Figure 18.5, in combination with the conductance nonlinearity of Figure 18.2 or similar.

The nonlinear function (NLF) detection nonlinearity (the rational-function sigmoid plotted in Figure 18.2)

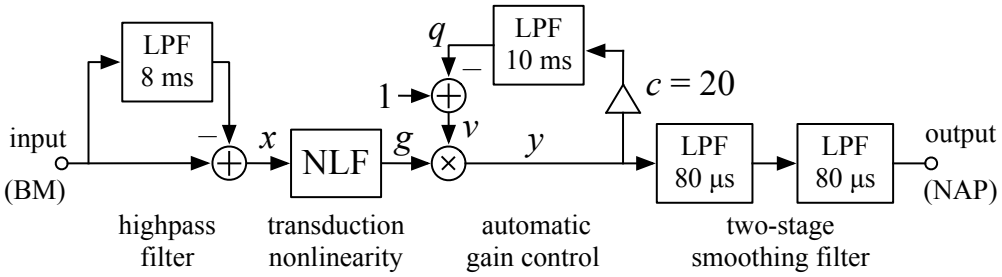


Figure 18.7: The Digital IHC block diagram, an adaptation of the Schroeder–Hall and Allen IHC models. The model uses four instances of the first-order IIR digital filter of Figure 7.1, configured as smoothing filters (lowpass with unity gain at DC). In the diagram, the lowpass filters are labeled LPF, with their respective time constants The first LPF is subtracted to make a highpass filter to suppress frequencies below 20 Hz that are generated from quadratic distortion in the BM wave propagation. The second LPF is the loop filter in an automatic-gain-control loop like the one of Figure 11.2, with $K = 1$ in the nonlinear gain-control function, but with rectifying nonlinearity before the variable gain rather than after. The rectifying nonlinear function (NLF) that converts BM motion to a membrane conductance is a soft rectifying rational-function sigmoid like the one shown in Figure 18.2. The variable gain v models the capacitor voltage of Figure 18.5. The final two LPFs smooth the output.

is computed this way:

$$u = \text{HWR}(x + 0.175)$$
$$g = \frac{u^3}{u^3 + u^2 + 0.1}$$

where x is a highpass-filtered (*AC coupled*, in EE parlance) version of the BM motion, from a highpass filter designed to suppress frequencies below about 20 Hz. In the real cochlea, such low frequencies would be mostly canceled by a reflection from the *short circuit* at the helicotrema, the hole at the apex of the cochlea where only subsonic frequencies arrive. Since we don't model reverse waves in the cascade filterbank, we instead suppress subsonic frequencies explicitly here, at the input to the ICS. HWR is half-wave rectification (positive part), the operator that makes the intermediate variable u to capture the break between the zero segment and the rational-function segment in the NLF. The constants 0.175 and 0.1 are part of the definition of this NLF.

With this conductance $g(t)$ as input, the adaptive-gain part of the model, with $v(t)$ as gain, works like the $K = 1$ AGC loop of Chapter 11 (though the signal names are different), this way:

$$v = 1 - q \quad \text{(using lowpass loop filter state } q\text{)}$$
$$y = gv$$
$$q_{NEW} = q + a(cy - q) \quad \text{(loop filter state update step)}$$

The coefficient $c = 20$ (see Figure 18.7) effectively defines the output level of the IHC's AGC as cy. The $K = 1$ gain function ($v = 1 - q$) keeps the average output level below 1, the level that would drive the gain v to zero, so it keeps the average IHC output y below 1/20. The LPF smoothing coefficient a implements the filter time constant RC ($a = 1/220.5 = 0.0045$, from time constant $RC = 10$ ms, at 22 050 Hz sample rate).

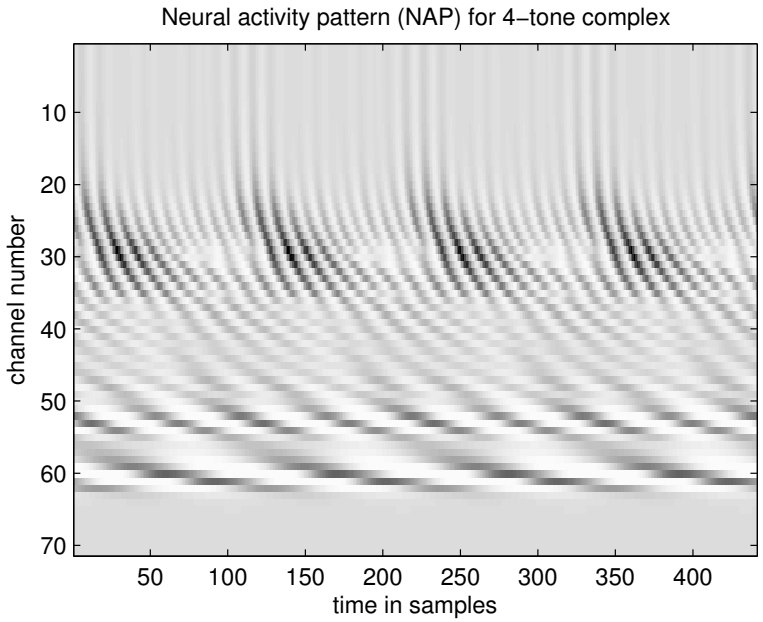


Figure 18.8: The neural activity pattern (NAP), the response of a bank of DIHCs in the context of the CARFAC, with the 4-tone stimulus of Section 17.4, at 60 dB SPL. The NAP represents, at least conceptually, the instantaneous firing rates of the groups of primary auditory neurons attached to each IHC.

In the CARFAC model, these values are specified as two time constants: 10 ms time constant of reservoir replenishment (input), and 0.5 ms corresponding to discharge (output) time when the output is saturated at $y = 1$ ($c = 20$ being the ratio of these times).

The output y is then lowpass filtered with two stages of first-order filtering with time constant 80 μs (corners at 12 500 rad/s or about 2 kHz). With the chosen formulation, the values of g, v, y, and the final smoothed neural activity pattern (NAP) are always between 0 and 1; they could be scaled and associated with physical units if desired, for example for comparison with biophysical models. The BM input x need not be bounded, since the NLF saturates, but it will usually be kept within a moderate dynamic range, with peak values not too far above 1, by the cochlea's compressive wave dynamics.

An example output of a bank of DIHCs in a CARFAC is shown in Figure 18.8, for the input being the BM response segment that was illustrated in Figure 17.6(a). The gray background corresponds to the nonzero rest-level output in quiet conditions.

Chapter 19

The AGC Loop Filter

> ... the output (BM displacement or velocity) varies much less than the stapes input displacement or velocity, for frequencies near the best frequency. The significance of this important finding will become clearer as we proceed, but, in my opinion, it is a precursor to an automatic gain control system which seems to be built into the cochlear filters.
>
> — "Cochlear modeling – 1980," Jont B. Allen (1981)

19.1 The CARFAC's AGC Loop

Allen (1979, 1981) was among the first to describe cochlear mechanics (as opposed to neural response) as having an automatic gain control (AGC) functionality, based on modeling Rhode's observations of nonlinear mechanical response (Rhode, 1971). Kim (1984) gets much more explicit, defining the roles of the inner- and outer-hair-cell subsystems, and of the medial olivo-cochlear (MOC) efferents in the cochlea's integrated nonlinear system:

> The function of the large medial OC neurons is to exert a gain control upon the biomechanics of the organ of Corti by reducing the amount of mechanical energy released from the OHCs via a synaptically mediated regulation of the membrane potential and conductance of the OHCs.

Unlike Allen's concept described in the chapter quote above, my original *coupled AGC* (Lyon, 1982) was not "built into the cochlear filters," but it was otherwise not far from the concepts of Allen and Kim and our present models. That is, it was a multiplicative coupled multichannel gain control that followed a linear filterbank, rather than achieving gain variation by varying the filter damping factors.

In this chapter, we describe how the CARFAC models the MOC feedback, and more local feedback, as an AGC loop filter feeding the modeled IHCs (detectors) back to OHCs (gain effectors) described in the previous two chapters.

Compared to the simple AGC model developed in Chapter 11, the AGC loop filter in the CARFAC has several aspects that make it interesting. First, it has four one-pole smoothing filters with their outputs combined, to span a range of time constants instead of having a single characteristic time constant or corner frequency. Second, each of these four filter stages has *coupling between neighboring channels*—and, in a binaural or multimicrophone version, also between the two or more ears. Third, to save computation cost, the loop-filter state is updated at a much lower sample rate than the rate used for the CAR filters themselves. We describe these aspects in turn.

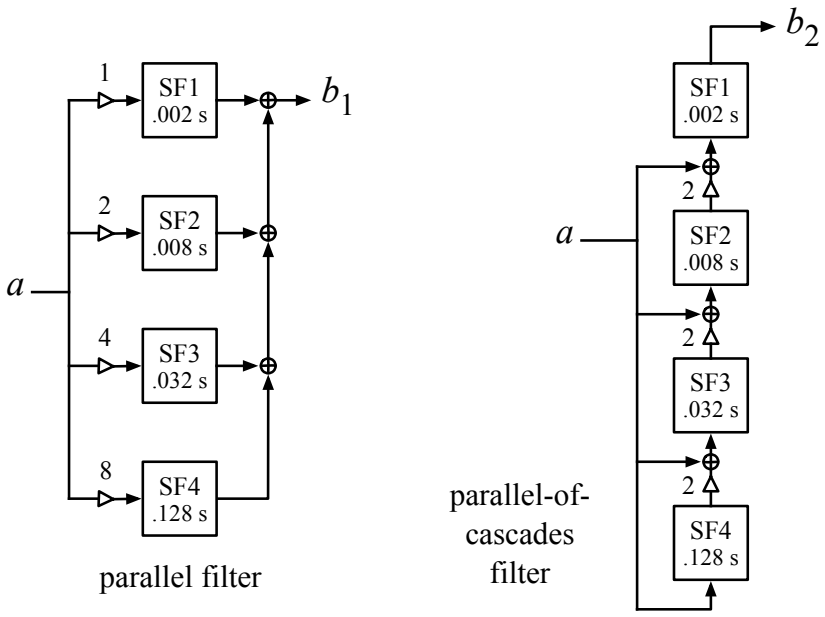


Figure 19.1: The filters in each AGC channel are conceptually made from four first-order lowpass smoothing filters with different gains and time constants, in parallel, as shown on the left, and as introduced in Figure 11.12. The individual smoothing filters (SFi) have transfer functions $1/(\tau_i s+1)$, with time constant $\tau_1 = 2$ ms, and increasing by factors of four up to $\tau_4 = 128$ ms. But we use the variant shown on the right: parallel combinations of cascades of these first-order filters, sharing pieces in cascade, because this arrangement makes it easier to run the filters with long time constants at lower sample rates.

19.2 AGC Filter Structure

The overall plan of the CARFAC, in Figure 15.2 hides most of the complexity, and shows the AGC loop filters as smoothing filters. In reality, they need not have unity gain at low frequency; rather, their low-frequency gain is adjusted to make the right interface between the range of levels expected out of the detection nonlinearity or inner-hair-cell model, and the range of damping factors desired in the filters. That factor will be adjusted in the end, to "close the loop."

The overall time–space filter network can be detailed in several different ways. We use four stages with different smoothing time constants, in parallel, as shown in Figure 19.1. We chose the method on the right because it most easily allows for an efficient scheme of decimation, running the slower stages at lower sample rates, as we discuss in Section 19.6.

19.3 Smoothing Filter Pole–Zero Analysis

The point of connecting smoothing filters of different time constants in parallel is to keep the gain-control loop stable, with good margins, by making a loop filter with a response that falls off gradually, not as steep as −6 dB/octave, and with considerably less than 90 degrees of phase shift. The gains of the four sections progress by factors of 2, while their time constants change by factors of 4; this puts the Bode-plot corners on a −3 dB/octave slope, as shown in Figure 19.2. The phase shift approaches 90 degrees only well beyond the highest corner frequency.

Consider what it means to arrange several filters in parallel; review Section 6.14. The transfer function of the four one-pole smoothing filters in parallel (as on the left in Figure 19.1) is found by adding the complex transfer functions of the one-pole filters:

$$H(s) = \frac{1}{\tau_1 s + 1} + \frac{2}{\tau_2 s + 1} + \frac{4}{\tau_3 s + 1} + \frac{8}{\tau_4 s + 1}$$

and, for the *parallel-of-cascades* filter (on the right in Figure 19.1), by adding the cascade transfer functions of varying order:

$$H(s) = \frac{1}{\tau_1 s + 1} + \frac{2}{(\tau_2 s + 1)(\tau_1 s + 1)} + \frac{4}{(\tau_3 s + 1)(\tau_2 s + 1)(\tau_1 s + 1)} + \frac{8}{(\tau_4 s + 1)(\tau_3 s + 1)(\tau_2 s + 1)(\tau_1 s + 1)}$$

In both cases, writing these over the common denominator defined by the four poles results in a third-order numerator whose roots are the zeros that are induced (here we simplify to one time-constant parameter τ_1, making others larger by factors of 4):

$$\begin{aligned} H(s) &= \frac{a_3 s^3 + a_2 s^2 + a_1 s + a_0}{(\tau_4 s + 1)(\tau_3 s + 1)(\tau_2 s + 1)(\tau_1 s + 1)} \\ &= \frac{a_3 s^3 + a_2 s^2 + a_1 s + a_0}{(64\tau_1 s + 1)(16\tau_1 s + 1)(4\tau_1 s + 1)(\tau_1 s + 1)} \\ &= \frac{a_3 s^3 + a_2 s^2 + a_1 s + a_0}{4096\tau_1^4 s^4 + 5440\tau_1^3 s^3 + 1428\tau_1^2 s^2 + 85\tau_1 s + 1} \end{aligned}$$

The coefficients a_i can be found from a little algebra, simplified by the stated progressions of time constants by factors of 4 and gains by factors of 2. For the simpler parallel case:

$$[a_3, a_2, a_1, a_0] = [7680\tau_1^3, 5520\tau_1^2, 690\tau_1, 15]$$

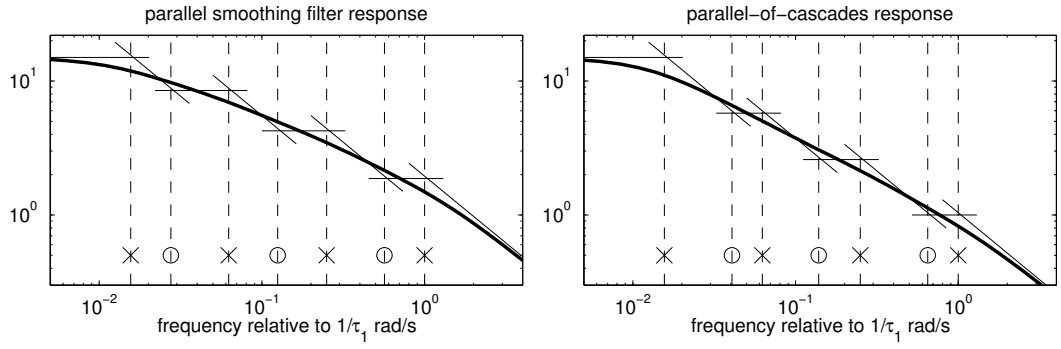


Figure 19.2: Since real poles and zeros induce a slope change of 6 dB per octave in Bode-plot asymptotes, we can use the calculated pole and zero frequencies to directly draw the skeletons of Bode plots for the four-pole smoothing filters. Each pole or zero corresponds to a corner between intersecting segments of slope 0 and –6; since the slopes are known, the frequencies are enough to easily construct the skeleton of the Bode plot as shown. The transfer function will be a smooth curve bounded alternately above and below by these corners, with slope between these asymptotic slopes. For the given time constants and gains, the zero frequencies and resulting slopes are slightly different between the parallel filter (left) and the parallel-of-cascades filter (right).

and for the parallel-of-cascades case:

$$[a_3, a_2, a_1, a_0] = [4096\tau_1^3, 3392\tau_1^2, 500\tau_1, 15]$$

The numbers are not so important, except to show that the results are not the same; but we can solve for the roots of those numerators and compare the locations of the zeros. In both cases, the result is a set of three real zeroes, interleaved between the four poles. The pole locations for both are

$$\left[\frac{-1}{\tau_1}, \frac{-1}{4\tau_1}, \frac{-1}{16\tau_1}, \frac{-1}{64\tau_1}\right]$$

while the zeros for the parallel structure are:

$$\left[\frac{-1}{1.76\tau_1}, \frac{-1}{8.00\tau_1}, \frac{-1}{36.23\tau_1}\right]$$

and for the parallel-of-cascades structure are:

$$\left[\frac{-1}{1.54\tau_1}, \frac{-1}{7.20\tau_1}, \frac{-1}{24.59\tau_1}\right]$$

Neither of these sets of zero positions is uniquely suitable, but the arrangement of interleaved real poles and zeros is a good way to make a filter with a moderate rolloff slope over a wide frequency range, as shown in Figure 19.2. The moderate slope corresponds to a moderate phase lag, well below 90 degrees.

Keeping the phase lag moderate (less than 90 degrees) until the gain magnitude drops enough is a good strategy to keep feedback control systems stable. With other delays in the loop, such as delay in the reaction of the CAR filterbank output to its damping-control input, and extra delay that we introduce by subsampling to run the AGC filters at a lower rate, the phase lag will eventually exceed 180 degrees at some high enough frequency. If that happens before the magnitude of the loop gain is less than 1, the control system will go unstable.

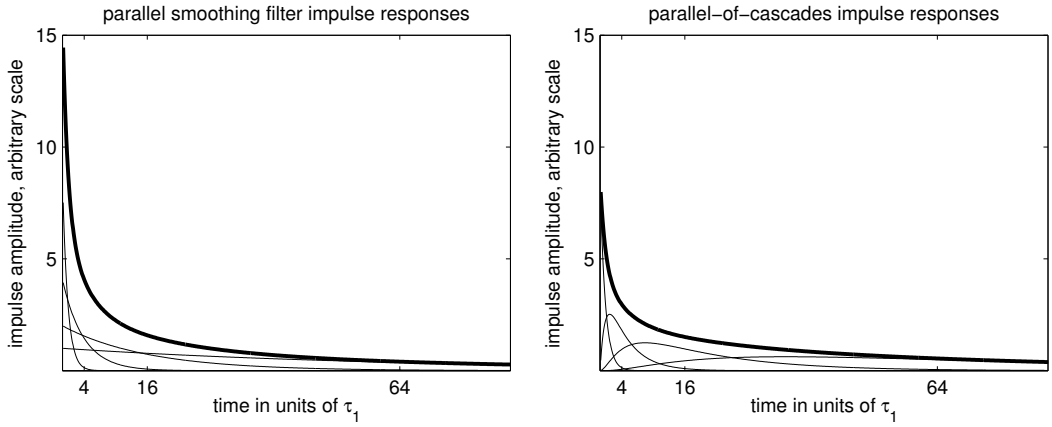


Figure 19.3: The impulse responses of the four-pole smoothing filters are easily found by adding up the impulse responses of their four paralleled parts, as shown here. On the left, the parts are the exponential decays of the first-order filters alone, while on the right they are the impulse responses of cascades of 1, 2, 3, or 4 first-order filters. Of course, since the two systems share the same poles, the total impulse response of either can be described as a weighted sum of those exponentials in the left plot, but with different weights. Therefore, the two structures can be completely equivalent if we generalize the gains; but for the default CARFAC we choose the parallel-of-cascades with gain factors of two.

Stability concerns also motivate the use of an AGC loop design with only a moderate "speedup factor" as introduced in Chapter 10. The speedup factor there is a reflection of a linearized loop gain that increases with signal level. The control loop might go unstable if a loud signal increases the loop gain too much, when there is delay in the loop. If the speedup factor is not bounded, then no amount of additional loop delay (delay that makes the net phase shift greater than the 90 degrees due to the AGC smoothing filter itself) can be tolerated. In the CARFAC, the IHC's saturating detection nonlinearity imposes a bound on the detected output level that goes into the loop filter, which helps keep the loop stable even with the inevitable extra delays of the CAR response and the smoothing filter implementation.

19.4 AGC Filter Temporal Response

The shape of the smoothing filter's temporal impulse response, shown in Figure 19.3, is in some ways like the shapes encountered in characterizing auditory responses in the nervous system, in that it has no clear time scale. This behavior is similar to processes with $1/f$ power spectrum (-3 dB/octave), or self-similarity, which have no characteristic time constant (Hausdorff and Peng, 1996). Phenomena such as this are sometimes analyzed into sums of exponentials, under the assumptions that different time constants come from different mechanisms, and that knowing the time constants will help to find or understand the mechanisms. But in real data, this analysis into exponentials is an ill-posed, or numerically difficult problem, especially if there are more than two time constants or if the system is nonlinear. In hearing, the notions of *rapid* and *short-term* and *long-term* adaptation are often invoked to describe such multi-timescale behaviors. Our time constants 2 and 8 ms are in the "rapid" range, while 32 and 128 ms are in the "short-term" range.

The "rapid" adaptation of outer hair cell activity is probably too fast to be efferent mediated. In recent years, researchers have found mechanisms local to the organ of Corti that may mediate such an AGC. Thiers et al. (2008) say, "The results suggest that a complex local neuronal circuitry in the OHC area, formed by the dendrites of type-II neurons and modulated by the olivocochlear system, may be a fundamental property of

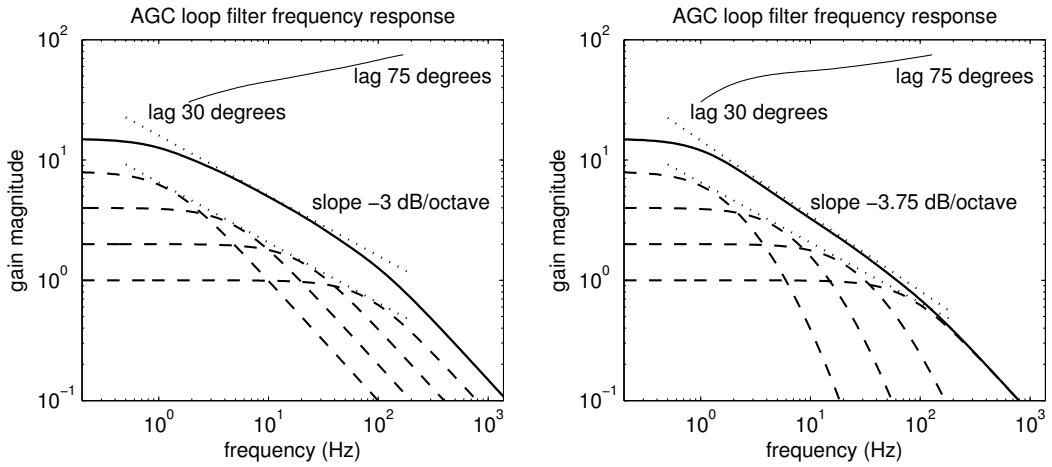


Figure 19.4: The frequency responses of the AGC smoothing filters (solid curves) can be found directly by adding up the complex gains of the four paralleled filters (dashed curves show their magnitudes), rather than via a pole-zero analysis leading to a Bode plot as in Figure 19.2. The purely parallel interconnection of first-order filters (the leftmost filter in Figure 19.1) gives the response on the left, while the parallel-of-cascades variants give the response on the right. The phase lag stays within a moderate 30 to 75 degrees over more than two orders of magnitude of frequency (upper curve, degrees of phase lag, on log scale). The results match Figure 19.2 using $\tau_1 = 0.002$ s.

the mammalian cochlea, rather than a curiosity of the primate ear. This network may mediate local feedback control of, and bidirectional communication among, OHCs throughout the cochlear spiral."

We could also extend the AGC to have some "long-term" adaptation, by adding more AGC filter stages with 0.5 s and 2.0 s time constants. Further time constants, out to "very-long-term" in the range of minutes are sometimes discussed. These longer time scales of adaptation may be more appropriate to a later stage of processing. As we learn more about what mechanisms have what time constants, we can trade off this feedback that affects the traveling wave with other adaptation mechanisms in the hair cells and auditory neurons. The confounded effects of different mechanisms are often not easy to distinguish in experimental data.

Numerically summing is another easy way to compute the frequency responses, as an alternative to the pole-zero analysis and evaluation. Summing (as complex numbers) the four all-pole transfer functions, at each frequency, results in the magnitude and phase Bode plots shown in Figure 19.4. In both cases, the smoothing filters maintain a moderate slope and phase lag well below 90 degrees from 1 Hz to 100 Hz, a range that encompasses the level fluctuations that the AGC will tend to react to.

The filter structure used in the AGC is further detailed in Figure 19.5, showing spatial coupling between channels at each stage, but still without showing exactly how the coupling is accomplished. The time-domain or frequency-domain aspect of the loop filters can be considered independently of the spatial smoothing. We make the cross coupling diffusion-like, such that it only has an effect where there is a spatial gradient; thus, it has no effect when all nearby channels have equal state variables, in which case the filters behave as the pure time-domain filters analyzed above.

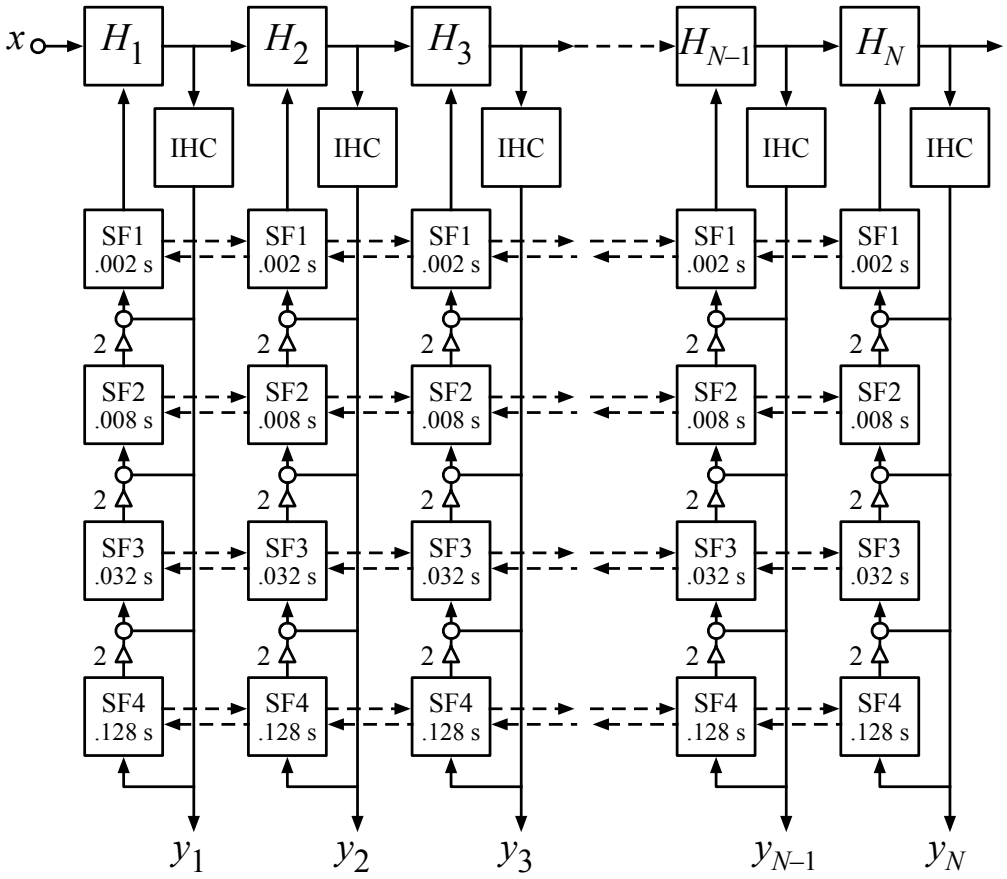


Figure 19.5: The filters in each AGC channel are based on the parallel-of-cascades version in Figure 19.1. In this configuration, the faster filters define a shorter (more local) loop, and it is easier to run the more-remote slower filters at lower sample rates, since their output steps will be smoothed by other filters before being applied to control the CAR filter stage damping. Neglecting the sampling approximations, the loop-filter transfer functions are as shown on the right in Figure 19.4. The lateral interconnections are shown dashed; the mechanism for spatial cross-coupling is detailed in the next figure.

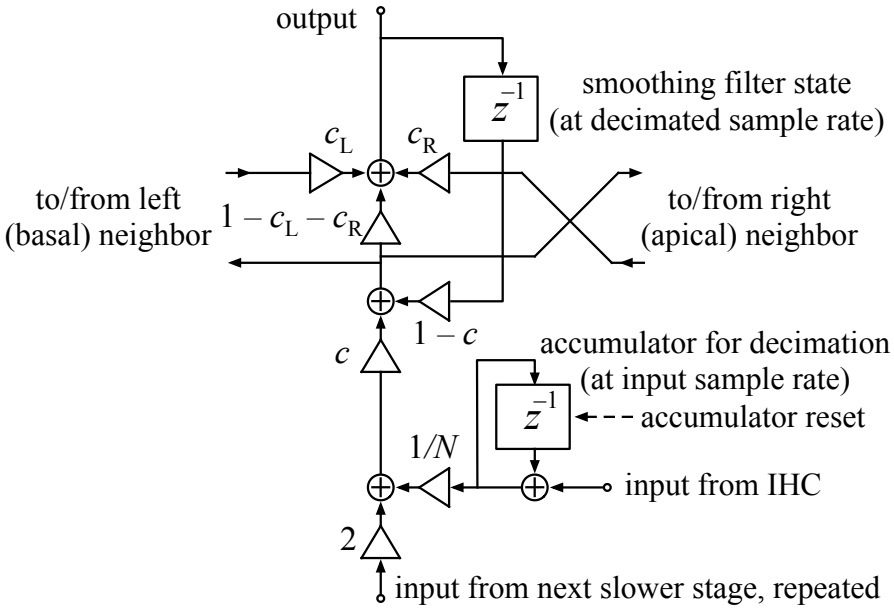


Figure 19.6: The unit smoothing filter (one stage, one channel) of the coupled AGC filter is drawn here in a bottom-to-top flow arrangement, as it is used in Figure 19.5. At the bottom right, input values at a high sample rate (from the IHC, at the CAR's audio sample rate) are accumulated until it is time for the unit to operate. After accumulation of N samples (N being the decimation factor), the accumulator value is divided by N and used as input to the smoothing filter, and the accumulator is reset. The input from the next slower stage, if there is one, is also added in, with a weight of 2; if that stage has a higher decimation factor, each of its output samples may be used as input more than once. The c coefficient controls the time-domain smoothing time constant. In the spatial smoothing part, the 3-point FIR filter $[c_L, (1 - c_L - c_R), c_R]$ applies weight c_L to the value from the left neighbor, c_R to the value from the right neighbor, and enough gain to the current channel to keep the total mixing gain equal to 1.

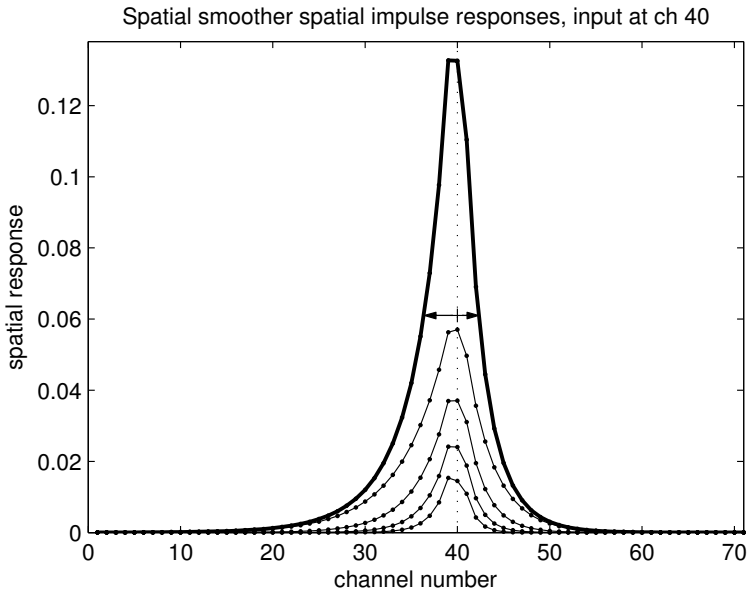


Figure 19.7: Spatial impulse responses of the AGC loop filter's spatial smoothing filters (four lower curves, which include their respective power-of-2 weights), and their sum (heavy upper curve), with an input at channel 40 only. The filters are designed with a moderate asymmetry of spread toward earlier channels (toward the base, or higher-CF end, of the cochlea) as indicated by arrows. The filters are 3-point FIR filters, applied to the AGC lowpass filter state arrays at each AGC update time, and effectively iterated many times by the long time constants of the temporal smoothing. For this figure, and the next one, the parallel paths were kept separate so that we could see the response of each part. For the fastest and most local stage, the coefficients used for this response are: $c_{\mathrm{L}} = 0.286$, $c_{\mathrm{R}} = 0.404$, $c = 0.166$. This c value represents a time constant of about 6 samples at the decimated rate, so the 3-point FIR smoothing is effectively applied about 6 times per time constant, resulting in the Gaussian-like spread seen on the lowest curve. Slower stages have more time to spread further.

19.5 AGC Filter Spatial Response

Smoothing across places, or *spatial smoothing*, is a key part of the coupled AGC concept. By allowing the response of the filter channels to reduce the gain to other nearby channels, the coupling helps to keep the gains of nearby channels close to each other. Locally similar gains preserve local spectral contrasts, while reducing dynamic range. Without the spatial coupling, spectral contrasts would also be compressed, or partially normalized away. In sensory systems more generally, this kind of spatially distributed feedback gain control is known as *lateral inhibition* (Békésy, 1967). Such inhibition has been described in the mammalian auditory nerve and cochlear nucleus (Rhode and Greenberg, 1994), and has been used as a model of processing of neural signals from the cochlea for enhancement of the representation of speech spectra (Shamma, 1985).

In the CARFAC AGC, the spatial coupling is implemented by a linear spatial filtering process that operates across the spatial samples within each of the four time-domain smoothing filters, replacing the filter states by spatially smoothed filter states, using the structure detailed in Figure 19.6. That is, this structure integrates a 3-point spatial FIR smoothing with the first-order IIR time smoothing in one unit of the AGC loop filter. The effective spatial impulse responses, computed by providing a steady input on a single channel while holding the others at zero, are as shown in Figure 19.7.

19.6 Time–Space Smoothing with Decimation

The AGC changes the filter parameters relatively slowly compared to most sound frequencies, so the AGC loop can be operated at a lower sample rate—*decimated*—as long as care is taken to keep it working smoothly and without much extra delay. In our default implementation, with the cochlear filters running at 22 050 Hz, we update the fastest AGC stage only every 8th sample period (about 2750 Hz sample rate, with its time constant of 2 ms), and subsequent AGC stages less often, by factors of 2. Based on the time constants growing by factors of 4, we could decimate even more, but that would make it harder to achieve the desired increases in spatial spreading with the 3-point FIR smoother.

High-rate inputs are averaged over the stage's sampling period to make an input sample, as shown in Figure 19.6; outputs from slow stages are simply repeated, adding to those new input samples; the resulting steps at update times are subsequently smoothed by going through the higher-rate smoothing stages. The final output drives an interpolation process that connects it as feedback to the DOHC, spreading each step change over 8 samples. The net result is that the AGC adjusts the CAR poles smoothly and quickly, with a low computational cost and fairly simple code.

At higher frequencies, where the smoothing filter response has decreased by more than 20 dB (see Figure 19.4), the smoothing filter rolloff slope increases to −6 dB/octave and its lag approaches 90 degrees; decimation by 8 adds a further delay of about 8 samples for the input averaging and output interpolation, or 90 more degrees at 1/32 of the sample rate (that is, at 687 Hz, where the smoothing filter is down by more than 35 dB in parallel-of-cascades version). If the original sample rate is 24 kHz, 8 samples is 0.333 ms, about 1/6 of the fastest stage time constant of 2 ms. The AGC loop is robustly stable with these parameters, since the loop filter's frequency response magnitude is quite low before the phase shift around the loop approaches 180 degrees.

We have implemented several approximately equivalent spatial smoothing methods: 3-point and 5-point FIR (nonrecursive) filters, and a two-pass (forward and backward) one-pole-each-way IIR smoothing filter. These are adjusted to give equal amounts of "spread" and "shift," which are design parameters that control how far the gain control spreads across channels, and with what asymmetry. These parameters correspond to the standard deviation and mean of the spatial impulse response when viewing it as a probability distribution. When the smoothing is effectively applied N_τ times (the effective number of times being the number of actual spatial smoothing applications per time constant of the temporal smoothing filter), the variance (squared

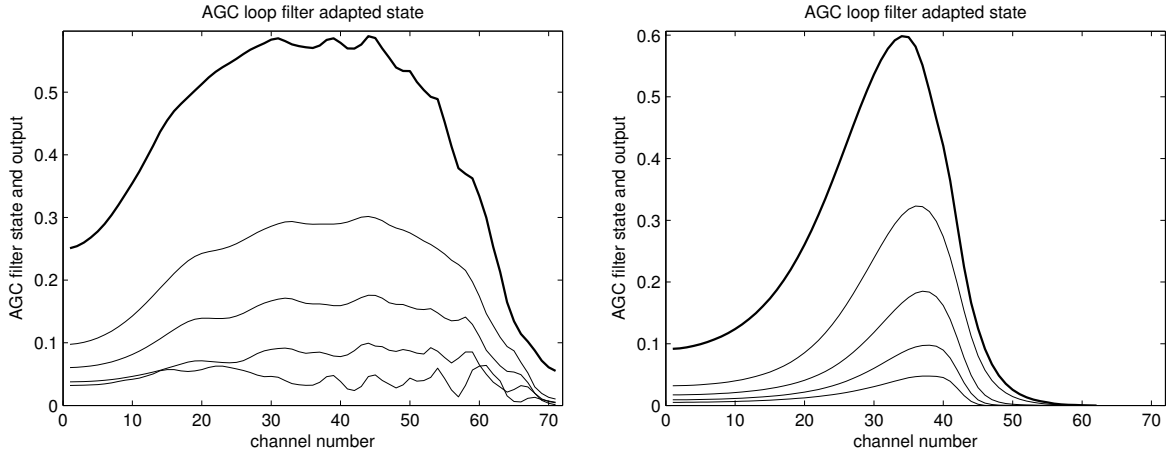


Figure 19.8: Typical states of four AGC parallel smoothing filters (from the left filter of Figure 19.1, lower curves), and their sum (the y_1 signal, upper heavy curves), for a speech sound (left), and a 1 kHz tone (right). We illustrate states of the parallel filter here, rather than the parallel-of-cascades form, so that each curve represents a time scale and its weight; the actual states in our parallel-of-cascades form of Figure 19.5 are roughly scaled cumulative sums of these, with a similar final result. The spatial smoothing is least on the lowest curve, the state of the fastest filter, and greatest on the state of the slowest filter. The spatial smoothing has been designed to be somewhat asymmetric, spreading more to earlier (more basal, higher-CF) channels than to later (more apical, lower-CF) channels, modeling the spread of MOC efferents toward more basal locations (see Figure 14.15). In particular, the place most responsive to the 1 kHz tone is channel 37, but the strongest AGC filter feedback comes at channel 34.

spread) and shift of the distribution are multiplied by N_τ, since convolution of impulse responses is analogous to the addition of i.i.d. random variables.

For some design parameters, a 3-point FIR smoothing filter, operating at some number of time samples per time constant, may not be able to produce the desired amount of spatial spread and shift. One option in such cases is to run the 3-point FIR step several times per AGC sample time, to get more spread; another is to use a 5-point FIR filter; yet another is to run an IIR filter to get more spread. Our AGC filter code switches to the 5-point FIR filter if it would take more than one iteration for the 3-point filter to give the specified amount of spread; if that is still not enough spread at one iteration per sample, then the IIR filter is used. In the default case, we run the filters at high enough sample rates that one step of 3-point FIR per sample time, as illustrated in Figure 19.6, does the job.

19.7 Adapted Behavior

When the AGC state adapts to a sound, the CAR filterbank changes its transfer functions to change how much gain it applies to the sound. Figure 19.8 show typical adapted states in reaction to a speech sound and to a tone, and Figure 19.9 shows the corresponding adapted transfer functions.

19.8 Binaural or Multi-Ear Operation

The coupled AGC has one more twist: coupling between the ears. At each time step, the states of AGC smoothing filters serving each ear are pulled toward each other, reducing the differences between them. This

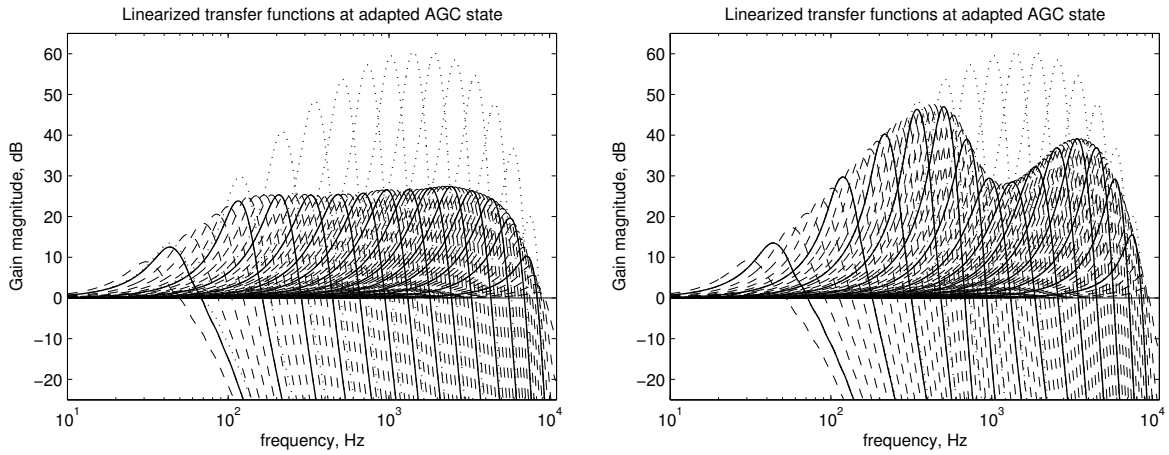


Figure 19.9: The linearized transfer functions of the CARFAC at the adapted AGC states shown in Figure 19.8. The gains of the middle high-gain channels have come down by more than 30 dB in reaction to the speech signal (left), compared to the gain in quiet (selected channels shown as upper dotted curves). The gain reduction is more localized to near the 1–2 kHz region when adapted to a 1 kHz tone (right).

cross-ear coupling models the physiology, in which most of the MOC efferents are driven by the ipsilateral ear (by the cochlea on the same side that they are controlling), but a significant minority, perhaps up to one-third, are driven by the contralateral ear. The effect seems to be mostly from the "uncrossed" olivocochlear efferents, which means that the information from the contralateral ear crosses the midline in its afferent pathway (Warren and Liberman, 1989; Guinan, 2010).

The effect of interaural coupling is to compress levels more than interaural level differences. That is, the gains of the two ears will somewhat track each other, so that when a sound in one ear is louder than the sound in the other, that loudness difference will be preserved, to a greater extent than the absolute loudness is. Thus the cross-ear coupling, another form of lateral inhibition, helps to preserve contrast between the levels in the two ears, to retain this cue for spatialization of sound sources. By keeping the gains close together in the two ears, the coupling will also help to minimize any differences in impulse-response timing (zero-crossing times), which will therefore help to preserve the binaural time-difference cue as well.

Each stage of the 4-stage smoothing filter does a cross-channel linear mix each time it updates, each with a mixing coefficient adjusted to its sample rate and time constant, controlled by a mixing specification that says what fraction of the output should come from the mean, as opposed to from the ipsilateral ear. Basing the coupling on the mean allows an easy generalization to systems with more than two "ears." We typically use zero coupling on the fastest stage, to save time, and because it might represent a more local mechanism in the cochlea as opposed to an MOC effect. For the other three stages we use 0.5 of the mean, which effectively gives 25% contralateral and 75% ipsilateral effect in a binaural system. It might be more realistic to cross-couple only the slowest stage of the AGC.

19.9 Coupled and Multistage AGC in CARFAC and Other Systems

The coupled AGC combines many physiological effects into a smoothing filter of only moderate complexity, with coupling between channels and between ears that helps the cochlea to preserve spectral contrasts and interaural cues while largely suppressing absolute level effects. It runs efficiently as part of the CARFAC digital system, because it can be operated at reduced sample rates.

The response of the CARFAC involves this AGC filter in the loop with the CAR, OHC, and IHC, and

produces outputs such as the average-rate cochleagram or NAP of Figure 19.10 (compare with linear-scale spectrogram in Figure 5.9 and mel-scale spectrogram in Figure 5.10). A zoomed-in NAP with fine temporal structure has been shown in Figure 18.8. The overall level-compression effect has been illustrated in Figure 17.3.

Similar multi-timescale gain-control models have been developed to model nonlinear visual adaptation. For example Hateren and Snippe (2001) say:

> The best model performs close to this maximum performance, and consists of a cascade of two divisive feedback loops followed by a static nonlinearity. The first feedback loop is fast, effectively compressing fast and large transients in the stimulus. The second feedback loop also contains slow components, and is responsible for slow adaptation in the photoreceptor in response to large steps in intensity. Any remaining peaks that would drive the photoreceptor out of its dynamic range are handled by the final compressive nonlinearity.

Such biological gain-control systems are a bit more elaborate than what engineers typically design, but their multiple mechanisms and multiple time constants do contribute to robust system behavior in the real world of high dynamic range.

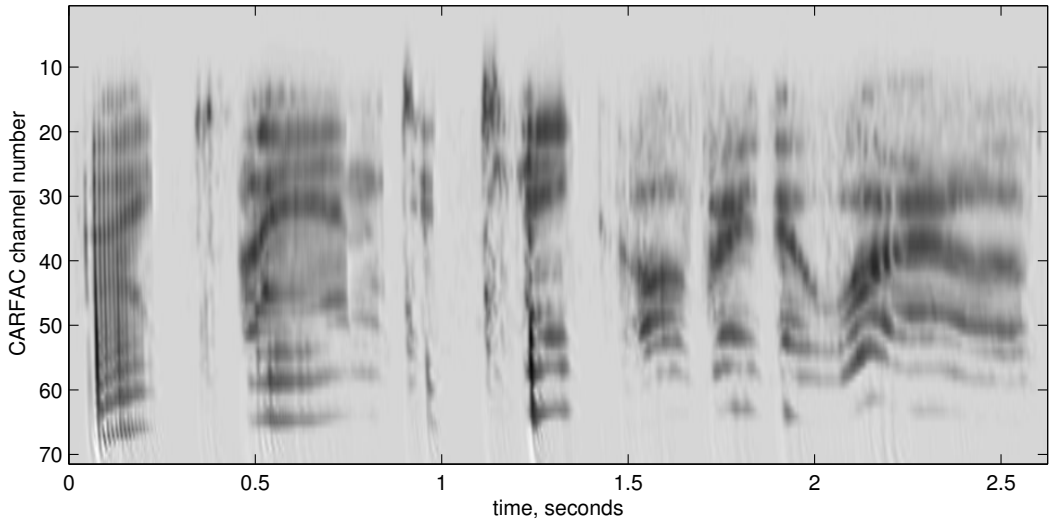


Figure 19.10: The average-rate (time-smoothed) neural activity pattern (NAP) of a few seconds of speech shows how the CAR, OHC, IHC, and AGC work together to make a clear alternative to a spectrogram. We sometimes subtract off the rest response level and clip to white to remove the gray background.

Part IV

The Auditory Nervous System

Part IV Dedication: J. C. R. Licklider

This part is dedicated to the memory of Joseph Carl Robnett Licklider (1915–1990). "Lick" is best known as one of the "fathers of the Internet" (Poole et al., 2005), based on his ARPA leadership and his writings such as "Man–Computer Symbiosis" and "The Computer as a Communication Device."

But before he was a computer network and systems guy, Lick was an auditory psychologist and modeler (November, 2012). His work on pitch perception, as represented in the "duplex theory," is the basis for much recent work in hearing, including my own, connecting the output of the cochlea to perception and neural processing of complex sounds.

I had the pleasure of meeting Lick just once, in 1984 at a Navy-sponsored workshop on "Artificial Intelligence and Bionics." I think he was a little surprised to see his duplex theory coming back as a practical computational approach, three decades after he came up with it. It has become even more practical since then, thanks partly to his computer innovations.

In this part, we discuss the levels of processing in the auditory nervous system. We develop the idea of auditory images, of the sort that are thought to be extracted by brainstem and midbrain for projection to auditory cortex.

We start where the last part left off, with the "cable" for the telephone theory of hearing, the auditory nerve, which transmits the vibrations as detected by hair cells in the cochlea to the first stop in the brainstem, the cochlear nucleus.

Several kinds of processing in the cochlear nucleus support both binaural hearing and the extraction of properties such as pitch and timbre that can be monoaural or binaural. We cover the extraction of such properties into the *stabilized auditory image*, a basis for sound representation in machine hearing systems as well as a model of representations in the inferior colliculus of the midbrain. We cover binaural spatial processing and the brainstem's olivary complex. We finish this part with a discussion of *auditory scene analysis* as the main aim of the auditory brain, and some ideas for how such analysis might be done in the thalamus and cortex to finally "extract meaning."

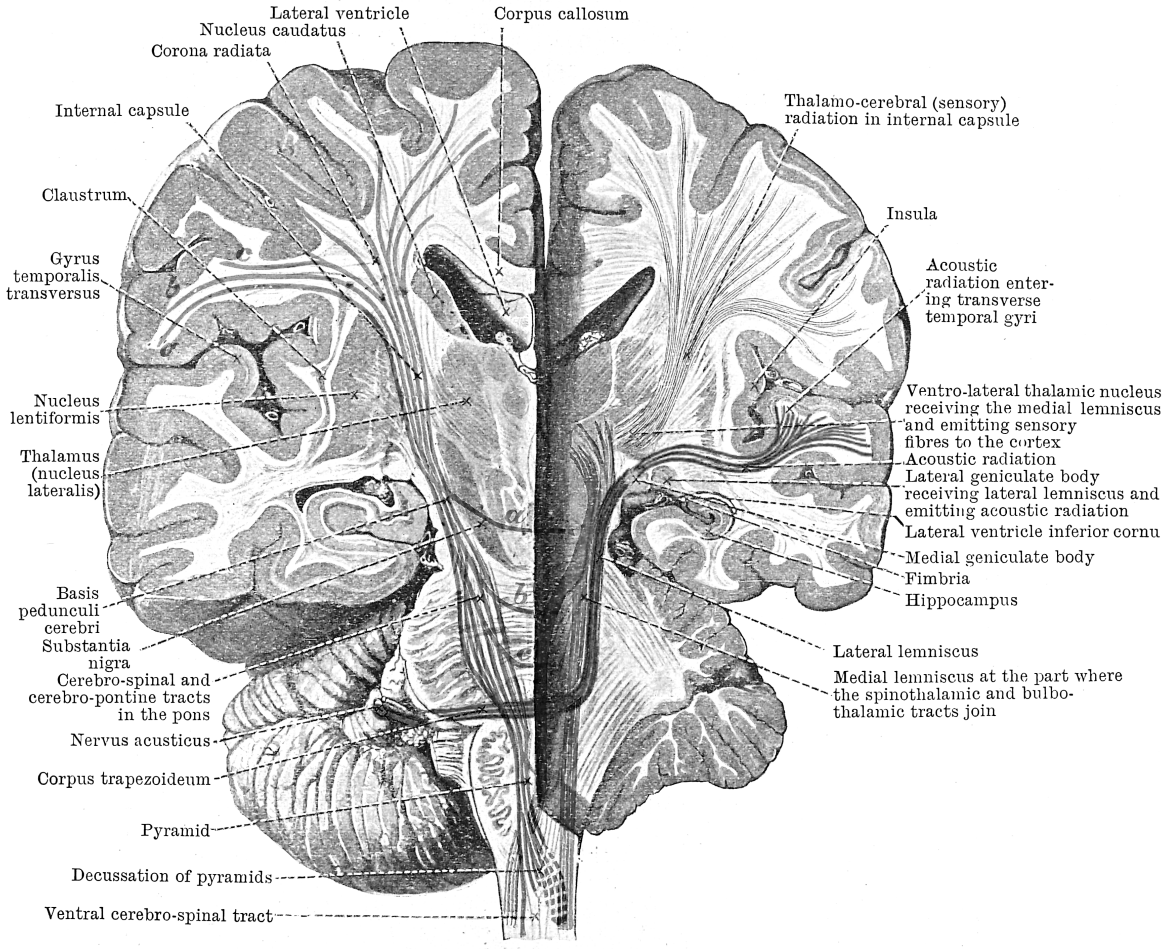


FIG. 506.—A VERTICAL TRANSVERSE SECTION OF THE BRAIN TO SHOW THE WHOLE OF THE CENTRAL ACOUSTIC PATH. The left hemisphere (right side of the figure) is cut on a plane posterior to that of the right. Motor fibres red. Sensory fibres blue. Acoustic fibres yellow.

The auditory nervous system was already fairly well mapped out a hundred years ago, as this color illustration from *Cunningham's Text-Book of Anatomy* shows (Cunningham and Robinson, 1918). Auditory fibers are dark gray here, yellow in the color plate.

Chapter 20

Auditory Nerve and Cochlear Nucleus

> I experimented in this way, and eventually found that I could send as many as 352 impulses per second along the nerve of a rabbit and get a note from the muscle of the pitch of 352 vibrations per second ... but when I tried by more rapid stimulation of the nerve to get a higher note from the muscle, I failed. ... Now, am I to conclude that, because I failed to get a higher note than one of 352 vibrations from the muscle, it is not possible to send more than 352 vibrations per second along a nerve? By no means ...
>
> — "A lecture on the sense of hearing," Rutherford (1887)

The auditory nerve (AN), originating in the spiral ganglion in the cochlea, carries the output signals of the cochlea's IHCs into the same-side (*ipsilateral*) cochlear nucleus (CN) in the brainstem, just a few centimeters away, as shown in Figure 20.1. To a large extent, the inputs and outputs of the CN tell us what information the brain is getting from the ear, and what the important first steps are in processing that information.

Observations of physiological behavior of the CN support our emphasis on the use of fine temporal structure in sound representation, as in the Fletcher (1930) "space–time pattern theory" and the Wever and Bray (1930b) "volley theory" that we reviewed in Chapter 2. The CN's inputs and outputs are events, synchronized to sound waveform structure in a very precise and robust way in some parts of the CN, but less synchronized in other parts, serving different subsequent processing pathways.

In the CN, pathways diverge in support of a number of processing functions in parallel, variously specialized for binaural processing, for periodicity detection, and for other monaural feature extraction. The *tonotopic organization* of the auditory nerve is maintained as a spatial dimension, and various subsequent brain areas use another spatial dimension to organize the extracted features into what we call auditory images, of various sorts.

20.1 From Hair Cells to Nerve Firings

That the AN fibers fire in synchrony to sound waveforms has been known since Wever and Bray (1930b) demonstrated cat ANs producing electrical signals that they could pick up with an electrode and amplify to reproduce intelligible speech waveforms. Synchrony of AN firings in response to tones, clicks, noises, speech, etc., has been widely studied and analyzed over the intervening decades (Galambos and Davis, 1943; Kiang, 1965; Rose et al., 1967; Young and Sachs, 1979; Palmer and Russell, 1986; Evans, 1989).

The *afferent* (toward the brain) signals on the AN can be interpreted as stochastic point processes (random events) approximately representing the continuous-time and continuous-amplitude outputs of the inner hair cells (IHCs), which in turn reflect the rectified waves on the basilar membrane. In any given short time

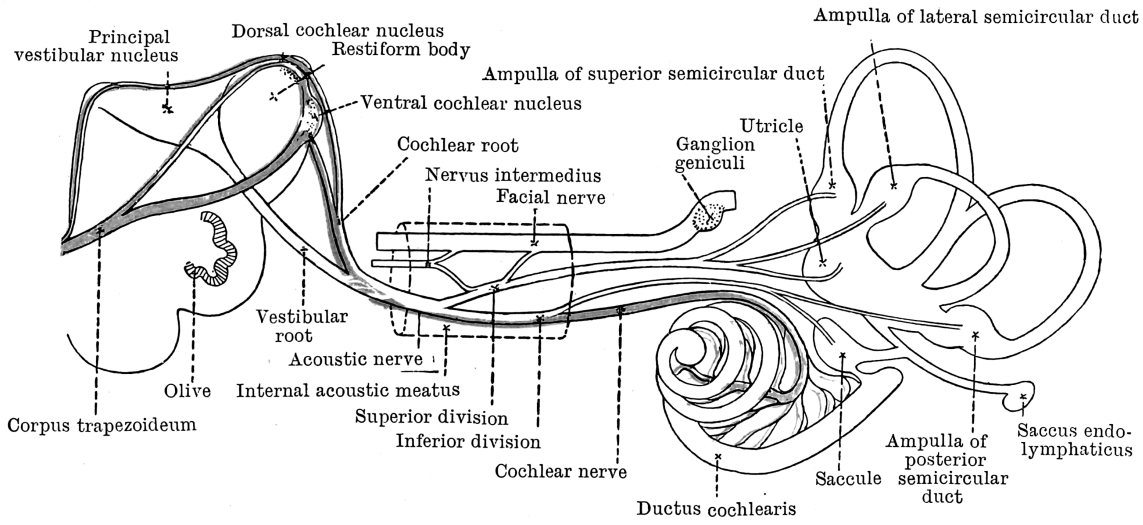


FIG. 587.—SCHEME OF THE ORIGIN AND DISTRIBUTION OF THE ACOUSTIC NERVE.

Figure 20.1: As shown in this color illustration from Cunningham and Robinson (1918), the acoustic nerve, or eighth cranial nerve, includes the cochlear division (dark gray here, blue in the color plate) that serves hearing, and the vestibular division (red in the color plate) that serves balance functions. After a stop at the dorsal and ventral divisions of the cochlear nucleus, the auditory pathway branches into the three acoustic stria, one of which, the ventral acoustic stria (the lower one here) goes to the superior olive on both sides, crossing via the trapezoid body. The facial nerve (yellow in the color plate) takes efferent signals back to the stapedius muscle in the inner ear, via the geniculate ganglion, to serve the protective acoustic reflex.

interval, the probability of an action potential, or firing, on an auditory nerve fiber is a function of the IHC output—a neurotransmitter amount that is itself a function of the mechanical response of the cochlea. These functions are complicated by the fact that they are also functions of the recent history, or state, of the hair cells, the nerves, and the synapses between them.

In our machine hearing systems, we tend to abstract away the firings, and use (nearly) continuous-valued but discrete-time signals to represent the outputs of the hair cells and primary auditory fibers. We give up detailed modeling of the auditory nerve, but gain the ability to use standard digital signal processing methods. On the other hand, to understand what functions we need to perform, it is very important to understand the variety of data from the real auditory nerve.

Other approaches to machine hearing do keep discrete events to represent analyzed sounds. For example, the AER-EAR system of Liu et al. (2010) is a hardware VLSI hearing system that uses an address-event representation (AER) to communicate and process analyzed sound as simple events.

Both the primary AN signals and later levels of nerve firings in the auditory nervous system are traditionally studied via firing-rate and synchrony measures. Firing rate is fairly simple, complicated only by what interval it is estimated over, and by what space of stimuli it is measured on. Synchrony is summarized by a variety of techniques, such as *post-stimulus-time histograms*, *period histograms*, and *interval histograms*, which display the relationships of nerve firing times or firing intervals to stimulus times or parameters. Figure 20.2 shows post-stimulus-time histograms in response to clicks; each histogram shows the times of an AN fiber's spikes relative to the time of the stimulus click sound, accumulated across many repetitions of the stimulus. For periodic stimuli such as simple or complex tones, period histograms are used to show spike times relative to the stimulus period; see Figure 4.8 for period histograms of several AN fibers to a vowel sound. Interval histograms are used to summarize the times between spikes, without reference to a stimulus. First-order

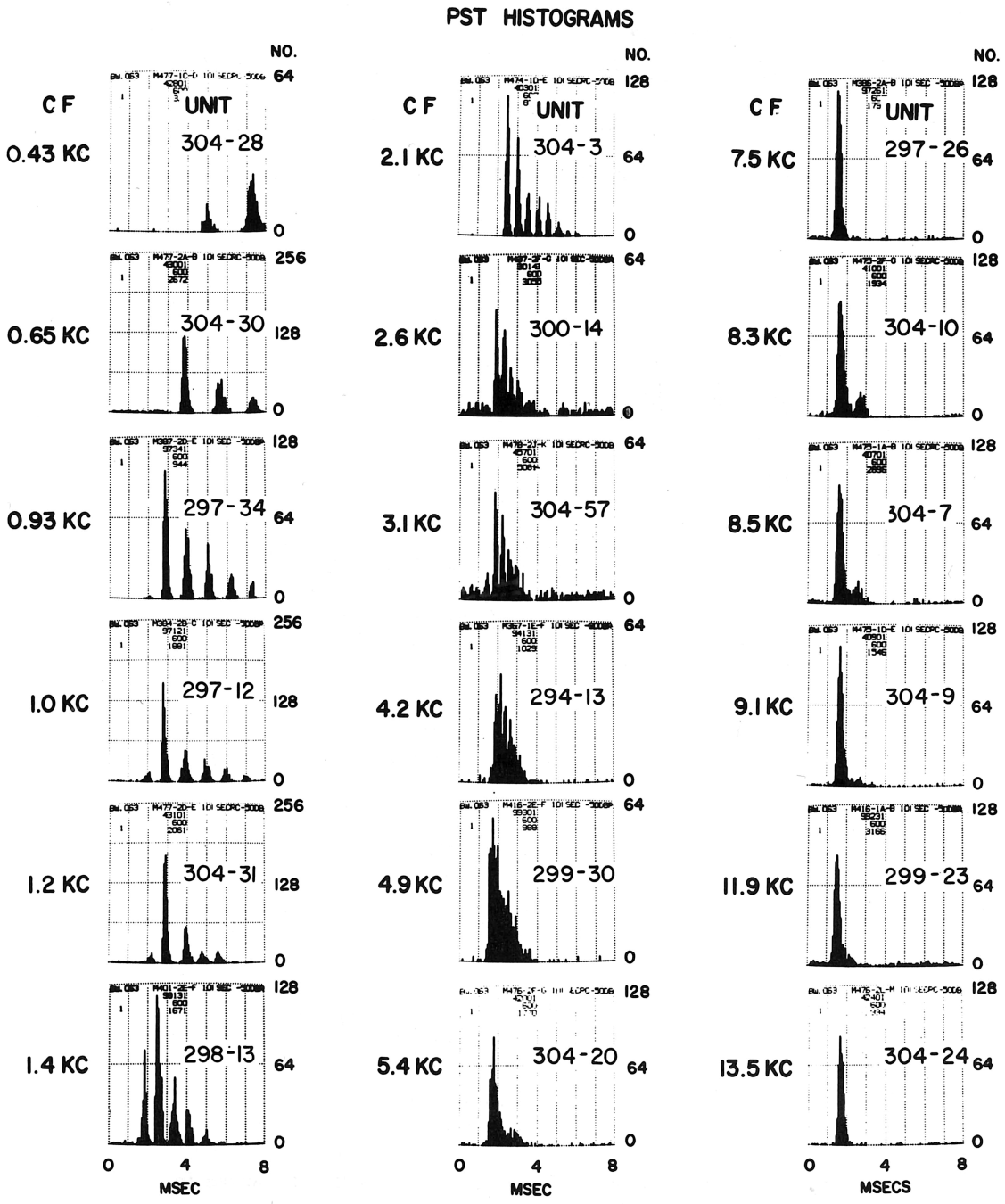


Figure 20.2: Kiang (1965) recorded times of action potentials on cat auditory nerve fibers, in response to brief clicks presented 10 per second, and summarized those firing times as post-stimulus-time (PST) histograms. Each histogram is labeled with the CF of the nerve fiber (the "unit") in KC, which is 1960s terminology for kHz (the unit numbers in each plot represent the animal number and the particular neuron). Notice that the fine time structure of the ringing of cochlear bandpass filters is reflected in the PST histograms for units with CF up to about 4 kHz. With plots sorted by cochlear place, or CF, as here, the graded latency to first response is also apparent, at approximately 1 ms plus 2 cycles of CF. [Figure 4.2 (Kiang, 1965) reproduced by permission of The MIT Press.]

interval histograms summarize the time from each spike to the next; all-order histograms (Delgutte and Cariani, 1992; Cariani and Delgutte, 1996a) also show more distant intervals, and are essentially autocorrelation functions of the spike trains—like rows of auditory images (see Chapter 21).

Many aspects of sound, such as pitch and direction, have been found to be much more robustly represented in time patterns than in rate-versus-place patterns.

20.2 Tonotopic Organization

The auditory periphery spreads out sound by frequency (or time scale) and delivers an array of waveforms to the brain via the auditory nerve. An amazing feat in the development of the auditory nerve and subsequent parts of the auditory nervous system is the maintenance of spatial order—known as tonotopy—in the auditory nerve and in subsequent brain areas. Somehow, the fibers of neurons that start together in the spiral ganglion stay together, near their neighbors of origin. One mechanism in the early development that helps to control this organization is the spontaneous spiking of inner hair cells. Mature inner hair cells do not generate spikes, though they are specialized neurons. But during early development, spikes that originate at an IHC are carried by all the AN primaries that innervate that IHC, and serve as markers to keep those primary fibers together (Kros, 2007). Exactly how the neighbors stay together to preserve a tonotopic organization is less clear, but it may rely on a changing burstiness of spike patterns along the basal–apical axis (Johnson et al., 2011).

Each cochlear place sends fibers with different spontaneous firing rates. High-spontaneous fibers fire at low sound levels, and even in quiet, providing a substrate of pulses that can convey the presence of a signal by modulation of their timing even at levels too low to increase their rate. But these fibers are in rate saturation at moderate and high sound levels, much as the rods in the eye, which mediate night vision, are saturated in daylight. The low-spontaneous fibers, like the cones in vision, take over at higher levels. At very high sound levels, they too are mostly saturated, leaving little or no rate-versus-place representation of sound spectrum.

Primary fibers bifurcate into the ventral and dorsal divisions of the CN. As pathways divide here and in subsequent stages, multiple tonotopically organized areas emerge.

20.3 Fine Time Structure in Cochleagrams

Though the cochlear filter bandwidths may be only tens to hundreds of hertz, each of the frequency channels in the AN can represent waveform detail, using volleys (groups of firings on multiple nerves, close together in time), up to several kilohertz. This relatively high neural bandwidth allows for a good representation of the half-wave rectified waveforms in low-CF channels, and a good representation of fast modulation in high-CF channels, and allows a timing precision sufficient to support binaural directional distinctions based on interaural time differences of a few tens of microseconds. Such time-precise or high-bandwidth waveform information is referred to as *fine time structure* or *fine temporal structure*, as we have already encountered.

A graphical representation of the information sent from the cochlea on the auditory nerve is called a *cochleagram*. Typically organized much like the time–frequency plane of a spectrogram, the cochleagram's rows each represent the time-domain signal conveyed by nerve fibers from one cochlear place, or the model output of one channel. Each column shows the signals from all places, at an instant in time. Along each dimension, we sample the information, typically using several samples per mm of cochlear place, or per critical band, in the place dimension, and tens to thousands of samples per second in the time dimension.

If a cochleagram is sized to fit one second or more of sound across an image, squeezing each channel into about a thousand pixels by resampling to 1000 or fewer samples per second, it will preserve only a few hundred hertz of bandwidth of temporal structure. The fine time structure will be smoothed away in the resampling filter, leaving not much more than a representation of short-time spectrum with pitch-period

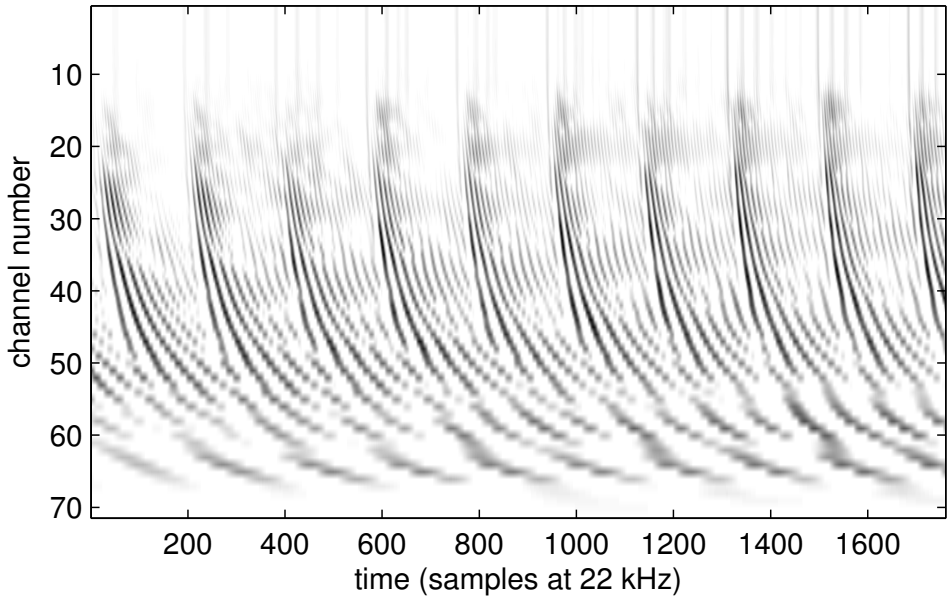


Figure 20.3: A segment of a cochleagram, showing 71 frequency channels as functions of time, in response to a spoken vowel. The IHC output is offset to put the rest response level at zero (white); positive excursions plot as dark regions, and below-rest excursions are clipped to white. This cochleagram spans less than a tenth of a second of sound, not even enough for one syllable of speech.

modulation, as in a wideband spectrogram. Cochleagrams are often used this way, as an auditory enhancement to spectrograms, often sampled at just 100 or 200 samples (spectral slices) per second.

As an alternative, the cochleagram can be displayed at a scale that keeps the fine time structure, say with ten thousand or more pixels per second of sound, as in Figure 20.3. In that case, the width of the image is too large to see much sound at once, so the image needs to be scrolled or paged for viewing. If the image is chopped into pieces, each representing one tenth of a second of sound, and the pieces are displayed in sequence, as a movie, the result can be a confusing mess. When the input sound is periodic with a period of 1/10 s (or 1/20 s, or 1/30 s, etc.), it will make about the same picture in each frame, so the image will be *stable*, and can be examined. But if the sound is periodic with a slightly longer or shorter repetition interval, the image will slide left or right from frame to frame. Strong motion patterns will result from small periodicity differences, and it will be very hard to see the fine time structure that appears at a different place in each movie frame. The overall effect will be very dependent on the chosen frame rate, imposing an arbitrary and artificial set of time-period preferences on the representation of sounds.

Thus, if we are to make a spatiotemporally coherent movie-like auditory image of sound, analogous to the moving image that the retina sends to the visual brain, we need some kind of *stabilization*—to convert the fine temporal structure to a stable fine spatial structure. The auditory nerve provides the temporal structure, and later parts of the auditory brain produce and work with the more slowly changing spatial patterns, as we discuss in upcoming chapters, especially Chapter 21.

20.4 Cell Types in the Cochlear Nucleus

Cells in the cochlear nucleus (CN) have a range of different behaviors in response to sound stimuli, and they are typically classified into different types based both on their patterns of behavior and on structural features. Globular and spherical bushy cells, stellate or multipolar cells, octopus cells, and fusiform or pyramidal cells

are the main structural types; each occurs primarily in a different division of the CN. They partially correspond to response types known as primary-like, chopper, onset, pauser, and others, but not one-to-one (Rhode and Greenberg, 1992; Young and Oertel, 2003). Rather than detail these patterns and types and their connections as many sources do, we have just a few comments.

There are three pathways, or *acoustic stria* from the CN to other parts of the auditory nervous system, from different regions with different response patterns. It seems clear that the different pathways are optimized for different features and functions.

The bushy cells of the anteroventral cochlear nucleus (AVCN) have a primary-like (like the primary fibers of the AN, emphasizing onsets) to onset (responding only at onset) response pattern, and project to the olivary complex, where binaural comparisons are handled. This pathway is probably optimized for speed and synchrony, to assist in precise interaural time comparisons. The onset-emphasizing responses seem well suited to be part of the precedence effect (see Section 22.7). The bushy cells combine inputs from several primary auditory afferents, via large synapses known as the *endbulbs of Held*, and actually exhibit tighter time synchronization to waveform fine structure than the primaries do (Joris et al., 1994). We discuss these in Chapter 22 as likely sources of the onset-trigger events that control the formation of binaural stabilized auditory images.

Other pathways bypass the olivary complex, so are probably optimized for monaural features. Chopper cells in the dorsal cochlear nucleus (DCN) produce spikes at a regular but level-dependent rate, and appear to be specialized for enhancing a rate-versus-place representation of sound spectra (Blackburn and Sachs, 1990).

Onset responses, such as the precisely synchronized firings from octopus cells in the posteroventral cochlear nucleus (PVCN), may be involved in pitch detection (Golding et al., 1995; Langner, 1997; Oertel et al., 2000; Winter et al., 2003). They could be used for generating the trigger events that a stabilized auditory image might be based on. The octopus cells detect coincident arrival of firings of several auditory nerve fibers, in groups spanning about one-third of the tonotopic range (Oertel et al., 2000). Hence, they are good for detecting sharp (wide-bandwidth) onsets precisely; their firing-latency standard deviations are usually about 20 to 50 microseconds (Golding et al., 1995). They are fast enough to fire not just at onsets, but at events such as pitch pulses that repeat a few hundred times per second. The octopus cells project to nuclei in the olivary complex and in the lateral lemniscus, doing primarily monaural processing before sending information to the inferior colliculus, but it remains unclear where in this process a pitch map may be computed.

According to Golding and Oertel (2012), the precise coincidence detection in the PVCN octopus cells includes delay compensation, to better detect wideband events propagating down the cochlea's BM. These cells use dendritic delays, with mechanisms partially similar to what the principal cells for the MSO use for detection of coincident events from two ears with a preferred interaural delay (see Chapter 22). The delays involved are generally less than 1 ms, but are still remarkable as part of a system that achieves a latency spread, or timing uncertainty, of only a few tens of microseconds.

20.5 Inhibition and Other Computation

The AN and CN are more than just transparent communication relays. As mentioned above, some CN cells enhance synchrony to waveform events. Inhibitory mechanisms that have been described in AN and CN implement a sort of *lateral suppression* (Rhode and Greenberg, 1994) that helps to maintain a rate–place spectral representation across sound levels and noise levels in several other CN cell types. While some of the suppression originates in the cochlea and AN, via rapid and short-term adaptation processes (Mountain and Hubbard, 1996)—which we have described in terms of coupled AGC—there is additional lateral suppression in specific circuits in the CN, especially for the chopper and pauser cell types.

Other cells in DCN, such as pyramidal cells and vertical cells, are thought to use lateral inhibitory connections to detect more specific features of sound, such as narrow spectral peaks and notches, or sharp edges

(Young and Oertel, 2003; Reiss and Young, 2005). Additional hypothesized roles of the CN include onset detection, periodicity analysis, and dereverberation (Werner et al., 2010). Some DCN cells project to areas that mediate the acoustic startle reflex, and it is speculated that they may be involved in providing directional information (from monaural spectral cues) about the startling event (Young and Davis, 2002).

20.6 Spike Timing Codes

In general, neurons encode information by when they fire. In both vision and audition, there seems to be a lot of information encoded in detailed spike timing patterns. Cariani (1999) shows responses of different cell types in different CN areas to a vowel sound. All of them show clear temporal synchrony; the AVCN's primary-like units show synchrony to both the pitch period and the fine structure of formant resonance, while the PVCN's choppers and the DCN's pauser units synchronize to the pitch period. CN chopper and onset units have been shown to *mode lock* to periodic envelopes and sound signals (Laudanski et al., 2010), meaning that they usually fire in the same time pattern on each period of the stimulus. Thus the functions of some of these cell types may be more about encoding envelope or periodicity than spectrum.

It remains to be seen whether machine models of such functions will benefit from using a similar spike-timing model, or whether we can continue to get away with ignore spiking in most cases, modeling the functional properties of these systems using the usual discrete-time signal processing approaches. At the level of auditory nerve, our several dozen channels represent the same information for which the auditory nerve uses tens of thousands of spiking fibers. Representing the spikes explicitly might be computationally expensive; but if spike densities are low, a spike-timing representation can achieve high timing precision more efficiently than a sampled-data approach can.

Chapter 21

The Auditory Image

> We must think of the neural arrangement, therefore, as extended in two spatial dimensions. The one corresponding to frequency is the x-dimension, or the dimension of the nervous tissue into which the lengthwise dimension of the cochlea projects. The whole arrangement for determining autocorrelation functions is replicated in the x-dimension. The τ-dimension is functionally orthogonal to the x-dimension, and we can think of it, at least for convenience of graphical representation, as being spatially orthogonal, also. The over-all system, then, yields a representation of the stimulus $f(t)$ in two spatial dimensions and time, a running autocorrelation $\phi(t, \tau, x)$ of the components in each of many frequency bands.
>
> — "A duplex theory of pitch perception," J. C. R. Licklider (1951)

> In contrast to the two-dimensional visual and somatosensory receptor surfaces, the cochlea provides only a one-dimensional rendition of the impinging acoustic energy distribution along the organ of Corti. Consequently, cortical frequency maps can expand along the second dimension of the cortical sheet, providing additional territory for signal processing while closely preserving receptor-related neighborhood relationships.
>
> — "Auditory cortex mapmaking: principles, projections, and plasticity," Schreiner and Winer (2007)

The stabilized auditory image (SAI) captures much of the salient information in the short units of sound that make up speech, music, and other important sound events. It complements the information in the traditional short-time spectral representation of sound, and allows the partial segregation of sound fragments that have different temporal patterns, whether periodic repetition or otherwise. An extension of the SAI concept to longer time lags can capture beat and rhythm in a form similar to how the SAI captures pitch.

21.1 Movies of Sound

Licklider explained his *duplex theory* in terms of a neural pattern of two spatial dimensions and one time dimension. The auditory image approach adopts this movie-like dimensionality more generally, as the sort of representation that must be present in various 2D *sheets* of neural tissue, but particularly in projections to primary auditory cortex, just as images from the retina project retinotopically to primary visual cortex.

We use it, as Licklider did, as a representation *of the stimulus* first; but in addition, the same concept can be invoked to create derived representations, further from the stimulus and closer to the percept or interpretation evoked by the stimulus.

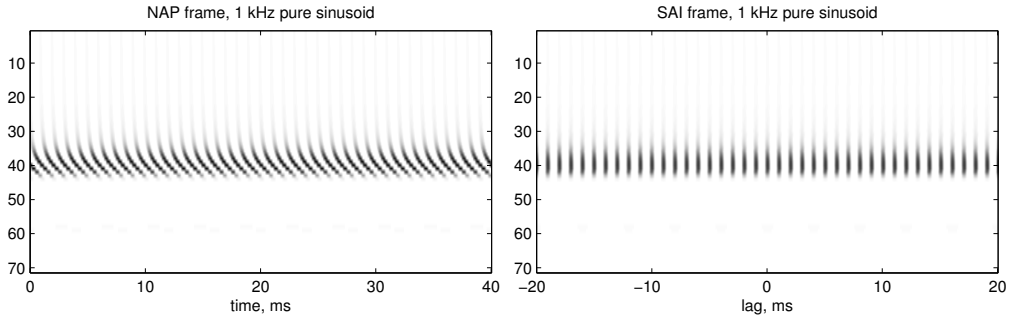


Figure 21.1: The NAP (left) and SAI (right) of a 1 kHz pure tone have a lot in common. Both are simple patterns with period matching the 1 ms tone period. The top part of each image shows a weak response of high-CF channels (low channel numbers) to the 1 kHz tone as it propagates through the cochlea; the middle shows a strong response for places with CF near 1 kHz; and the lower part shows essentially no response for channels with lower CFs. The NAP shows curvature patterns from propagation delays of the cochlear filtering, as we saw in Figure 18.8, and has no certain time origin that could be used to make a stable spatial pattern. The SAI's stabilization process straightens the pattern and stabilizes it, aligning peaks at the 0 lag point.

In this chapter, we focus primarily on the initial Licklider-type auditory image, as a representation of monaural stimuli, or of the signals at each ear independently. We explore the relationship between Licklider's *running autocorrelation* and Patterson's *triggered temporal integration* as approaches to auditory image stabilization (Licklider, 1951; Patterson et al., 1992; Patterson and Holdsworth, 1996).

When the stimulus is as simple as a pure sinusoid, the neural activity pattern (NAP) and SAI are also very simple, as shown in Figure 21.1. For this or any steady stimulus, the movie will be steady, too; that is, every frame will look alike.

21.2 History

In a nutshell, the auditory image can be thought of as a version of the running autocorrelation in Licklider's duplex theory: a map of activity as a function of both frequency (or cochlear place) and temporal periodicity (or time lag), all changing in time, like a movie. The image is stabilized, in that prominent temporal features are anchored to a specific zero-lag place in the movie frame.

The term *auditory image* has been used by different authors for somewhat different concepts over the last several decades. Our usage descends primarily from the Patterson et al. (1992) SAI, a realization of Licklider's concept that is closely related to what we had previously called an *auditory autocorrelogram*, or simply a *correlogram* (Duda et al., 1990; Slaney and Lyon, 1990, 1993).

Earlier alternative uses of the term include those of Altman and of McAdams. To Altman and Viskov (1977) a *fused auditory image* was a spatial sound source perception in binaural listening—a spatially located sound source percept. McAdams (1982) also used *auditory image* to describe the perception of an auditory object or source, as opposed to our use for an internal representation of a sound mixture. Yost (1991) adopted a similar interpretation in terms of sound source perception. These are not "movie-like" representations, and not what we mean by *auditory image*.

Patterson et al. (1992) say that the auditory image is "intended to represent our initial impression" of a complex sound, rather than represent an inferred source, and it is literally in the format of a moving image. They describe how they construct SAIs from the output of a cochlear model via a stabilizing *trigger* mech-

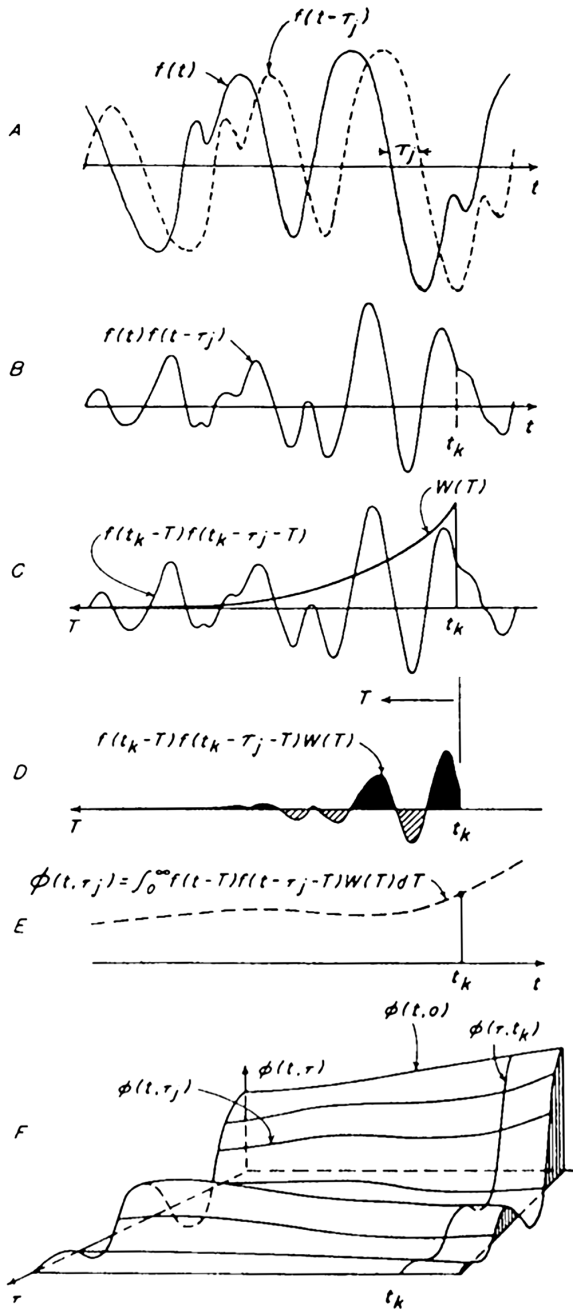


Figure 21.2: Licklider's illustration of the computation of a running autocorrelation function by smoothing the product of a signal $f(t)$ times a delayed version of the same signal (he uses T as the dummy integration variable, where we used u). The panels show: (A) the input signal and delayed input signal, for a particular delay, or lag, τ_j; (B) the product of the signals from panel A; (C) the product waveform, relabeled in terms of time offset T from a particular time t_k, superimposed with the exponential weighting function $W(T)$ (the time-reversed impulse response of a first-order smoothing filter); (D) the weighted product, in this case showing larger areas under the positive parts than under the negative parts; (E) the integral of the signal in panel D, $\phi(t,\tau)$ for the particular τ value, as t changes; (F) the two-dimensional surface $\phi(t,\tau)$, showing slices at various values of t (time) and τ (time lag). Notice that this surface is changing slowly in time, but captures fine temporal information in its lag dimension. [Figure 4 (Licklider, 1951) reproduced with permission of Springer.]

anism and *temporal integration*. They contrast this approach with earlier Licklider-style auditory images that were described with different terminology (Lyon, 1984; Assmann and Summerfield, 1989; Meddis and Hewitt, 1991).

Cooke (1993) noted that what he called *ACGs* (autocorrelograms) had been mentioned by most of these authors as representations from which sound source separation could begin, as opposed to representations of the outputs of sound separation. Mitch Weintraub and I had done early monaural and binaural sound separation experiments with variants called auto-coincidence functions (Lyon, 1983, 1984; Weintraub, 1984, 1987).

21.3 Stabilizing the Image

In Section 20.3, we referred to the need for some kind of stabilization, if we want to make a graphical representation of the fine temporal patterns that are present on the auditory nerve. Short-time autocorrelation, or "running autocorrelation" as Licklider formulated it, is one method that converts fine temporal pattern relationships in a signal $f(t)$ to a slowly time-varying function of a new parameter τ: $g(t, \tau)$. That is, applying a running autocorrelation analysis to every frequency channel (each x coordinate, or place) of a cochlear model output $f(x, t)$ will produce a slowly changing *stabilized* auditory image $g(t, \tau, x)$.

There are various ways to define a short-time autocorrelation. Consider Licklider's method: the running autocorrelation $g(t, \tau)$ for the delay τ, as a function of time t, is a weighted average, over times shortly before t, of the product of $f(t)$ times a delayed version of itself, $f(t - \tau)$:

$$g(t, \tau) = \int_0^\infty f(t - u) f(t - \tau - u) w(u) du$$

where w is a weighting function applied to past values of the product. An examination of the integral reveals that it is a convolution with $w(t)$, representing a simple linear filtering of the product $f(t)f(t - \tau)$:

$$g(t, \tau) = (f(t)f(t - \tau)) * w(t)$$

See Figure 21.2, Licklider's illustration of this method. It shows the steps to get from $f(t)$ to the two-dimensional surface $g(t, \tau)$ (he calls it $\phi(t, \tau)$). This function varies slowly in t and captures the fine temporal structure of $f(t)$ along the τ dimension. He uses an exponential weighting, so the convolution is easily implemented as a one-pole smoothing filter.

This method has the advantage that it is causal; that is, the values of $g(t, \tau)$ depend only on $f(t)$ for earlier values of t, so they can be computed easily in real time by nonlinear circuits (or in discrete approximation by digital computation).

We illustrate Licklider's method on one channel of a neural activity pattern (NAP) or inner-hair-cell output from the CARFAC model in a sequence of figures: Figure 21.3 shows the input waveform, which comes from one channel of a CARFAC cochlear filterbank; Figure 21.4 shows the 2D instantaneous product of the input signal with delayed versions of itself; Figure 21.5 is the time-smoothed product; and Figure 21.6 shows four sample slices of that $g(t, \tau)$ at 20 ms intervals.

21.4 Triggered Temporal Integration

Patterson introduced an alternative method of computing a Licklider-style auditory image, based not on autocorrelation, but on a closely related process that he calls *triggered temporal integration* (TTI). A good way to think of it is as a modification of the autocorrelation into a cross-correlation of the signal with a sparsified version of itself. The sparsified signal is just a set of impulses at the trigger times, where trigger times are

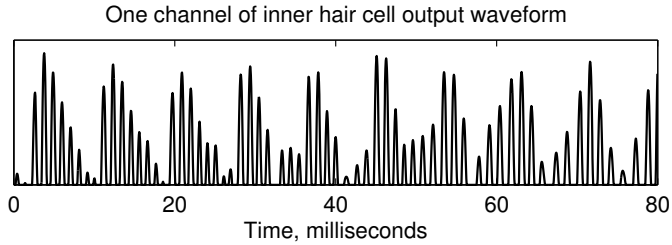


Figure 21.3: The waveform that comes from a single inner hair cell is the nonlinearly detected view of the basilar membrane motion at one place corresponding to one characteristic frequency: one frequency channel. Here we show an 80 ms segment, for channel 42 (a place with CF near 800 Hz) of the cochleagram in Figure 20.3, responding to a spoken vowel of about 120 Hz pitch. Several levels of temporal structure are apparent. This signal represents the input to the calculations illustrated in subsequent figures, analogous to Licklider's input signal, the solid curve in Figure 21.2(A).

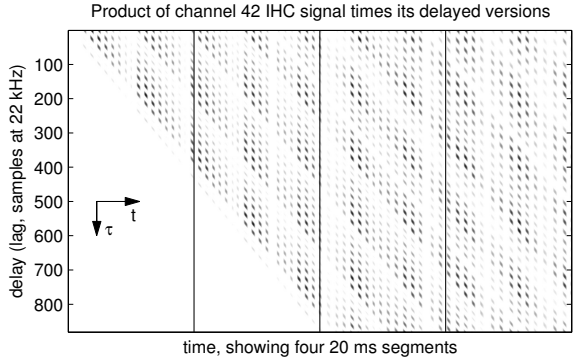


Figure 21.4: The instantaneous products of the channel-42 signal times its delayed versions, for up to 880 samples of delay (out to 40 ms). Vertical lines indicate boundaries between 20 ms segments. This image is analogous to Licklider's Figure 21.2(B), except that we show it in two dimensions for many values of lag, with more-positive values plotting darker, as opposed to his single illustrated lag value.

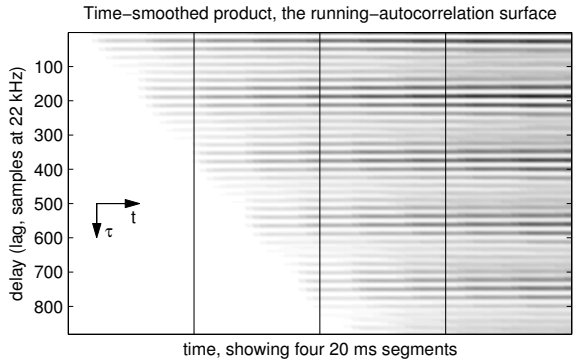


Figure 21.5: Smoothing the products along the time dimension, using a first-order filter with 60 ms time constant, results in this $g(t, \tau)$ running autocorrelation image; the fine time structure is now in the τ dimension, while the function changes only very slowly in time. This image is analogous to Licklider's slowly changing short-time autocorrelation surface of Figure 21.2(F), which is the all-lags version of the single slowly changing correlation coefficient of Figure 21.2(E), which is the signal from (B) smoothed using exponential weighting as in (C) and (D).

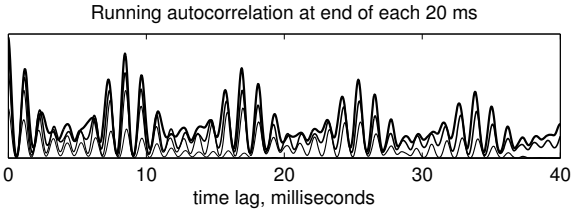


Figure 21.6: Four slices of the running $g(t, \tau)$, at the end of each 20 ms segment, show how the function of lag changes slowly. Later slices are shown with heavier lines. Each slice is a good estimate of the one-sided short-time autocorrelation function of the channel-42 signal. The τ values shown as positive here represent correlations of the current time with the past. In some other figures, we turn the lag axis around and put the past on the left. These four slices, taken at marked times spaced 20 ms apart, are analogous to time slices at times t_k in Licklider's Figure 21.2(F). They all have maxima at the zero-lag position.

chosen to coincide with some of the most prominent peaks in the waveform. The other way to look at the process is as the sort of triggered synchronization that an oscilloscope display gives you. We examine both of these views.

21.4.1 Cross-Correlating with Trigger Impulses

In Licklider's running autocorrelation function:

$$g(t, \tau) = (f(t) f(t - \tau)) * w(t)$$

we can replace either $f(t)$ or $f(t - \tau)$ with a nonlinearly derived trigger signal $\hat{f}$ that is just a sequence of impulses at the trigger times, to generate these alternative versions for the elements (pixels) of the SAI:

$$g_1(t, \tau) = (\hat{f}(t) f(t - \tau)) * w(t)$$

$$g_2(t, \tau) = (f(t) \hat{f}(t - \tau)) * w(t)$$

The function with argument $t - \tau$ is the output of the delay line at delay τ. The g_1 version therefore represents the case of the continuous f being put through the delay line, while the g_2 version represents the case of the trigger events being put through the delay line. We use the former, though the latter can potentially be done with less delay state memory, especially in implementations that use sparse representations (Weintraub, 1984). For our plots with a lag axis, we typically define lag as $-\tau$, so that activity preceding the trigger event (from earlier times $t - \tau$) is shown to the left of the zero-lag point.

Figure 21.7 shows an example of finding trigger points for $\hat{f}(t)$ for one row (place or CF channel) of a NAP, using the simple algorithm of picking the highest point in each 20 ms segment as a trigger point, and shows the relationship between four segments without and with alignment to those trigger points, as well as a summary of those four segments by averaging them after alignment to the trigger points.

The g_1 version is best for displaying correlation in signals that lead up to the trigger points, while the g_2 version shows patterns that follow the trigger points. The patterns are in many ways similar, especially when the signals are periodic, related to the running autocorrelation patterns that both derive from.

In either alternative, the nonlinear trigger function $\hat{f}(t)$ can consist of either unit impulses, or impulses scaled according to the size of the $f(t)$ peaks that produced them.

Figure 21.8 extends this example to show trigger points on all channels of the NAP, for a selected segment, along with a resulting SAI frame and summary SAI.

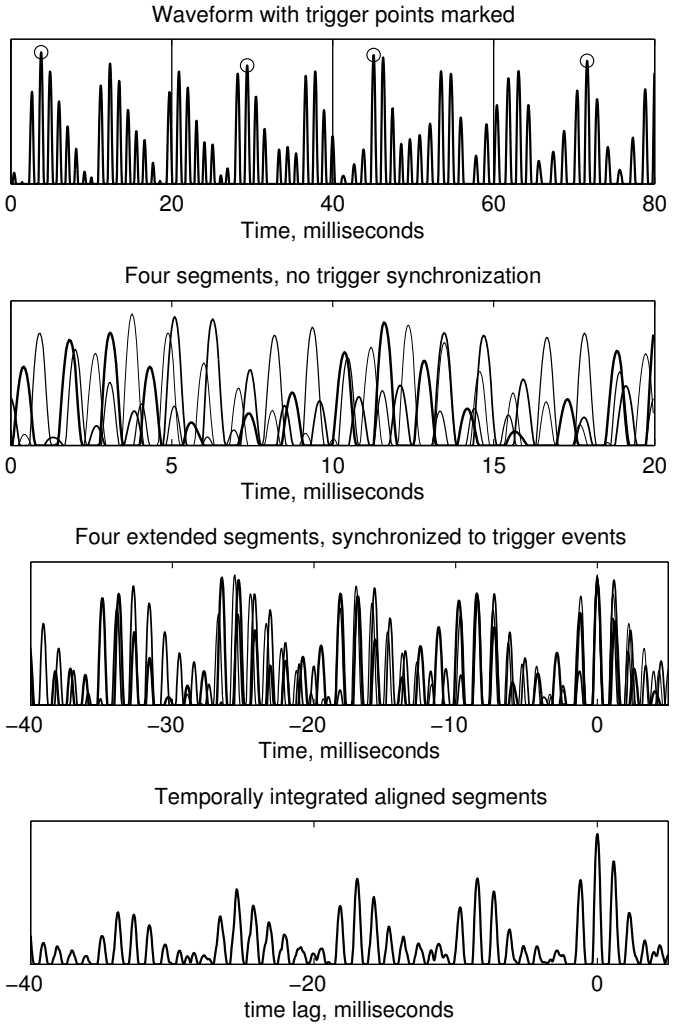


Figure 21.7: The signal of Figure 21.3 is shown divided into 20 ms segments, with the highest value in each segment indicated by circles (top panel). We take these time points as trigger events. If we display the 20 ms segments of the waveform together, they do not line up, and look like a mess (second panel). If instead we summarize the waveform's changes in time by displaying together segments that are aligned based on the trigger events, the picture is much less confusing (third panel). Here, the relative maximum in each 20 ms segment was aligned to the τ origin, and the input signal from 40 ms before to 5 ms after each trigger event was plotted. The signal is not exactly periodic, so the segments do not quite stay aligned away from the time lag origin, but the approximate repetition is clear in the approximately consistent pattern. The aligned segments are then *temporally integrated*, or averaged across the four different trigger times to make one row of one frame of an SAI (bottom panel).

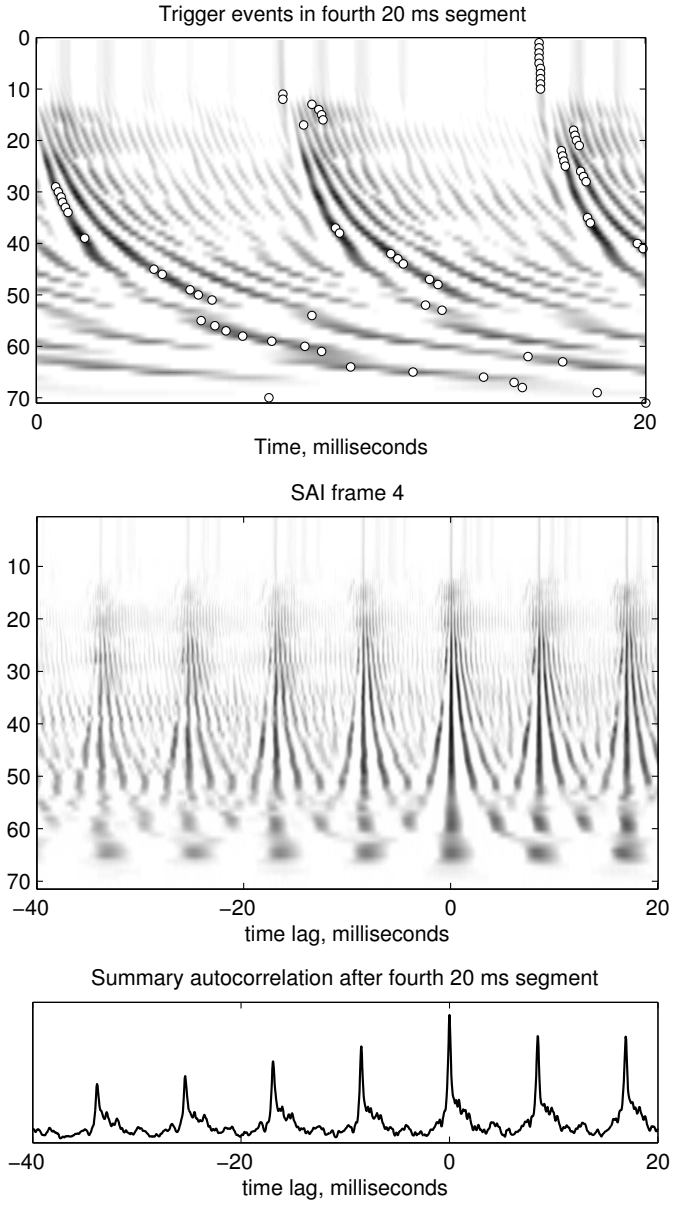


Figure 21.8: The simple triggering algorithm of picking the maximum point in a segment, for each channel, results in these irregular trigger events, shown as circles overlaid on one 20 ms segment of cochleagram (top). The SAI made with the simple trigger algorithm (middle) shows the pitch clearly, but also shows some discontinuities between rows. Even with more sophisticated trigger algorithms, some of this effect will be seen. The average along columns of the auditory image is the summary SAI (bottom), sometimes called the summary autocorrelogram (SACG) especially if the rows are computed by autocorrelation. The trigger irregularities have little effect on the summary.

21.4.2 Triggered Synchronization

Speech signals and other waveforms are often displayed on an oscilloscope screen through a triggering process—starting the horizontal sweep when the voltage signal crosses a trigger level. This kind of synchronization stabilizes the display so that the structure of the signal can be viewed easily, relative to the trigger point. In classic oscilloscopes, the display would show only what happens *after* the trigger, but in modern scopes, digital sampling and buffering allows a display of what happened *before* each trigger event as well. These are the same operations that we can get by displaying positive and negative lag ranges of the cross-correlation of the signal with the trigger event signal as described in the previous subsection.

The temporal integration part of Patterson's TTI concept is the time-averaging of these traces. We do this averaging in our eyes when we look at an oscilloscope display. For the purpose of SAI generation, an explicit averaging over intervals of around 10 ms or more may be sensible, possibly using higher averaging weights for traces with higher values at the trigger points.

The previous example is extended in Figure 21.9 to show multiple trigger events per segment. The resulting SAI is more coherent (has less tearing or discontinuity artifacts) compared to the one with only one trigger per segment, due to the extra averaging, or temporal integration. Its summary SAI is not significantly different.

21.5 Conventional Short-Time Autocorrelation

A typical approach in acoustic analysis is to process windowed segments of waveforms. The short-time autocorrelation function (STACF), which is the autocorrelation function of a windowed segment of a waveform, is an alternative way to stabilize the description of repetitive structure in a signal. It uses a windowed segment of each NAP row, as shown in Figure 21.10, and generates from it a symmetric STACF, as shown in Figure 21.11.

Whichever way the short-time autocorrelations are computed, Licklider's method, TTI, autocorrelation of windowed segments, or otherwise, they can be stacked together as rows to make an auditory image, a function of two spatial dimensions, slowly varying in time.

21.6 Asymmetry

Patterson and Irino (1998) have pointed out that the asymmetry of the SAI based on TTI captures the subtle phase sensitivity of human hearing, and that the symmetric autocorrelogram will lose that sensitivity. They did experiments with "ramped" versus "damped" sinusoids, periodic sound signals with a strong time asymmetry about their peaks, in opposite directions through their periods, and found that listeners could easily distinguish these signals with identical power spectra. Irino and Patterson (1996) reported (for 40 Hz modulation) that, "whereas the damped sinusoid produces a unitary perception, like a drummer playing on a hollow wood block, the ramped sinusoid produces a two-component perception—a drummer playing on a nonresonant surface with a soft, continuous sinusoid in the background." Correlates of these descriptions are subtle but visible in the SAIs of Figure 21.12, where the ramped signal's SAI has more of the signature of a continuous sine tone throughout it. In that figure, autocorrelograms are shown for comparison; subtle differences are visible, but the relationship to the perceptual description is more remote.

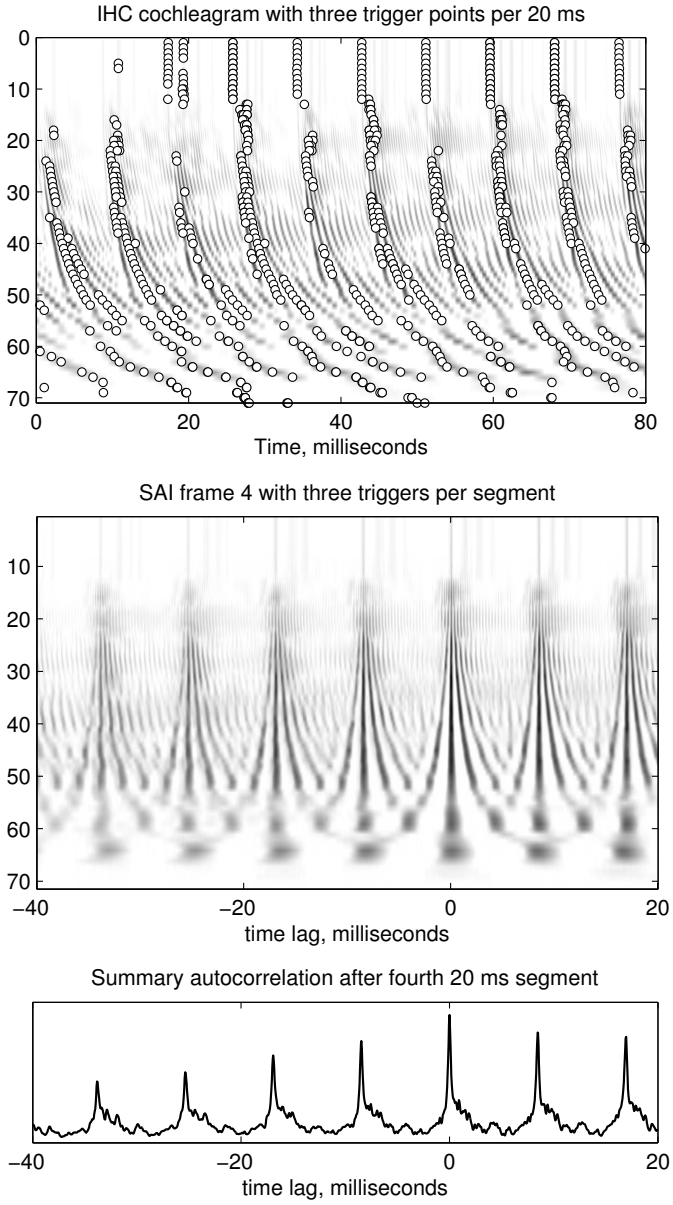


Figure 21.9: The trigger points for 80 ms of the cochleagram (top), using three peaks per 20 ms segment, selected as points of maximum value after weighting with overlapping sine windows, the windows being two segments wide and spaced one-third segment apart. With this method, the points chosen are sometimes not exactly at peaks of the original signal (due to the slope of the window), and the same time point is sometimes chosen more than once (due to the overlapping windows). In the resulting SAI (middle) discontinuities are less prominent with this larger number of trigger points contributing to the temporal integration. The summary SAI (bottom) is not much different from the simpler one shown in Figure 21.8.

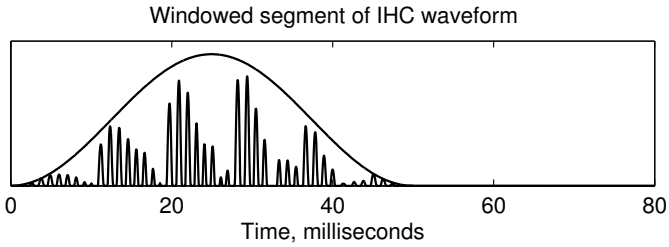


Figure 21.10: The 80 ms segment in Figure 21.3 has been multiplied by the 50 ms raised-cosine (Hann) window to make this windowed segment. Such a segment can be made again 20 ms later, or on whatever frame times are desired.

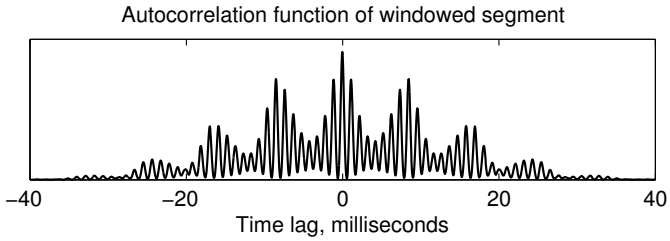


Figure 21.11: The autocorrelation function of the windowed segment in Figure 21.10 is this symmetric function of the time lag parameter. Notice that it resembles the result of TTI in Figure 21.7, at least for moderate lag magnitudes, but the TTI version is sharper and not symmetric.

21.7 Computing the SAI

The descriptions above in terms of continuous-time signals stay close to the formulation of Licklider and to conventional mathematics of correlation. For machine computation, we switch to a discrete-time formulation, both for the input NAP and for movie-like output frames at a much lower sample rate.

We typically produce SAI movies at a constant frame rate, for example tied to the segment rate by processing all the trigger events in all channels for a segment. The place or frequency dimension is indexed as channels of the cochlear model; we use x here for this channel index, sometimes writing $f(x,t)$ for $f(t)$ to be explicit about the channel dimension.

We adopt a formulation like the trigger-based g_1 above, with $\hat{f}(t)$ being a sparse sequence of trigger events, nonzero at times t_{trigger}, and $f(t-\tau)$ being values from the input buffer. Whenever the trigger sequence is nonzero, the input values $f(t_{\text{trigger}} - \tau)$ from the input buffer are used, for a fixed selected range of discrete τ values, spaced at the input sample rate. If the selected range is nonnegative, then only past values (before the trigger time) are needed; if the selected range includes negative τ values, then NAP samples after the trigger event are needed, so the input buffer needs to include these future values, essentially by deferring the trigger generation, adding a little latency to the process. The same τ range is used for every channel, to make a rectangular image.

Whether we wait until the end of a segment or not, we can consider the image to be updated at every trigger event, rather than continuously or at the end of each segment. Instead of defining the temporal integration by a smooth weighting function $w(t)$, we use a discrete update rule for row x of image $I(x,\tau)$ each time we get a trigger event on that row:

$$I(x,\tau) \leftarrow \alpha f(x, t_{\text{trigger}} - \tau) + \beta I(x,\tau)$$

where the coefficients α and β might be constants, or might depend on the time since the last trigger, or on the

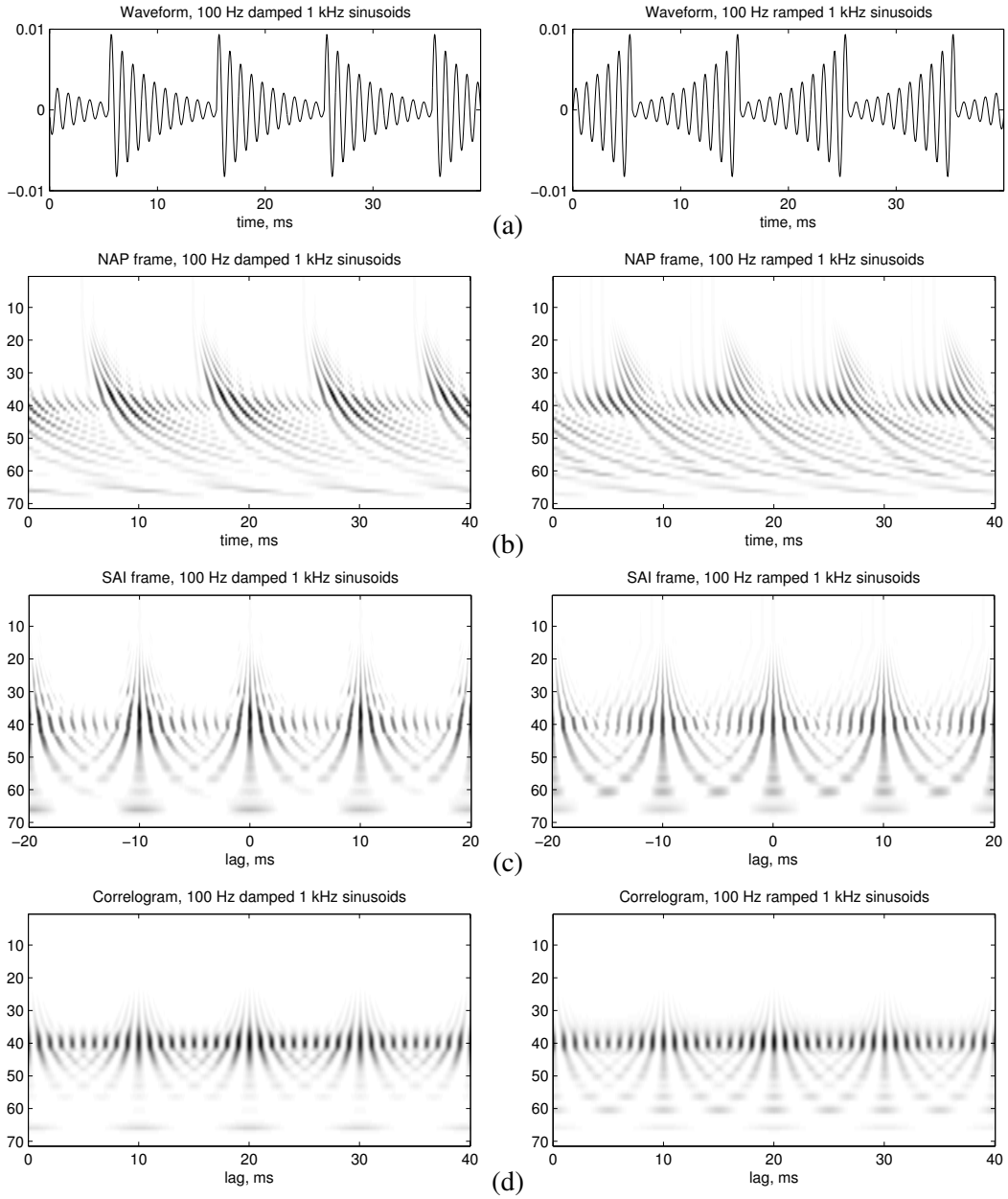


Figure 21.12: Damped (left) versus ramped (right) sinusoids (a), snippets of the resulting NAPs (b), SAI frames (c), and autocorrelogram frames (d). The damped and ramped signals are time reversals of each other and thus differ only in the phase of their Fourier components. Their SAIs show subtle differences that correspond to subtle perceptual differences. Compare with the SAI of a 1 kHz pure tone in Figure 21.1. The correlograms, though symmetric in lag, also show some differences in ramped versus damped, because the NAP rows are not time-reversals of each other. This difference is due to the AGC's lagging gain variation, which results in a stronger fundamental-frequency distortion tone in the damped case, but stronger harmonics in the ramped case.

amplitude at the trigger time, or both.

If β is an exponentially decreasing function of time since last trigger, then this update rule can be a good approximation to the exponential smoothing that Licklider describes. Alternatively, we can get a fair approximation to exponential smoothing by using a fixed β; for example, with $\beta = 0.9$ and four trigger events per frame, we get a smoothing time constant of about 10 events or about 2.5 frames.

If we choose $\alpha = 1 - \beta$, independent of the amplitude of the trigger event, then the filter is essentially a smoothing filter, so the output movie frame has the same average value as the NAP in each row. If instead we include a trigger-peak amplitude factor in α, that makes the result more correlation-like, putting more emphasis on the time structure relative to the larger trigger events, and expanding the output dynamic range to be proportional to the square of the input range.

We have chosen a method that emphasizes larger trigger events without expanding the output dynamic range, by making $0 < \alpha < 1$ an increasing function of trigger event amplitude and $\beta = 1 - \alpha$, to make a first-order smoothing filter with variable time constant. The result is that larger trigger events update the image with a shorter time constant, so their results show up more immediately, while smaller trigger events have relatively less effect on the image, allowing the effect of stronger events to last longer. Thus an isolated strong event will be captured and "stretched" to show up for more than a few frames in the SAI movie. Due to the stretching, the average output will be slightly higher than the average input, but the peak output will be about the same as the peak input.

This approach requires picking an amplitude scaling, and mapping it such that "strong" and "weak" trigger events are distinguished by large and small α values (say 0.5 versus 0.1 for time constants of 2 events versus 10 events). Very weak input will lead to long-time-constant smoothing, and very strong input with most α values being near 1 will lead to just reproducing the pattern of the latest trigger event in each frame, so some extra adaptive range compression may be needed to keep the range of α values in a good region; it is likely that efferent feedback tunes such mappings in the auditory nervous system. For now, the open-source CARFAC SAI code uses this fixed mapping (where the f values, the outputs if the IHC model, are mostly less than 1, but can briefly be much higher):

$$\alpha = \frac{f(x, t_{\text{trigger}})}{1 + f(x, t_{\text{trigger}})}$$

21.8 Pitch and Spectrum

The two dimensions of the auditory image represent frequency (or spectrum, in the Fourier sense that Ohm and Helmholtz emphasized) and pitch (or temporal structure, periodicity, etc. in the sense that Seebeck emphasized), and the 2D image shows their interaction. Summing along rows is one way to get a spectral estimate, while summing along columns gives a *summary autocorrelogram*, which will have peaks at likely pitch periods.

These two dimensions represent the "duplex" that Licklider speaks of. The spectral dimension that characterizes things like vocal formants and instrument timbre also carries some aspects of pitch. The temporal dimension carries the other important aspect of pitch, repetition in time, which may be the more important aspect in most cases.

Figure 21.13 shows cartoons of the characteristic local shapes that arise as the ringing of the bandpass cochlea channels interacts with the repetition-time analysis in the auditory image.

21.9 Auditory Images of Music

Almost any kind of sound can be used to make music, so one frame of an SAI of music can look like the SAI of almost any sound. Typically, however, sounds with well-defined pitches are used to carry a melody, and

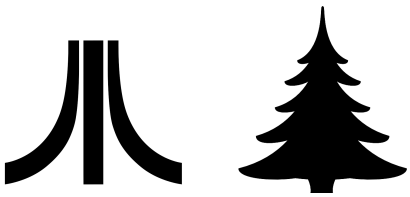


Figure 21.13: The local structures in SAIs are sometimes said to resemble Atari logos or Christmas trees. The wider fringe spacing toward the bottom is caused by the decreasing ringing frequency as waves propagate through the cochlea from the base (channels near the top of the picture) toward the apex (channels near the bottom of the picture).

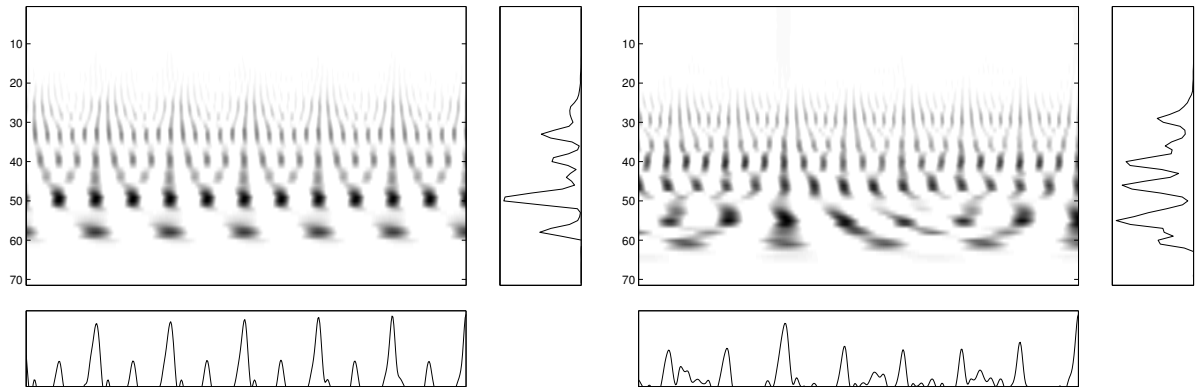


Figure 21.14: These SAIs of piano notes—an isolated note on the left and a chord on the right—show the activity *before* the trigger events, so the trigger times are aligned at the right edge. In the chord, the temporal profile (the average along columns) plotted at the bottom shows the root pitch period at 5 times the period of the highest note. The "auditory spectrum" plotted on the right is indicative of a sort of overall timbre. Compare with the SAIs of one, two, or three steady notes shown in Figure 4.10. The slightly higher pitch of high harmonics, a characteristic of piano strings, is apparent in the uppermost parts of these SAIs.

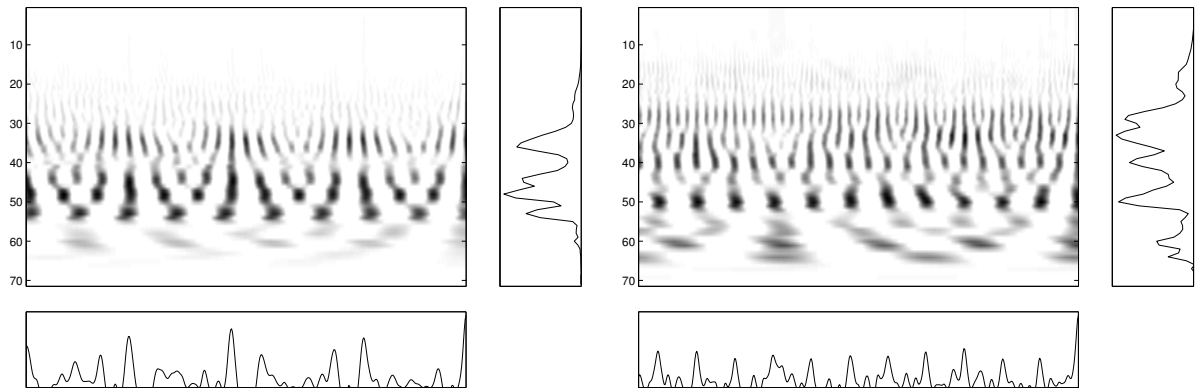


Figure 21.15: Two SAI frames of a jazz music piece. The sound represented is primarily a cowbell-like percussion note on the left, and a more complex mixture with multiple pitches on the right.

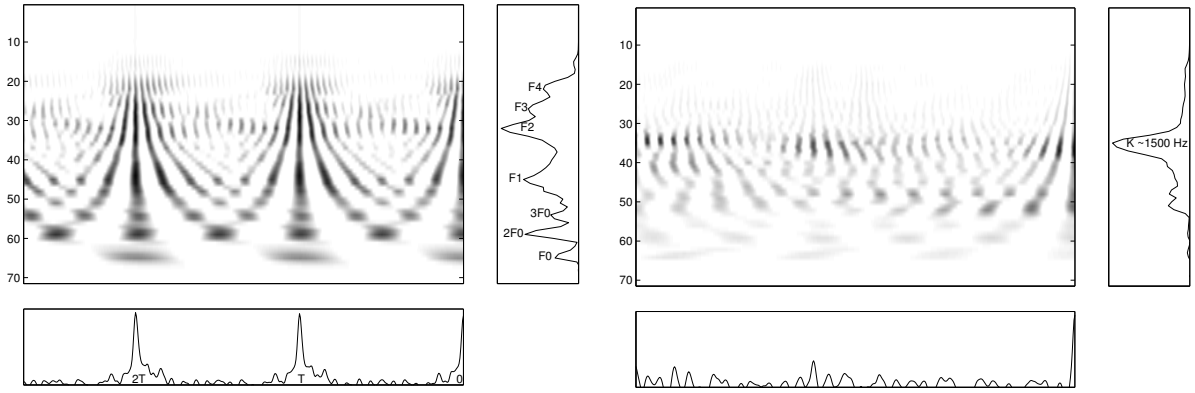


Figure 21.16: SAI of a spoken vowel /æ/ (in "plan") with a pitch of about 122 Hz, on the left, and of a /k/ (voiceless velar stop consonant) release burst, on the right. In the left panel, both the row and column averages show the pitch (frequency F0, period T) of the vowel, but the temporal profile (the average along columns, at the bottom) shows it more clearly and explicitly. The auditory spectrum clearly shows the formants, especially the first and second formants F1 and F2, which are most important in determining the perceived vowel category. In the right panel, the temporal profile of the /k/ shows no periodicity. The auditory spectrum shows the compact k-release burst resonance in the F2 region, typical of a velar stop release.

the SAI provides a clear indication of the movement of the melodic pitch over time. See Chapter 27 for how we exploit this property of SAIs in melody recognition.

Figure 21.14 shows two sample SAI frames from a piano piece. Figure 21.15 shows two sample SAI frames from a musical piece with percussion and strings present. At any given time, the important pitches in such frames are sometimes ambiguous, but in the "movie" version the movements and relations between pitches are more clear.

21.10 Auditory Images of Speech

The acoustic properties of speech are clearly visible in SAIs. This is to be expected, because speech evolved to suit auditory processing and representations, and because the SAI is designed to capture what we know about such representations.

Figure 21.16 shows a typical SAI frame of a vowel and an unvoiced consonant. The average along rows of these SAIs is an auditory spectrum that captures the kind of information typically used in speech recognition. It is not necessary to compute an SAI to get this information, since it's approximately equivalent to the short-time average of the neural activity pattern. The extra information that the SAI provides is in the more clear representation of pitch in the average along columns, and in the way that multiple concurrent speech sounds can be partially separated in the full SAI.

When several speech sounds are present concurrently, the picture is more complicated, as illustrated in Figure 21.17. Portions of the SAI movie will be seen to move together, following the pitch of one voice or the other. Figure 21.18 shows how one frame may be interpreted as showing portions of each of two vowels, as well as some "ghost" features, interrelations between pitch pulses of the two vowels that are not stable from frame to frame.

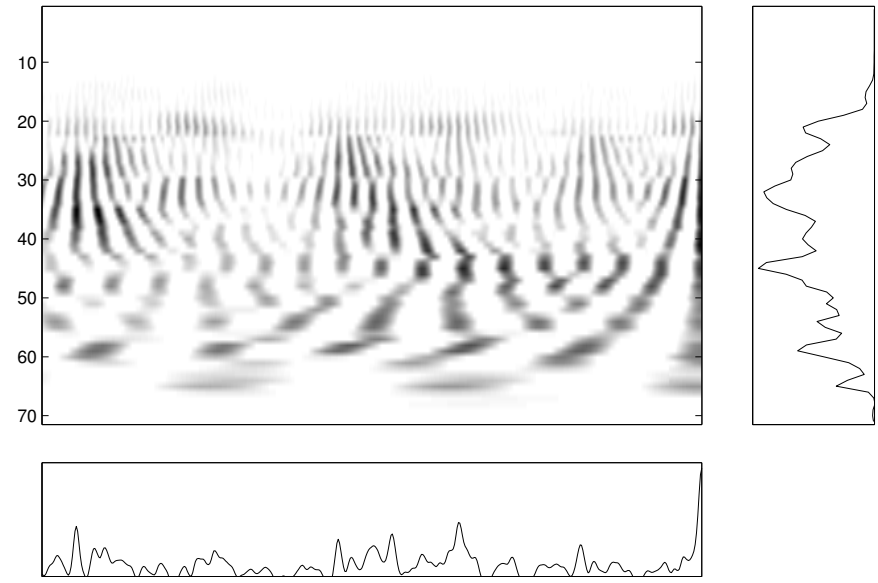


Figure 21.17: SAI of two concurrent vowels: the /æ/ of Figure 21.16 added to an /aɪ/ diphthong from the same speaker but at a lower pitch. The row and column summaries are no longer very informative, but the partial "Atari logo" structures in the image provide locally separated responses to the two vowels—along with some ghost responses at time lags equal to the intervals between pitch pulses of the two vowels. Viewed as a movie, the SAI shows coherent motion of each vowel. See Figure 21.18 for an analysis of this mixture SAI image.

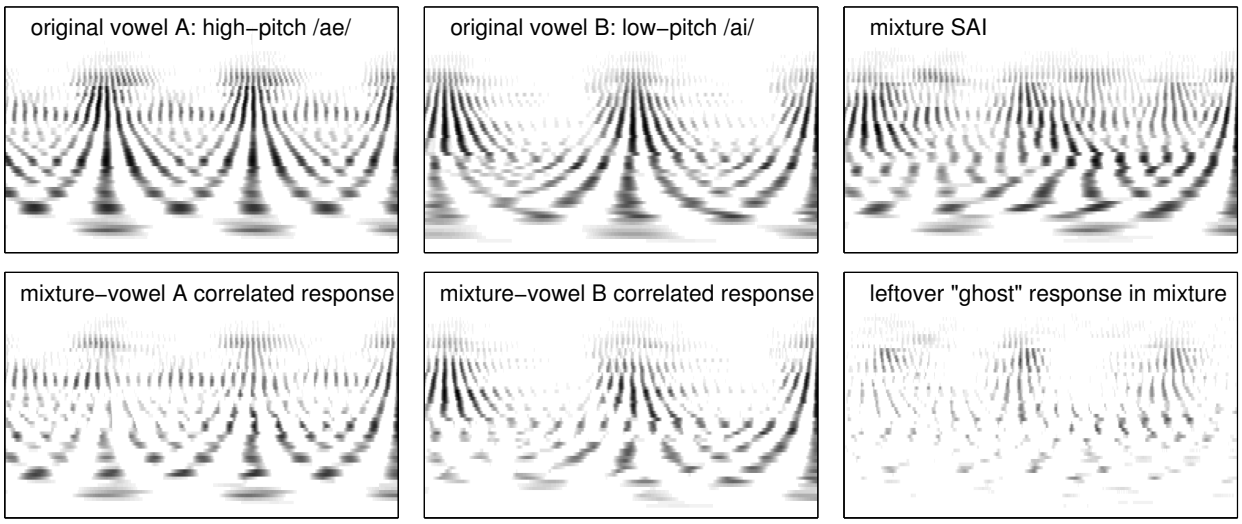


Figure 21.18: The SAI of the concurrent vowel mixture in Figure 21.17 is analyzed to show how it relates to the SAIs of the original clean vowels that were mixed. The SAIs of the two vowels and the mixture are shown on top. Below the originals, the portions of the mixture that match are shown (extracting as the max of each point in the pair original and mixture SAIs). Below the mixture SAI is shown the difference between the mixture SAI and the max of the two original SAIs, which yields the pattern that does not match either original sound's pattern—the "leftovers" or "ghosts" caused by one vowel's pulses corrrelating with the other's.

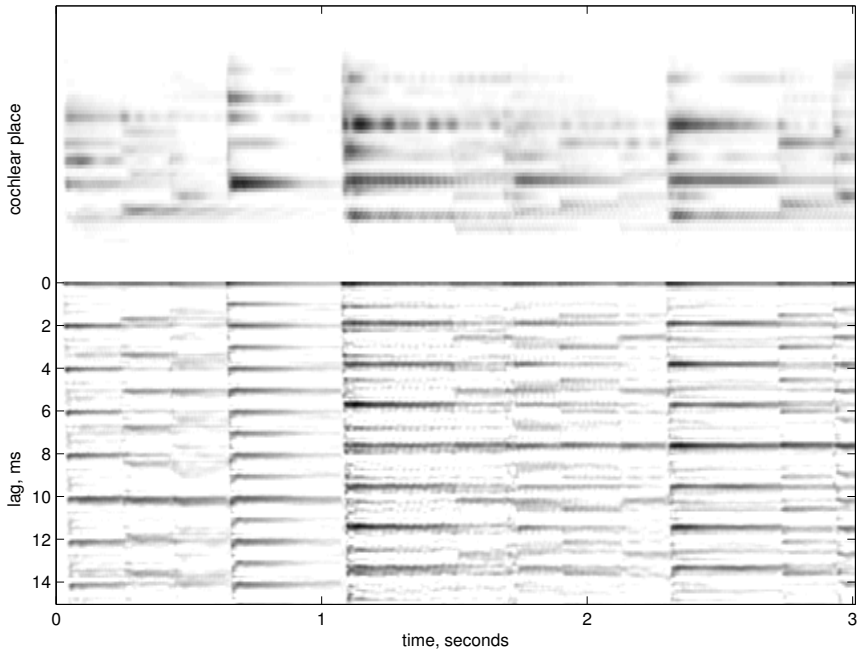


Figure 21.19: Combination cochleagram and pitchogram of 3 s of piano music. The period relationships that correspond to harmonic chords and consonant pitch intervals are more apparent in the lower part, the pitchogram, than in the cochleagram.

21.11 Summary SAI Tracks: Pitchograms

Time-stacking the SAI averages along rows makes a sort of auditory spectrogram, or cochleagram, while stacking the averages along columns, the summary SAIs, makes what we call a "pitchogram." Each of these is useful for visualizing the temporal evolution of a sound, though together they are still not as complete as an SAI movie representation. Figure 21.19 and Figure 21.20 show combination cochleagram and pitchogram plots for short segments of piano music and guitar music, respectively. Figure 21.21 shows a corresponding plot for a few seconds of speech.

For speech, a clear pitch track is usually evident in the pitchogram (along with tracks at multiples of the pitch period). For music, with multiple harmonically related pitches present at once, the pitchogram is usually more complicated, and more structured. Other sounds make their own characteristic pictures.

21.12 Cochleagram from SAI

The cochleagram formed by averaging along rows of a triggered-temporal-integration (TTI) SAI is very similar to the cochleagram formed by simply smoothing the NAP, since each row of the SAI is just a time-shifted NAP row, or an average of several.

For an SAI formed by conventional short-time autocorrelation, the zero-lag column is a good spectrum estimate, as it represents the short-time power in cochlear channel outputs. But the zero-lag column of the TTI SAI is not a very good spectrum estimate. It reflects the peak amplitudes, where the trigger events are, but does not reflect the fact that channels matched to the spectral peaks or resonances of the sound have a more sustained output with higher average relative to peak value. The higher power associated with sustained ringing is easy to see in the original NAP and in the TTI SAI, but is not picked up in the zero-lag column.

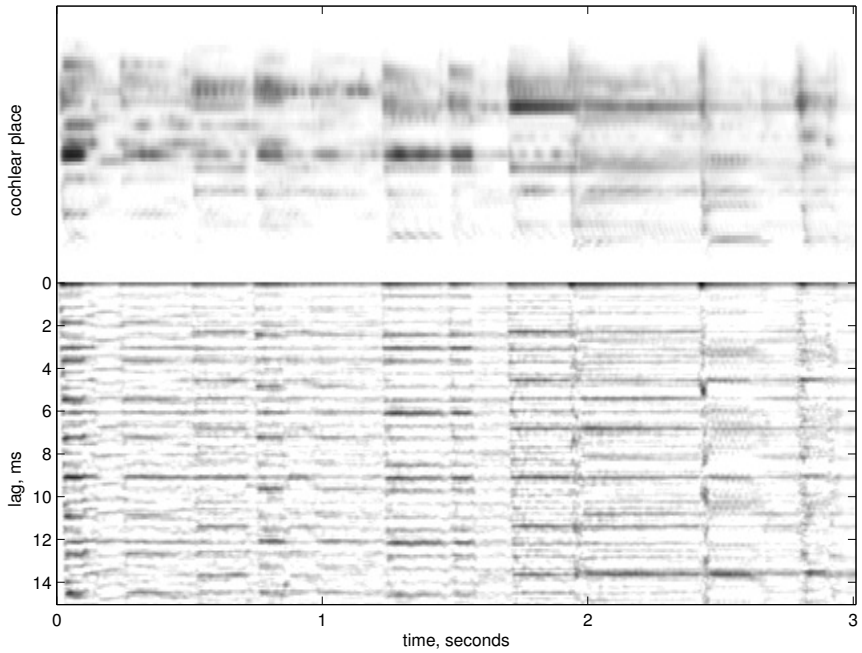


Figure 21.20: Combination cochleagram and pitchogram of 3 s of guitar music. A common period of 9 ms is apparent for many of the notes and chords.

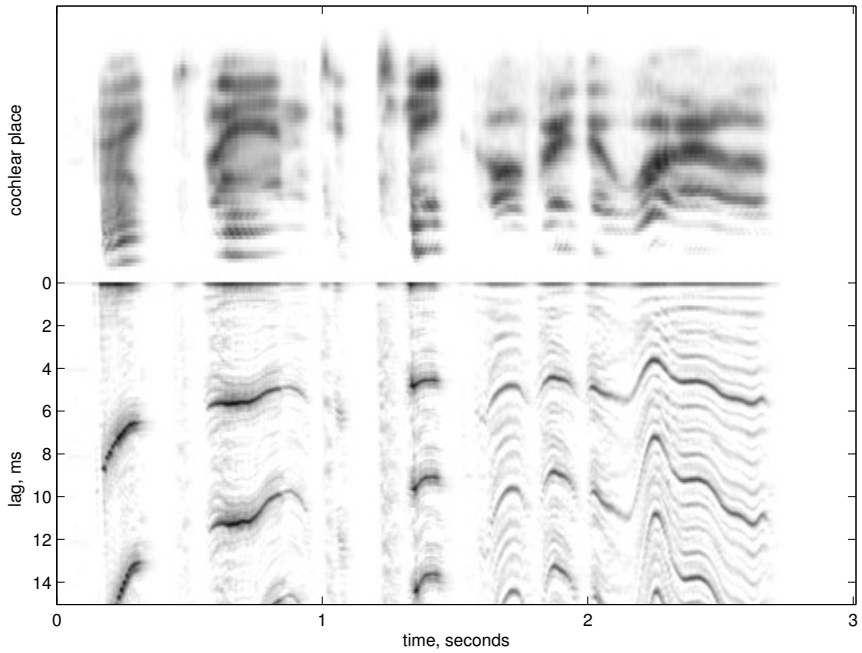


Figure 21.21: Combination cochleagram and pitchogram of 3 s of speech. The top part, the cochleagram, resembles a speech spectrogram, while the pitchogram on the bottom clearly shows the pitch contours of the voiced segments.

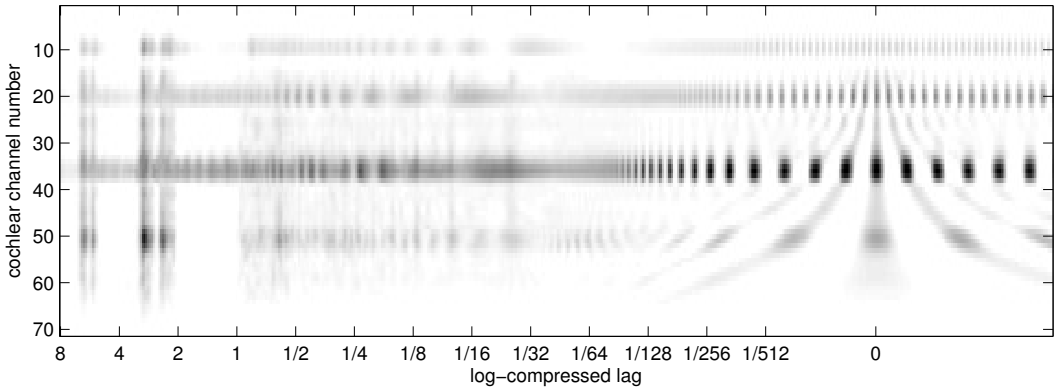


Figure 21.22: An SAI frame of a telephone ringing (file BelgiqueBellPhone.mp3 from freesound.org), with longer lags logarithmically compressed. The telephone sound exhibits structure on many time scales, out to its 3 s period and 6 s double period, and has somewhat different structure in different frequency regions.

That is why an average along rows of a TTI SAI is useful, like an average along columns is for a pitchogram. However, cutting off that average at some arbitrary number of columns will create an undue interaction of the resulting value with the pitch, as multiples of the pitch period may be just inside or just outside the range averaged. Therefore, a weight window that is maximum at zero lag and declines to zero weight at the extreme lag end(s) of the SAI is useful in forming a better spectrum-like feature.

While averaging along rows of an SAI makes some sense for generating a spectrum-like feature, it is simpler to just average the NAP directly, which is what we have actually done for the cochleagram part of the images in this chapter.

21.13 The Log-Lag SAI

The time-delay (τ) or lag dimension of the SAI can potentially be extended to very long lags, but having the number of columns in the SAI movie grow linearly with the maximum lag is not convenient. So we sometimes nonlinearly warp, or resample, the lag axis. To achieve a nearly constant percentage delay change per sample, the axis will be distorted such that the column position is proportional to the log of the lag, for large lag. We call this a log-lag SAI.

If the τ axis is resampled to make an approximate $\log(\tau)$ axis extending to long delays, with appropriate smoothing in the resampling to prevent aliasing, the finest time structure in the SAI will be smoothed away. The longer-time correlations that represent texture and rhythm will not show up well if the trigger impulses come at a constant average rate and constant size; some average amplitude information needs to be kept. That is, if we wish to see good approximations to autocorrelation over long time delays, then we need to either keep the amplitude scaling, or produce trigger impulses at a rate that reflects the local average amplitude.

One way we have made log-lag SAIs is to generate SAIs with multiple scales of pre-smoothed and decimated NAPs, and then warp and blend the corresponding different scales of SAIs. At longer lags, we use more pre-smoothing, and then generate a few trigger points per longer segment. At lags near 1 s that represent beat and rhythm, we only generate a few trigger points per second, so the rhythm information shows up well in the same framework that shows pitch, though the repetition rates are several orders of magnitude apart.

At long lags, the summary SAI (the log-lag extended pitchogram) resembles the *rhythmogram* introduced by Jensen (2005, 2007). At the coarser scales, the time-smoothed NAP channels that we detect trigger events in resemble Jensen's *perceptual spectral flux* (PSF) feature that signals note onsets at its peaks, except

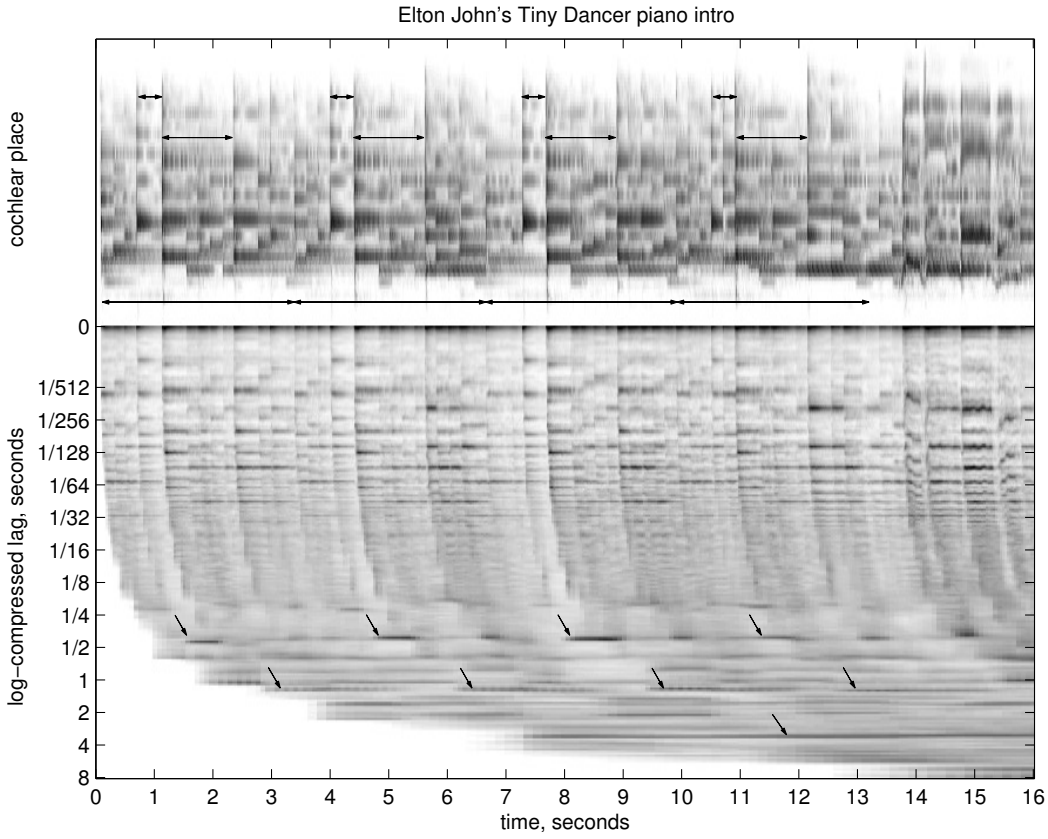


Figure 21.23: Combination cochleagram and log-lag pitchogram/rhythmogram of the piano opening of Elton John's "Tiny Dancer," with vocals in the final 2 s. Prominent time intervals between strong chord onsets, indicated by arrows from top to bottom, correspond to the durations of eighth notes (0.4 s) and of three eighth notes (1.2 s); measures (3.2 s) are also marked. The sweeping curves at the bottom reflect the delay of causal buffering, correlation, and averaging at exponentially increasing time scales.

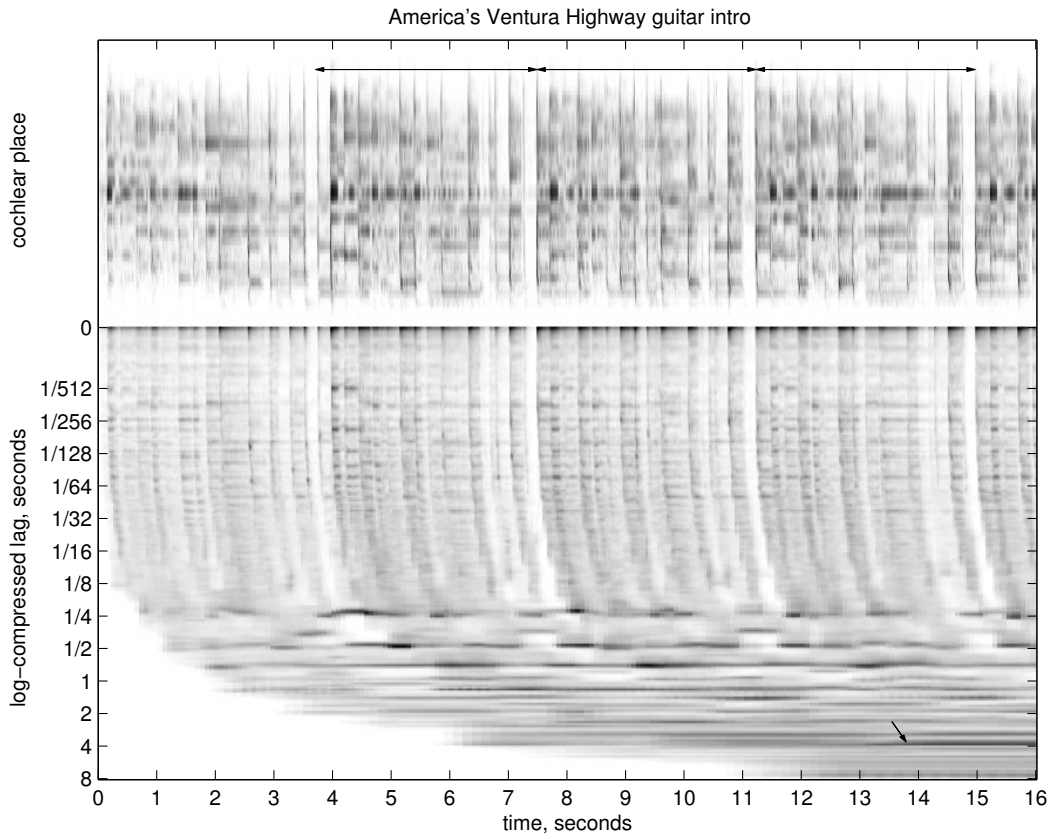


Figure 21.24: Combination cochleagram and log-lag pitchogram/rhythmogram of the popular opening guitar riff of America's "Ventura Highway." Prominent time intervals of one-quarter, one-half, and three-quarters seconds between notes show up, as does the phrase repetition near 4 s (indicated by arrows), but there is not much at 1 or 2 s. These time patterns summarize the rhythmic structure of the riff.

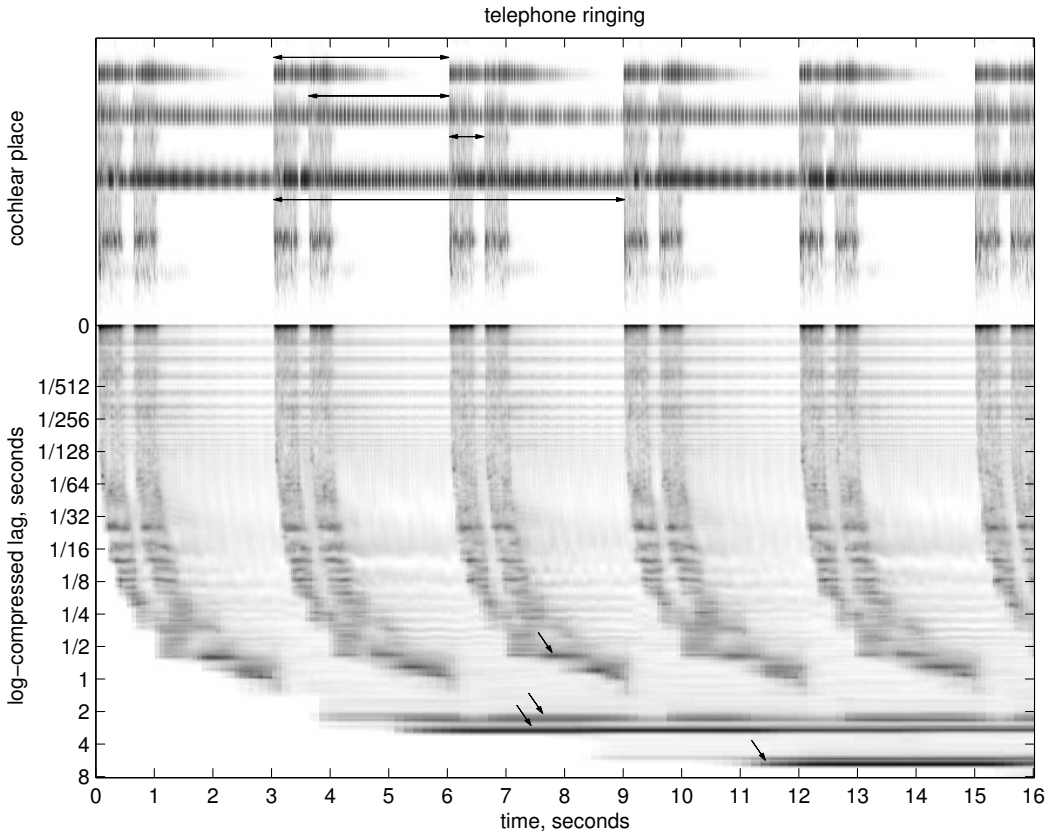


Figure 21.25: Combination cochleagram and log-lag pitchogram/rhythmogram of 16 s of the telephone ringing of Figure 21.22; two rings 0.7 s apart, every 3 s. The lag region that is not much used in music, between pitch periods and beat periods, is here filled with the pattern of the telephone's clapper intervals, at about 1/32 to 1/8 s. The rhythmic 0.7 s, 2.25 s, 3 s, and 6 s intervals are also prominent, as indicated by arrows.

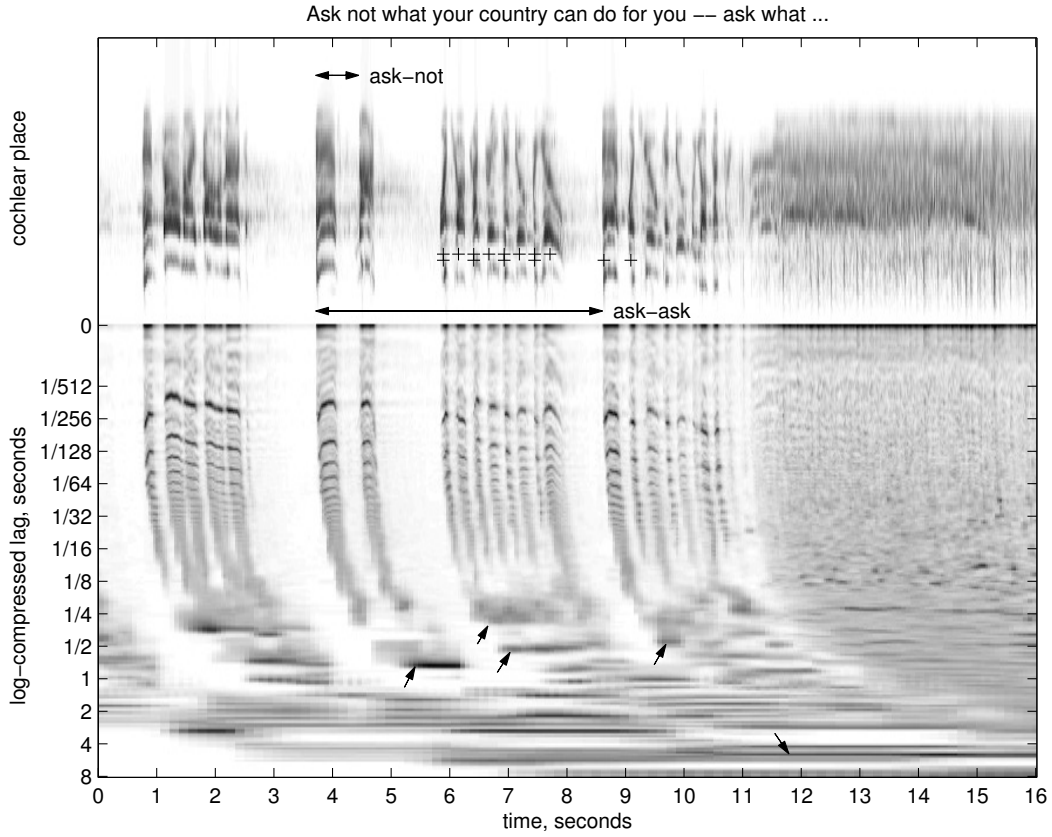


Figure 21.26: Combination cochleagram and log-lag pitchogram/rhythmogram of 16 s of John F. Kennedy's inaugural address: "And so my fellow Americans: Ask not what your country can do for you—ask what you can do for your country [applause]." English tends to have a regular cadence, but in oratory like this the rhythm is greatly altered. The interval between the syllables *ask–not* is lengthened to about 0.75 s, as marked on the plot, while the syllables in *what your country can do for you* come at about four per second, slightly syncopated, with a strong regularity at 0.5 s; these regular intervals are marked by crosses and arrows, as is the 0.4 s *ask–what* interval. The 5 s interval between the two occurrences of *ask* after a pause shows up relatively precisely.

that we keep different versions in different frequency channels. We call the extended log-lag pitchogram a pitchogram/rhythmogram.

Figure 21.23 and Figure 21.24 show combination cochleagram and log-lag pitchogram/rhythmograms for musical pieces, piano and guitar, respectively. Figure 21.25 shows how the telephone of Figure 21.22 shows up in this representation. Notice that the musical pieces have little structure in the middle "roughness" lag range between pitch periods and beat periods, about 1/32 to 1/8 s, but that the telephone ringing uses that range to attract attention. Finally, Figure 21.26 shows a few seconds of speech, from the most famous line of JFK's inaugural address. Prominent timing patterns show up clearly in the rhythmogram region, and the roughness of the applause shows up at the end.

Chapter 22

Binaural Spatial Hearing

> A failure to distinguish between phase expressed as angle, ϕ, and phase expressed as time, t', has led to some confusion in the literature.
>
> — "A place theory of sound localization," Lloyd A. Jeffress (1948)

> To anyone who is familiar with our modern knowledge of room acoustics, one of the most puzzling questions must be how it is that sounds can be localized at all in a reverberant room, let alone heard in the rather precise positions that are often reported.
>
> — "The precedence effect in sound localization," Wallach, Newman, and Rosenzweig (1949)

The binaural auditory system is essentially the *where* pathway: the brain's mechanism for figuring out where things are happening around the person or animal, based on comparing the sounds arriving at the two ears. The lower levels of the binaural system are probably the best characterized of any processing in the auditory brain. Neural signals from the two ears, after one stop in the *cochlear nucleus*, come together and interact in the *olivary complex*. Here, differences between the signals from the two cochleae are extracted, and put into a form that higher levels can handle.

The binaural system doesn't just localize sound—it also helps us interpret sound mixtures that include signals from different directions. For example, speech in a reverberant room is easier to understand by listening to a binaural recording than by listening to just one channel (Gelfand and Hochberg, 1976). Exactly how binaural directional cues are exploited to enhance the accuracy of the processing in the *what* pathway is not yet well understood, but there are many good clues.

One of the important lessons in the history of ideas about binaural processing is that we should not confuse a time difference with a phase difference—note the chapter opening quote by Jeffress. Unfortunately, this problem persists to this day; so we explore the distinction, and the history, in some depth in this chapter.

22.1 Rayleigh's Duplex Theory: Interaural Level and Phase

In the 1870s, Lord Rayleigh (Strutt, 1876, 1877) published his experimental observations on the ease of localizing all kinds of sounds. He pointed out that it is easy to tell if a sound is from the left or the right, relative to the head, and made several other key observations and calculations. In particular, he calculated the intensity differences that would be expected from a head-shadow effect, for sounds from off-center directions, as a function of frequency, and showed that this intensity difference between the two ears is a cue with good explanatory power, except for very low-frequency tones (below about 200 Hz). Such low frequencies have wavelengths that are large compared to the head, so his diffraction calculations showed there is not much

head-shadow effect; nevertheless, he found that these tones were very easy to lateralize. What remained a mystery was what cue we could be using to do so.

Thirty years later, he published a follow-up paper (Strutt, 1907), in which he proposed that a *phase difference* cue was the explanation for our ability to lateralize low frequencies. His theory that includes both the intensity difference and phase difference cues has come to be called the *duplex theory of binaural localization*. He theorized that phase-difference cues dominate at low frequencies, while intensity-difference cues dominate at high frequencies. He was mostly working with pure tones, because the tuning fork, a source of fairly pure sinusoidal sound waves, was one of the premier instruments among acousticians of the time, and because it was generally assumed that analyzing systems in terms of sinusoidal inputs was a good idea. He calculated that at about 640 Hz, a sound from the far left or right would arrive at the two ears 180 degrees out of phase, so the phase would not distinguish left from right in this case. Because of this ambiguity, Rayleigh concluded that phase could not be a good cue at 640 Hz and higher frequencies:

> Thus, although there might be right and left sensations from sources obliquely situated, these sensations would fail when most needed, that is when the source is really in the line of the ears. In this case a perception of phase-differences would seem to do more harm than good. At a pitch a little higher, ambiguities of a misleading and dangerous kind would necessarily enter. For example, the same sensations might arise from a sound a little on the left and from another fully on the right.
>
> On the whole it appears that the sensation of lateralness due to phase-difference disappears in the region of pitch where there would be danger of its becoming a misleading guide. ... It is fortunate that when difference of phase fails, difference of intensity comes to our aid.

This "duplex" conception of two types of comparisons of the signals from the two ears is still a standard model of how binaural hearing works, though the concept of interaural phase difference (IPD) is usually replaced by interaural time difference (ITD), which works more generally than just for sinusoids. The interaural level difference (ILD) cue has been a fairly stable idea, though generalizing it as interaural spectrum comparison along an independent frequency dimension makes it much more complicated, interesting, and informative.

In the decades following Rayleigh's first observations, the Helmholtz phase-blind Fourier-analysis conception of the cochlea seems to have suppressed thinking about phase and time differences, until Rayleigh (Strutt, 1907) eventually took the step himself to the phase-difference cue, a precursor to the more explanatory time-difference cue. In the book *Studies in Auditory and Visual Space Perception* by Arthur Henry Pierce (1901), there was still nothing about the fact that sound waves from the side must reach one ear before the other. Pierce examined existing theories in depth, and put a tremendous effort into evaluating and comparing the intensity-difference and semicircular-canal theories of binaural localization—the latter were based on the idea that these little-understood parts of the inner ear must be doing something useful with respect to sound. The concepts of phase and time difference never came up, because Helmholtz's model of the cochlea as extracting the magnitudes of Fourier components still permeated his thinking.

The three "History Connection" boxes in this chapter examine some of the trials and tribulations of more than a century in getting to an accepted view of how ITD works. We are still not quite finished with this difficult journey.

22.2 Interaural Time and Level Differences

To first order, the cues that binaural spatial hearing uses are the differences in the times of arrival of sounds at the two ears, and the differences in the sound levels at the two ears: the interaural time difference (ITD)

History Connection: Phase "Unwelcome"

The idea that the phases of the signals at the two ears could interact had been examined in the context of *binaural beats*, and had been discarded, with the conclusion that the effects must instead be due to intracranial sound propagation.

Sylvanus P. Thompson (1877) reported observing these beats between tones presented to the two ears. His hypothesis that "the tone-stimulus is transmitted along each auditory nerve to some common cerebral centre and that at this centre the beats arise" was ignored or rejected at the time, and was "unwelcome" even after Rayleigh's duplex theory. For example, Wilson and Myers (1908) critically assess "the influence of binaural phase differences":

> The importance of the transmission of stimuli by bone conduction from ear to ear is well seen in an experiment described by Thompson, in which two tones were generated in different rooms and were led by tubes, one to each ear of the observer. These tones were produced from two tuning-forks, having a pitch of 246 and 256 vibrations per second, respectively. Under these conditions, as is well known, beats are audible, just as if the two tones were presented to a single ear. Thompson concluded that under the conditions of binaural hearing above described, the tone-stimulus is transmitted along each auditory nerve to some common cerebral centre and that at this centre the beats arise. But this and the following interesting fact, also observed by Thompson, can be explained without recourse to such an unwelcome hypothesis, if we suppose that each tone is transmitted by bone conduction to the opposite ear and that the beats heard are due to the play of the two series of vibrations of different frequency on one and the same sense organ.

Wilson and Myers had discussed the possibility that the phase sensitivity implied that the auditory nerve carried the waveform directly; but then rejected that idea:

> In the case of vibrations of sound,—despite the fact that they are much slower than vibrations of light,—it is nevertheless just as difficult to suppose that such characters of the stimulus are actually communicable to the auditory nervous impulse. It is very hard to believe that every crest and every trough of each sound wave produce an exactly corresponding crest and trough in the impulses transmitted along each auditory nerve. Were this so, we could no longer regard the sensory nerve as an intermediary, knowing no more of the exact nature of the external stimulus than the telegraph wire knows of the mental processes of the operator who transmits the telegram. We could no longer regard the sensory nerve impulse as being determined solely by the method of response of the end organ with which it is connected.
>
> We hope to show that such a radical change in our views is unnecessary.

Eventually, researchers had to accept such a radical change, and admit that the nerves do convey waveform details to central sites for comparison. An accumulation of evidence, and Wever and Bray's 1930 *volley theory* that we discussed in Chapter 2, made the ideas less "unwelcome."

In the meantime, however, there were still plenty of efforts to explain binaural hearing without any central phase comparison, relying on level differences alone. One of these, by Henry J. Watt (1920), actually proposed an auditory-image-like model: a two-dimensional pattern of activity reflecting tonotopic organization along one dimension, and interaural intensity relationship along the other, narrower, dimension of neural tissue.

History Connection: From IPD to ITD

Rayleigh (Strutt, 1877) had made an observation that practically begged for someone to propose a cue based on the different times of arrival of sounds at the two ears:

> When one ear is stopped, mistakes are made between [tuning] forks right and left; but the direction of other sounds, such as those produced by clapping hands or by the voice, is often told much better than might have been expected.

The conceptual replacement of the phase differences of sine waves ("pure tones" as the acousticians always glorified them) with the time difference of more general or transient sounds started about 1908, but took a long time to catch on. Mallock (1908) noticed the strange apparent direction of the sound of a bullet, due to its bow shock wave (its miniature sonic boom), and after some experiments and analysis concluded (see Figure 22.1):

> A sound which is caused by the detached waves, such as those which accompany a bullet, can scarcely be said to have a pitch, but the wave-length is certainly small compared with the distance between the ears, and is indeed comparable with the dimensions of the bullet itself. It would seem, therefore, that the ears can determine the direction of a sound, not only by difference of phase, but by the actual difference in the times at which a single pulse reaches them.

He found a consistency of observations to within a few degrees of the calculated wavefront direction (corresponding to a time-difference error of a few tens of microseconds). Then Hornbostel and Wertheimer (1920) found time differences between clicks to be effective down to 30 microseconds, and even smaller "under favorable conditions." Otto Klemm (1920) published much more detailed experimental results, also in 1920, and found a time-difference threshold of about 20 microseconds in one subject, and even less than 10 microseconds in another! Several researchers published ITD thresholds in 1921: Aggazzotti (1921) found 70 microseconds, and Pérot (1921) found between 55 and 80 microseconds (and more at lower levels).

In spite of all these investigations, the paradigm shift to thinking of ITD as a genuine cue was slow to come. For example, Hartley and Fry (1922) analyzed the localization of complex tones, but interpreted it all in terms of the independent localization of sinusoidal components by their phase and amplitude differences.

While the ideas were still being debated, during World War I scientists on both sides were putting the "binaural sense" to use in military applications, for finding the directions to airplanes and submarines and tunnel diggers, using human listeners with binaural sound horns and time-delay compensators to steer their listening direction (Yerkes, 1920; Drysdale, 1920; Ferry, 1921). The 1920 report by Hornbostel and Wertheimer (1920) was based on their German wartime devices. They filed for a patent in 1915 on the "Richtungshörer" (directional listener) that used a wide spacing between listening horns to exaggerate the time difference cue (King and Wertheimer, 2007); similar to the one in Figure 22.2, it was popularly known for its inventors as the "Wertbostel."

In 1931, Erich von Hornbostel (1931) restated his "time theory," as part of a "discussion on audition," a discussion in which others argued as strongly against it. He pointed out that phase was a poor alternative to absolute time difference. He tried to push the field away from the overreliance on tones, saying, "Theory, and also experiment, must take into consideration the fact that noises are more important in life than musical sounds (complex and pure tones) which are of rare occurrence in Nature; the fact, therefore, that noises are better localized than tones is a useful one."

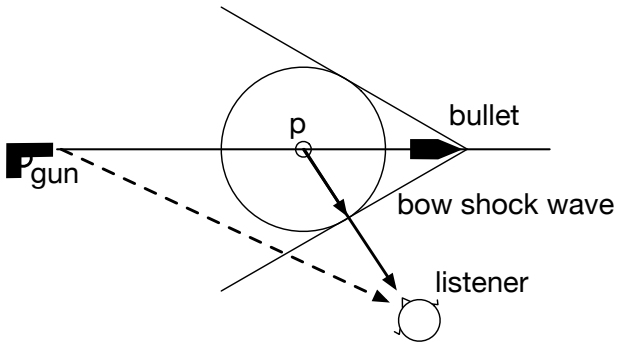


Figure 22.1: Mallock (1908) noticed that to a down-range listener, the crack of a supersonic bullet seemed to come from a direction different from the direction of the gun. The bullet's "sonic boom" or detached bow shock wave, moves at the speed of sound, while its apex at the bullet moves faster than sound, resulting in a wave angle as shown. When a two-eared listener is facing normal to the wave, it arrives at both ears at the same time, resulting in the apparent direction toward the point p.

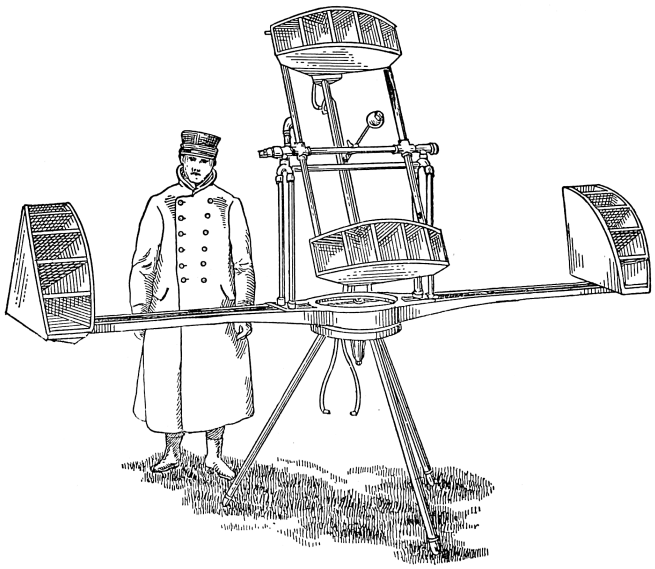


Figure 22.2: A World War I era acoustic goniometer for locating "invisible aeroplanes" (Ferry, 1921). Operators rotate pairs of pickup horns about vertical and horizontal axes, until a sound seems to be straight ahead. The device is clearly designed to exaggerate an interaural time difference cue.

History Connection: Getting Away from a Focus on Sinusoids

In 1936, while acknowledging that "phase difference is but a special case of time difference," Stevens and Newman (1936) concluded "that the localization of low tones is made on the basis of phase-differences at the two ears, and that the localization of high tones is made on the basis of intensitive differences. There is a band of intermediate frequencies near 3000 cycles in which neither phase nor intensity is very effective and in which localization is poorest."

The experiments of Scherer (1959) compared the ability to lateralize sinusoids, versus broadband signals, based on an ITD cue. Rayleigh had shown that the ITD cue for sinusoids, that is, a phase difference, can be totally ambiguous above about 640 Hz; but Scherer found that the ability to detect the insertion of a 20 μs delay is only reduced a little at 800 Hz, and falls gradually between about 800 and 1600 Hz. By 1600 Hz, with sine waves, his subjects had no ability to distinguish 0 from 20 μs ITD. But with noise filtered to a band around 1600 Hz, or even 3000 Hz, his subject were just as good at detecting the 20 μs ITD as at lower frequencies.

In light of Hornbostel's and Scherer's results, we know that the dip around 2–4 kHz that Stevens observed is purely an artifact of using tones, with their inherent cyclic time ambiguity, like those annoying beepers on carts in airports that you never notice coming up behind you because their beeps can't be localized. For more typical sounds, we localize quite well in that frequency range.

When Jeffress (1948) looked at the science, he concluded, "We may therefore reasonably assume that the basis for our ability to localize clicks and low frequency tones is the time difference." He wasn't denying that intensity difference is important, too, but made the point that time difference works across the frequency range, at least for sounds that contain time-localized events, as clicks do.

Even noises are not always so easy to localize; the normal or Gaussian amplitude distribution typical of some noises has few strong "outlier" features to drive robust localization. For sounds with "long tail" amplitude distributions or otherwise strongly fluctuating envelopes, frequent distinctive events in the tail of the amplitude distribution provide especially good points to localize based on time difference, using "envelope cues" and "onset cues" (Kollmeier et al., 2008). Onsets of pitch pulses in spoken vowels are such points. As McFadden and Pasanen (1976) say, "the auditory system obviously can be just as sensitive to this temporal difference at high frequencies as it is to cycle-by-cycle differences at low frequencies."

In spite of observations of precise neural synchrony to onsets and the ease of lateralizing sounds with transients, we still see papers stating that ITD works as a cue only at low frequencies, or that we're not very good at localizing in the 2–4 kHz region. These statements are correct, but only when applied to sinusoids or narrow-band signals; they can be compared to the view of *Flatlanders* (Abbott, 1884), people whose world is missing a dimension, limited to the infinitesimal slice of the sound space defined by sine waves.

Commenting on the brain area that extracts ITD cues, Karino et al. (2011) betray the usual preconception that ITD should be dominated by lower frequencies, as in Rayleigh's original conception: "Surprisingly, the tonotopic distribution of the afferent endings indicate that low characteristic frequencies are under-represented rather than over-represented in the MSO." Hopefully, we will get over being surprised that ITD is important at higher frequencies.

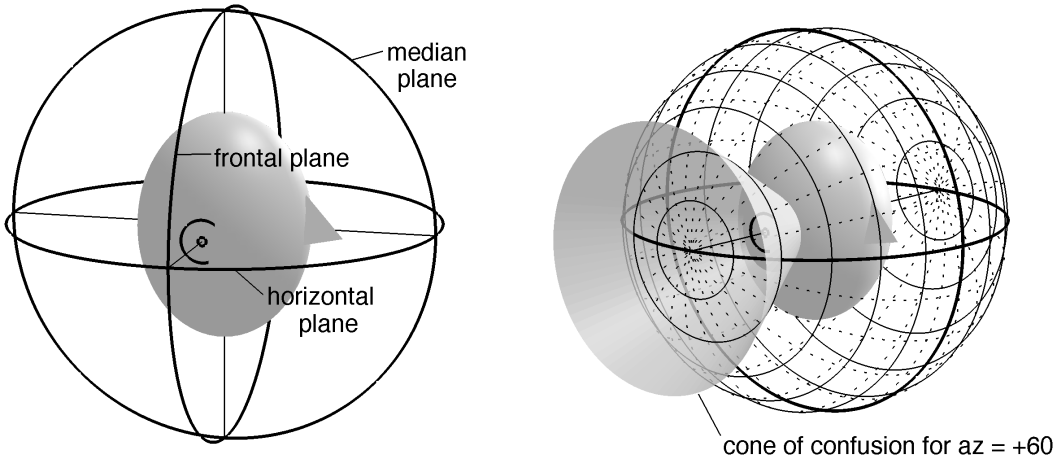


Figure 22.3: Sound directions are often described within three planes, depicted on the left by circles that determine the planes around an ellipsoidal head, aligned with the axes through the ears and the up–down and front–back directions. On the right, the *interaural-polar coordinate system* (Brown and Duda, 1998) is shown. In this system, elevation is like longitude, the angle about the polar axis between the ears, measured from the prime meridian in the horizontal plane; lines of constant elevation are shown dotted. Azimuth is like latitude, measured from the equator in the median plane; circles of constant azimuth are shown solid. One "cone of confusion" is also shown: the set of sound directions with a constant azimuth, or approximately a constant ITD.

and interaural level difference (ILD, also known as interaural intensity difference). Both ITD and ILD are functions of frequency, and can be thought of as being sensed separately in different cochlear filter channels (or groups of several nearby channels) before being fused into a percept of direction. These cues may also be changing in time, even if there is only one sound source present, as the source moves, or as echos from the floor, walls, the torso, other objects and people, etc., follow the onset of a direct sound from a source.

The binaural hearing problem can thus be seen as mostly the problem of how to extract, and how to combine, ITD and ILD cues. As part of the auditory-image approach, these cues are assumed to be represented in organized maps. Patterns in these maps are associated with directions in space. Azimuth is not alone in affecting the patterns of ITD and ILD versus frequency; elevation and front–back differences also make distinctive patterns.

A consistent spatial percept corresponding to an external sound source requires a consistent combination of these cues. For lateralization, or determining how far a sound source is from the median plane (see Figure 22.3), the ITD alone is a good cue, determining a left–right sense but leaving source elevation ambiguous. In recognition of ITD as the primary cue, we typically use the interaural-polar coordinate system shown in that figure, which associates ITD, or path-length difference, with azimuth (approximately, assuming a spherical head). The other dimension, elevation, or angle around a cone of confusion, then corresponds to more subtle frequency-dependent cues.

22.3 The Head-Related Transfer Function

The time difference in a wavefront's arrival at the two ears is actually somewhat frequency dependent, due to the effects of sound diffraction around the head. At high frequencies, the time difference corresponds closely to the path-length difference, including a curved piece of path wrapping around the head to the far ear, as illustrated in Figure 22.4. For a sound in the horizontal plane at an azimuth angle θ, traveling around

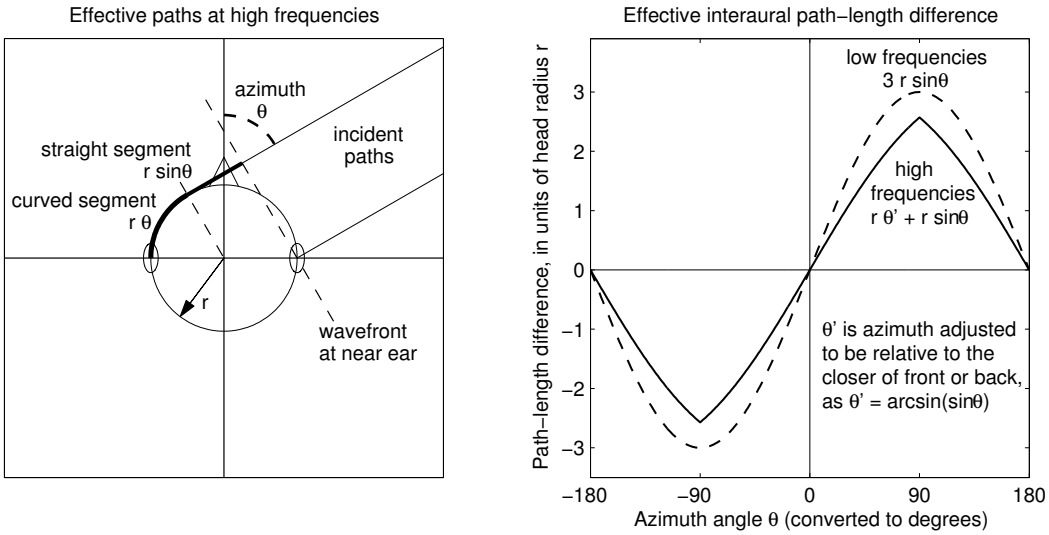


Figure 22.4: High-frequency sounds follow the shortest path around the head. In the horizontal plane, for sounds incident from an azimuth angle of θ, the extra path length to the far ear is $r\theta + r\sin\theta$ (using the azimuth angle modified as shown to be an angle of magnitude less than 90 degrees in the interaural-polar system). Due to diffraction effects, low-frequency sounds experience a somewhat larger time lag than this estimate would suggest; about 50% larger for small angles (Kuhn, 1977). The angle illustrated on the left is on the +60 degree cone of confusion shown in Figure 22.3.

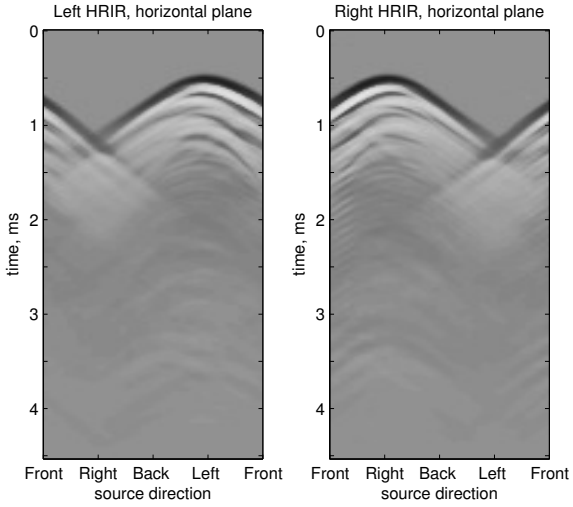


Figure 22.5: Head-related impulse responses of a dummy head, for sounds from various directions in the horizontal plane. The impulse responses are mapped to gray levels and displayed from top to bottom, from an arbitrary time origin shortly before the arrival of a sound impulse at the head; the x axis represents the azimuth angle of the sound source. Separate arrivals can be seen when the sound goes around the front and back of the head as in the left ear when the sound is from the right. For a sound from an intermediate azimuth (not a multiple of 90 degrees), differences in pinna echos, the ridges near the top of the plot, help distinguish front from back.

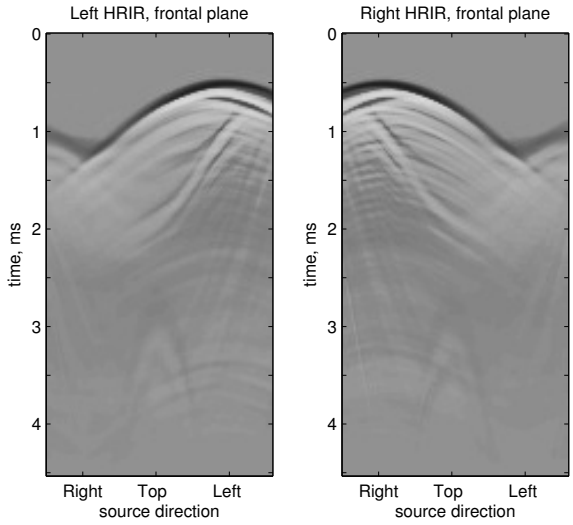


Figure 22.6: Head-related impulse responses of a dummy head, for sounds from various directions in the frontal plane. The x axis represents the angle of the sound source, from low on one side, over the top of the head, to low on the other side. The steeper patterns near the center, in the 1–2 ms range, represent shoulder-bounce arrivals, mostly at the ipsilateral ear, up to about 1 ms delayed from the main arrival (Brown and Duda, 1998).

a spherical head of radius r, this path-length difference is $r(\theta + \sin\theta)$ (for θ in radians). So the ITD is near $r(\theta + \sin\theta)/c$ at high frequencies, where c is the speed of sound. At low frequencies, however, the ITD is closer to $3r\sin\theta/c$, about 50% more than the high-frequency ITD at small angles; the frequency transition between these different delays is gradual, between about 400 and 2000 Hz for a human head (Kuhn, 1977). At low frequencies, the maximum ITD is nearly 1 ms; the frequency at which it corresponds to a half cycle is very near the 640 Hz that Rayleigh had calculated.

The sound spectrum and the ILD–frequency pattern encode elevation and front–back cues, as they are affected by how the sound waves interact with the floor, torso, shoulders, pinnae, and head shape, among other things. The acoustics of such sound propagation around the head and into the ears has been well studied and modeled (Wenzel et al., 1993; Duda et al., 1999). Acoustic propagation is nearly perfectly linear, so it can be characterized in terms of a head-related transfer function (HRTF). Every direction from which a wave may approach the listener has corresponding left-ear and right-ear filters, or transfer functions.

Figure 22.5 and Figure 22.6 show *head-related impulse responses* (HRIR, a convenient way to represent HRTFs, for two different paths that a sound source could take around the head; the data are from Algazi et al. (2001b). In this representation, the ITDs between near and far ears are the dominant motif in the image, but many more subtle cues are also apparent. Figure 22.7 shows both HRIRs and HRTFs for the median plane, where there are no interaural differences, but strong monaural cues for elevation. In general, such cues are helpful for resolving the elevation angle around a *cone of confusion*—a set of directions all having the same ITD.

Kulkarni and Colburn (1998) have studied listeners' ability to distinguish real from synthetically spatialized sounds in the median plane, using the listener's own HRTFs and smoothed versions of them. They found that moderate spectral detail in the HRTF filtering (for example, modeling the HRTF using only 32 cepstral coefficients) is sufficient to make synthetic sounds indistinguishable from real. Baumgartner et al. (2014) have modeled the particular spectral features needed for localization around the cones of confusion ("in sagittal planes"), and report that "positive spectral gradients"—that is, frequencies at which the spectrum

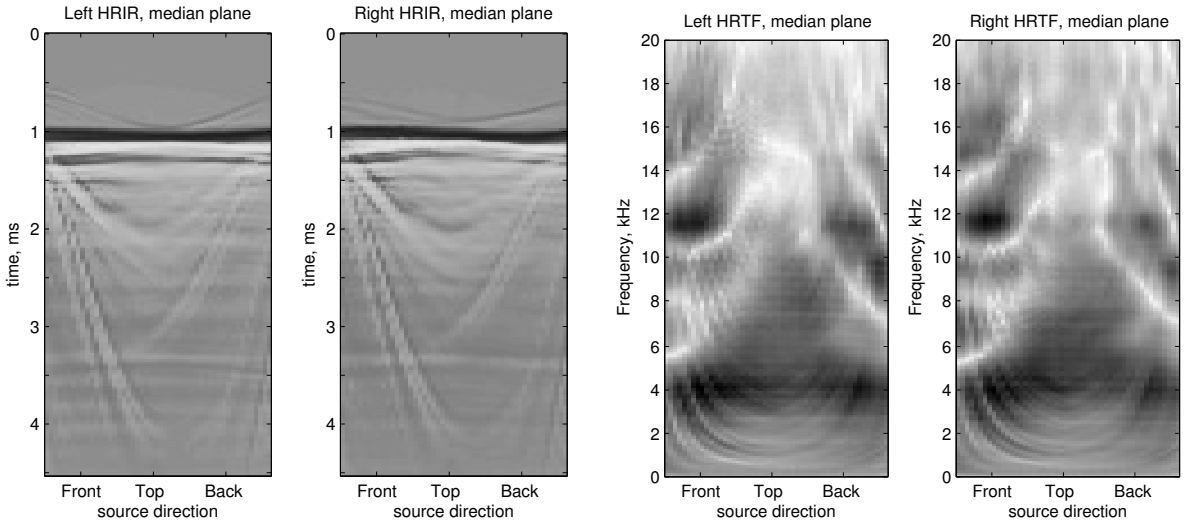


Figure 22.7: For sounds in the median plane, the responses of the two ears are essentially identical; the zero ITD and ILD cues indicate a sound straight ahead, or overhead, or straight behind, but there is no interaural cue for elevation angle. The HRIR (left panels) and HRTF (right panels) do show prominent elevation dependence, but the cue is essentially monaural, not based on a difference between the ears. The prominent spectral notches (white areas, since as always, we plot larger values as darker) above 5 kHz come from pinna diffraction, and the ripples below 4 kHz from torso (chest, shoulder, and back) echoes. Both provide useful cues to elevation (Algazi et al., 2001a), and both might be detectable via either temporal or spectral patterns. These data are from a real person, not a dummy head, which is why the signals at the two ears are not quite identical.

is sharply increasing along the low-to-high frequency dimension—make a good starting point for modeling human psychophysical data.

22.4 Neural Extraction of Interaural Differences

Jeffress's proposal for how the brain might detect and map ITD was shown schematically in Figure 2.1. A similar idea was as actually proposed in brief by Bowlker (1908), forty years earlier:

> In order to explain the existence of a movable image of the sound within this zone, we may suppose that the transmission of the sound impulse through some specialized part of the auditory apparatus or brain takes a definite time from each ear, and that the point where the impulses meet is the focus that gives rise to the sensation of a sound-image.

The signals from the two ears are compared where they first come together in the brainstem: in the olivary complex (OC) in mammals. The medial superior olive (MSO) extracts ITD by coincidence detection, in the manner described by Bowlker and Jeffress, and the lateral superior olive (LSO) extracts ILD by responding to excitatory input from one side sufficient to overcome inhibitory input from the other. Other parts of the OC, the lateral and medial nuclei of the trapezoid body (LNTB and MNTB) are also involved. The MNTB provides the inhibitory inputs to LSO; the LNTB perhaps provides secondary inhibition that helps to implement the precedence effect (Yin, 1994); see Section 22.7. The basic circuit involving the olivary complex and cochlear nuclei is as schematized in Figure 22.8.

Knowledge of neural circuits for binaural coincidence detection, or cross-correlation, has been reviewed recently, in the case of mammals (Joris et al., 1998) and birds (Konishi, 2003). In both cases, the function and

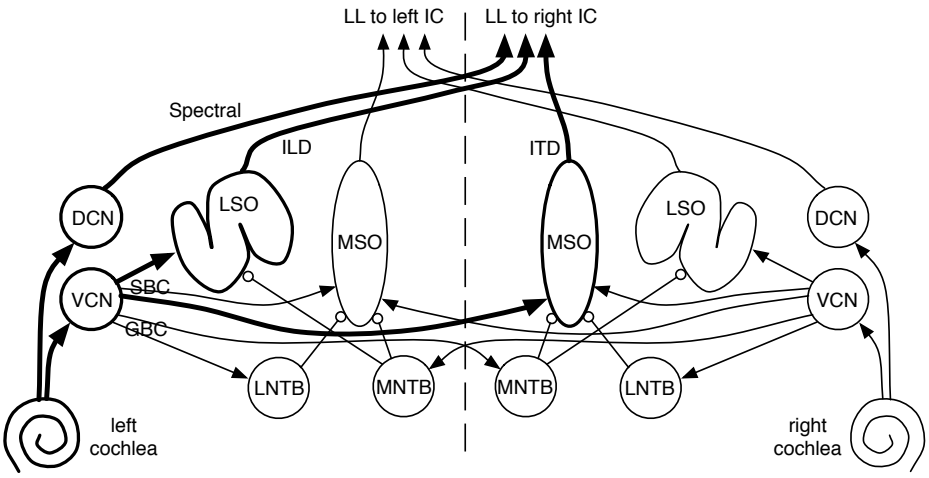


Figure 22.8: The ascending binaural circuits from the cochlea, through auditory nerve and ventral and dorsal cochlear nuclei (VCN and DCN), and through the olivary complex. Main excitatory pathways for a sound on the left are shown bold. A sound on the left primarily activates the left LSO and the right MSO, both of which project upward via the right lateral lemniscus (LL) and its nuclei to the right inferior colliculus (IC). Inhibitory connections are indicated with bubbles at their ends. The DCN is thought to provide spectral cues to IC, to help with vertical localization. Spectral, ILD, and ITD cues are probably integrated in IC. The division of VCN into AVCN and PVCN is not shown; the bushy cells (SBC and GBC) are mostly in AVCN.

organization are found to be largely as Jeffress predicted. In terms of the bilateral organization of brains, we localize sounds on the contralateral side. That is, a sound from the right causes neurons on the left to respond. This lateralization is accomplished by having relatively constant delays from the left cochlea through the left cochlear nucleus (CN—or nucleus magnocellularis, NM, in birds) to the left coincidence detecting neuron in the medial superior olive (MSO—or nucleus laminaris, NL, in birds), while fibers crossing from the right CN have varying axon lengths to contribute a range of delays before reaching their points of synapse in MSO. That is, the variations are in the longer delays, from the contralateral side, such that earlier signals from the contralateral ear coincide, when they arrive at MSO cells, with ipsilateral signals caused by sounds arriving later at the ipsilateral ear. For signals near the midline, there is some overlap in response between the two MSOs.

To a large extent, Jeffress's model is confirmed. From their various experiments with cells in the cat MSO, Yin et al. (1987) concluded:

> All of our results support the idea that the central binaural neurons perform an operation very similar to cross-correlation on the inputs arriving from each side. These inputs are transformed from the actual acoustic signal by the peripheral auditory system, and these transformations are reflected in the properties of the cross-correlations.

Yin and Chan (1990) confirmed spatial mapping of ITD in MSO, and tuning to envelope ITD at high CF. As Joris et al. (1998) say of circuits in the central nervous system (CNS):

> Physiologically, neurons in the MSO are sensitive to microsecond differences in their afferent signals. This is one of the few sensory circuits in the mammalian CNS for which a strong functional hypothesis can be formulated and the mechanisms underlying its physiological properties are relatively well understood.

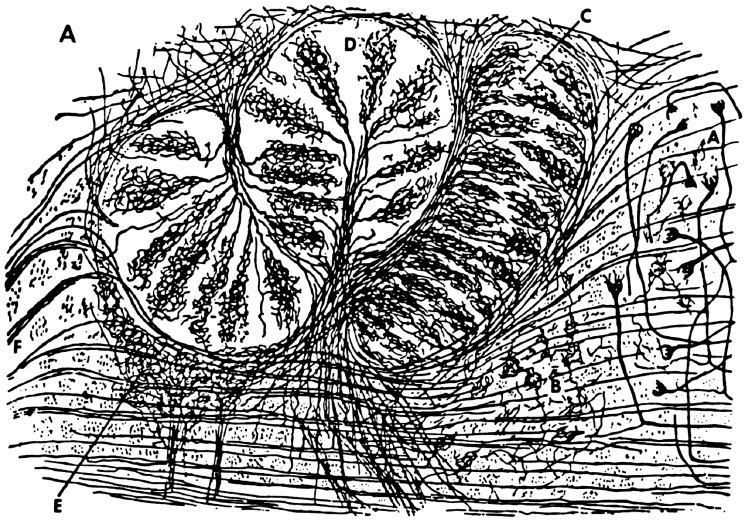


Figure 22.9: Superior olivary complex neurons as sketched by Cajal (1909), showing the S-shaped LSO and the sausage-shaped MSO, with trapezoid body neurons around them.

> ...
>
> Just as Hubel and Wiesel's model of orientation selectivity has provided a bull's eye for vision research, the simple model introduced by Jeffress has served as a focal point in audition and has sparked many computational and experimental studies, illustrating the heuristic value of such qualitative models in neuroscience.

Though the model is useful, and the mechanisms and functions are "relatively well understood," the details are not yet. The exact mechanism by which the MSO neurons preferentially fire on coincident inputs from the two sides is not well understood. It probably involves the addition of contributions to postsynaptic potential (PSP) from the two sides, using receptor channels with a very fast action and recovery. The PSP in response to a pulse from one side is not enough to reach the firing threshold, but two pulses, close enough in time, will make a supra-threshold PSP and cause the neuron to fire. Such an additive mechanism can be effective as a multiplier when the inputs are discrete pulses. The firings of MSO coincidence-detection neurons are then taken as input to subsequent processing, including probably aggregation across frequency channels, and smoothing in time.

An analysis by Grothe (2003) of the differences between the ITD extraction circuits of birds and mammals, which evolved independently, suggests that they employ rather different circuits, and use inhibition differently, deviating from the simple picture that Jeffress hypothesized. He concludes, "In mammals, exquisitely timed hyperpolarizing inhibition adjusts the temporal sensitivity of coincidence detector neurons to the physiologically relevant range of interaural time differences. Inhibition onto bird coincidence detectors, by contrast, is depolarizing and devoid of temporal information, providing a mechanism for gain control."

MSO neurons typically also have a moderate sensitivity to monaural sound, that is, to sound presented to only one ear. It is likely that the basic mechanism cannot distinguish signals from left versus right, and just requires enough nearly simultaneous input to trigger it. Another possibility that has been mentioned is that the random firings from one side may provide enough random coincidences with monaural signals from the other side to explain the monaural response. In this respect, it is not quite like the multiplier or left–right coincidence detector in Jeffress's circuit.

Circuits that extract ILD, in the lateral superior olive (LSO—or the nucleus angularis, NA, in birds)

work by excitatory–inhibitory (EI) neurons. The neurons are excited by ipsilateral signals, and inhibited by contralateral signals, so fire preferentially for ipsilateral sounds of sufficient ILD (unlike the MSO neurons, which usually prefer contralateral sounds). LSO neurons then project across to the contralateral inferior colliculus (IC), while MSO neurons project to IC on their own side, such that IC neurons combine both ITD and ILD cues and mostly prefer contralateral sound directions. There are also inhibitory projections from LSO to IC on the same side, which tend to suppress the IC response to contralateral sound when the LSO detects an ITD favoring the ipsilateral side.

According to Joris and Yin (1995), the LSO extracts not just level difference, but also envelope time differences, via its EI neurons; Tollin and Yin (2005) showed that LSO is also sensitive to interaural phase differences at low frequencies. These observations may explain why the LSO gets precisely timed inputs from the cochlear nuclei, like the MSO does. An LSO neuron fires when the excitatory input, from the ispilateral CN, is sufficiently stronger, or earlier, than the corresponding inhibitory input, from the contralateral CN via the MNTB. Thus, the LSO neuron will tend to fire more for ipsilateral sounds, and less for contralateral sounds. The inhibition works by *shunting*. That is, an inhibitory input effectively turns down the gain on the excitation, bleeding off the excitatory postsynaptic signal by opening a chloride-channel path that *shorts out* the dendrite of the LSO cell, about where it connects to the cell body (Zacksenhouse et al., 1998). If enough excitation gets to the cell body before the inhibition shorts it out, the cell fires. The effect is more nonlinear than taking a difference of excitation and inhibition. Whether this nonlinearity matters to practical modeling remains to be investigated.

22.5 The Role of the Cochlear Nucleus and the Trapezoid Body

The MSO and LSO get their inputs via *spherical bushy cells* (SBCs) and *globular bushy cells* (GBCs) of the AVCN. These cells are specialized to aggregate input from several primary fibers of the auditory nerve, and to preserve or enhance synchrony to the waveform, with appropriate emphasis on onsets.

Where the auditory nerve reaches the AVCN, the axons connect to SBCs through some of the largest synapses in the mammalian body: the *endbulbs of Held*. The GBCs receive inputs through somewhat smaller, but still large, endbulb synapses. Some of the GBCs project to the MNTB, where they connect through the ultimate large synapse, the *calyx of Held*; these synapses are so large that they surround the target (postsynaptic) neuron's cell body, in a cup-like or flower (calyx) shape; only one fits on a typical MNTB neuron. The endbulbs of Held are not quite so huge, fitting typically two on each target SBC. The size of these synapses is a clue to their function: they're fast. They support the low-latency and low-dispersion transmission of detailed timing information across the brainstem, so that the signals from the two ears can be compared for their relative arrival times.

The bushy cells are often characterized as "primary-like," based on the shapes of their post-stimulus-time histogram (PSTH) in response to tone bursts. Like the fibers in the auditory nerve, they tend to have a high firing probability right at the onset of a tone burst, falling off rapidly to a much lower steady-state level. But the bushy cells differ in the details. They synchronize to the phase of low-frequency tones better than the primaries do, almost always firing within a very narrow phase range; and at high enough levels, the bushy cells fire with near certainty at the onset of a tone burst, after which they may be refractory (not able to respond) for about a millisecond, resulting in a "notch" in the PSTH. When this notch is seen, which is usually in GBCs, they are sometimes characterized as "primary-like with notch," even when the existence of the notch is acknowledged to be level dependent.

The GBCs project to the contralateral MNTB, which in turn projects to the LSO on that side, and provides the inhibitory input to the EI cells there. The SBCs, on the other hand, do not usually exhibit the notch in their PSTH, probably because they do not have such a preferential firing at onsets; they provide the excitatory inputs to MSO for timing comparisons.

This binaural auditory circuitry is probably the most time-critical of any mammalian neural circuits. Each SBC probably tends to fire on receiving nearly-coincident input from two primary fibers. This nonlinear function, a monaural coincidence detector, is likely part of the strategy for getting well-timed pulses as input to the binaural coincidence detector in the MSO, where each cell receives excitatory inputs from just a few SBC axons. It is less clear why such fast GBC cell types and fast synapses are used in MNTB, in the path providing inhibitory input to the LSO.

There is not a lot of data on how these cells respond to signals more interesting than tone bursts and noises, but the evidence suggests that the SBCs are optimized to fire in precise synchrony to interesting events in the sound signal. For example, a recent study with broadband noise shows that the bushy cells show better synchrony to envelope features than the primary fibers do, as would be expected if they were doing a sort of monaural coincidence detection (Louage et al., 2005). Presumably, synchrony to events would be even more clear if the sound stimulus had more event-like structure of occasional outliers, as speech has for example. That is, if the sound has a more long-tail distribution compared to that of tones and steady noises.

As GBC neurons from AVCN project across the midline to the MNTB, others project ipsilaterally to the LNTB. Complicating the traditional simple picture of MSO taking excitatory inputs from SBCs from both sides, both the MNTB and LNTB provide inhibitory inputs to MSO on their own side (Grothe and Koch, 2011). As a result, the MSO has both excitatory and inhibitory inputs from both sides. The inhibition may act quickly enough to be part of the basic time-comparison mechanism, as Grothe and Koch (2011) suggests. That is, the coincidence detectors of the MSO may use the early arrival of precisely synchronized inhibitory inputs (from "primary-like with notch" GBCs via MNTB and LNTB) to sensitize the MSO cell to just the right combination of excitatory inputs (from "primary-like" SBCs) immediately thereafter. According to this theory, delay tuning can be done without the axonal delay lines proposed by Jeffress (Grothe et al., 2010; Grothe and Koch, 2011).

Joris and Yin (2007) reject the idea that the fast inhibition can replace axonal delay as the primary mechanism for ITD tuning. Whatever the mechanism turns out to be, the basic idea of Jeffress that ITD is extracted centrally from signals precisely synchronized to the waveforms at the two ears remains the basic model.

Further tuning of coincidence detection neurons, so they fire just often enough as the input levels and spike rates vary, may be accomplished by an adaptation process, either involving fast synaptic plasticity, as found in NL in birds (Cook et al., 2003), or via inhibitory feedback from the next layer, a common feature of many levels of neural processing. It seems likely that such a process may operate at the bushy cells in AVCN, such that at low levels a single spike arriving at one endbulb of Held would be enough to fire the cell, but that at higher levels a near coincidence of two spikes might be required.

In a machine model, since we typically represent firing probability as a function of time, rather than go all the way to modeling action potentials, a monaural coincidence detection is probably unnecessary. But an expansive nonlinearity or peak picking to emphasize the extreme peaks, combined with some adaptation to keep the signal in a reasonable range, would be a good functional equivalent. It might be worth converting to a spike code for the binaural cross-correlation, for good resolution at low computational cost, in which case getting sparse spikes with good correspondence between the different ear (microphone) signals is the key problem.

In humans, as opposed to smaller mammals with less propagation delay between their ears, the LSO is relatively small compared to the MSO (Heffner and Masterton, 1990). Kulesza (2007) finds about 15500 neurons in a human MSO compared to 5600 in an LSO. ITD is the predominant cue for lateralization of sounds by human, but we know that ILD is also an important cue that can trade to some extent against ITD, so ignoring the MNTB/LSO pathway in modeling human spatial hearing would probably be a mistake.

22.6 Binaural Acoustic Reflex and Gain Control

Two small muscles in the middle ear, the *tensor tympani* and the *stapedius*, reflexively contract in response to loud sounds (above about 40 sones), and also in anticipation of vocalizing, to reduce the transmission of sound energy through the middle ear and into the cochlea. With respect to the larger one, the tensor tympani, it was discovered by Pollak (1886) that "when only one ear perceives a sound, the muscle on the other side reacts also" (Barth, 1887).

The smaller muscle, the stapedius, is able to respond much more quickly, and contracts at the onsets of sounds as a protective mechanism; if the sound's steady level is not too high, it quickly relaxes again. In echolocating bats, the stapedius is contracted tightly during chirp vocalizations, but relaxed a few milliseconds later when echoes return, in a cycle that can repeat up to 100 times per second (Borg and Counter, 1989).

The tensor tympani is innervated by the mandibular division of the fifth cranial nerve, the trigeminal nerve, and is activated primarily by chewing, but also by loud sounds. The stapedius muscle is innervated by a branch of the seventh cranial nerve, the facial nerve, and is primarily activated by vocalizing. These activations by chewing and vocalizing are bilateral, essentially equal on the two sides. Activations by loud sounds are also bilateral, but not quite symmetric. Though the response is somewhat greater to a sound on the same side than on the opposite side (Møller, 1962), the contractions do tend to track, probably so that the dynamic range of sounds can be greatly compressed without compressing the interaural intensity difference cue so much. Similarly, the efferents in the auditory nerve, the eighth cranial nerve, which reduce sensitivity via the outer hair cells, come from the binaurally sensitive olivary complex and tend to keep the gains of the two ears about equal, preserving the sensitivity to interaural differences. The muscles have a somewhat frequency-dependent effect on gain, and introduce phase shifts, so keeping them matched also helps to preserve interaural phase and time differences (Guinan, 2010).

The muscle-controlled middle-ear effects are followed by cochlear mechanical and neural gain control: efferents from the LSO inhibit the inner hair cells, while efferents from the MSO turn down the gain of the outer hair cells. Brugge (1992) interprets contralateral suppression effects as a sort of auto-tuning of the system, "a possible 'binaural gain control' mechanism to maintain a balance in sensitivity between the two ears in the face of interaural threshold fluctuations." Darrow, Maison, and Liberman (2006) find that "lateral olivocochlear feedback maintains the binaural balance in neural excitability required for accurate localization of sounds in space." Kim et al. (1995) make a similar point for the medial efferents: "Perhaps such a requirement for equalizing the cochlear amplifier gains in the two ears is the functional reason why most MOC neurons exhibit binaural facilitation and/or excitation ... and why some MOS neurons project to the two cochleas." When we build gain-control mechanisms into multimicrophone channels, we should similarly make sure the gains track, at least partially, if we want to use an intensity-difference cue.

22.7 The Precedence Effect

Wallach, Newman, and Rosenzweig (1949) noted that "... localization within a reverberant room is both common and useful. The problem of how this is possible remains, however, unsolved." The *precedence effect*, sometimes called the *law of the first wavefront* or the *Haas effect*, must be a big part of explaining this ability. We perceive the direction of a sound based on time and level differences at its onset, and ignore differences that follow about 2 to 40 ms later, which are likely "contaminated" by echos from the floor, walls, etc. We give the first wavefront "precedence" in the comparisons.

Both echo-localization suppression and source–echo fusion are known as the precedence effect (Burkard et al., 2007). The latter effect, that an echo would reinforce a direct sound rather than be heard as a separate echo, was reported by the head of the Smithsonian Institution, Joseph Henry (1851); he found that an echo delay of about 1/20 to 1/15 second (50–67 ms), from a wall behind the source, was needed to make a distinct

echo. For a short sound ("a sound produced by an instrument which gave a sudden crack, without perceivable prolongation"), the echo delay threshold was somewhat shorter; for a more sustained sound, longer.

About a hundred years later, Helmut Haas (1951) found that the fusion of speech with its echo would still work, for echos in the range of about 10 to 40 ms delay from the direct signal, even if the echos came from very different directions. This observation was a boon to the sound reinforcement business, since it meant that loudspeakers could be placed almost anywhere in a big crowd; as long as an audio delay was inserted to allow the direct sound to arrive before the amplified sound, listeners would hear the sound from the right direction (limited by the *Haas breakdown*, where the later sound is too much louder or the net delay is outside the workable range).

Wallach et al. (1949) coined the term *precedence effect* when they investigated how different directions of a direct sound and an echo influence the perceived direction of the source, in the time interval range where a separate echo is not heard. Using dichotically presented click pairs (that is, presented to the two ears by headphones), they investigated what interaural delay of the first click would make the sound seemed centered, as a function of the interaural delay of, and inter-click interval to, the second click. The result was that very little interaural delay in the first click was needed to counter the effect of an opposite-signed interaural delay in the second click, unless the inter-click-interval was very short. That is, the first click takes precedence, or priority, in determining direction, unless the two clicks come within about a millisecond of each other.

Efforts to model the precedence effect are usually either by modifications to the basic Jeffress cross-correlation model, or by modifications to the stages providing its inputs. Some researchers find a good step in the right direction by incorporating fast onset emphasis in the hair-cell or cochlear-nucleus model that provides the input signals to the correlators (Hartung and Trahiotis, 2001). Recently, a strong correlation has been found in humans between psychoacoustic data, click-evoked otoacoustic emissions, and auditory brainstem responses, suggesting that echo suppression in the 1–4 ms region is primarily due to the nonlinearity of cochlear mechanics (Verhulst et al., 2013; Bianchi et al., 2013). Other researchers find that the physiological suppression observed in the inferior colliculus depends on the ITD of the leading click in a way that implies that it can't come from levels that precede the correlators (Yin, 1994).

Recently, using amplitude-modulated (AM) 500 Hz tones with interaural phase shifting throughout the modulation cycles, Dietz et al. (2013) have shown that the portion of the signal that most drives the lateralization percept is the rising part, not the peak, in the range of 4 to 64 modulation peaks per second. The correlates of this psychoacoustic effect are seen in the responses of binaural auditory neurons in the MSO and IC. Dietz et al. (2014) conclude that, "A comparison of two models to account for the data suggests that emphasis on IPDs during the rising slope of the AM cycle depends on adaptation processes occurring before binaural interaction."

This story is probably not finished, but at least we have reason to believe that emphasis of the rising slopes of onsets in the auditory nerve response is likely to be one key piece of the precedence effect puzzle.

22.8 Completing the Model

Given all these ideas about how different parts of the auditory brainstem process binaural information, and the considerable remaining uncertainty of the exact functions and mechanisms, we nevertheless need to choose a model and implement it if we want to take advantage of the signals available from multiple microphones. We are not constrained to have two microphones mounted in a head-like arrangement, but that is one reasonable approach. An alternative popular approach to multimicrophone input has been to use many microphones in an array, and use adaptive beam steering to combine them into a single monaural signal that gives a good signal-to-interference ratio for a chosen source or direction. That approach is well explored elsewhere, and is not really a hearing approach, so we focus here on what can be done with just two microphones with reasonable separation to get an ITD cue, possibly mounted in a head-like baffle to get some ILD cue.

We address this problem similarly to how we did the monaural stabilized auditory image: extract sparse "trigger" events from one ear to correlate against the signal from the other ear. For bilateral symmetry, we do this in both directions. In general, the ear on the side that sound is coming from will have the cleaner signal, and will thus have a better chance of identifying the key wavefront features that we want to trigger on to correlate with the signal from the other side.

There have been several major reviews of binaural processing models and how they incorporate the precedence effect (Stern and Trahiotis, 1995; Colburn, 1996; Clifton and Freyman, 1997). However Litovsky et al. (1999) note that "there is no model currently published that is able to accommodate available data satisfactorily. In addition, none of the models can account for phenomena such as the buildup or breakdown of the precedence effect, which are thought to be more cognitive." We hope that a good onset-emphasizing cochlear model in combination with a good conception of trigger detection in CN will provide a starting place for working on this lack of applicable models. But a higher-level interpretive or feedback component is still necessary to explain the various aspects of precedence, according to many detailed analyses (Hartmann, 1997; Hafter, 1997; Blauert, 1997).

22.9 Interaural Coherence

To select good times at which to pay attention to ITD and ILD cues, Faller and Merimaa (2004) suggest using a measure of *interaural coherence*: how "similar" the signals at the two ears are, within each frequency band. Wilson and Darrell (2006) suggest a generalization of that approach to learn a dynamic weighting of cues. These studies were in the context of a linear gammatone filterbank with half-wave detection, and spectrograms, respectively, so did not benefit from the natural onset emphasis of the AGC and IHC parts of a more realistic model such as the CARFAC.

It remains to be seen whether the interaural coherence is a complementary, versus redundant, step in a realistic binaural model. Zurek (1987) suggests that both onsets and coherence are needed: "...in addition to the abrupt onset, interaural coherence may be necessary to elicit the precedence effect." Later, however, Zurek and Saberi (2003) show good modeling results using onset emphasis only, in a rather abstract model based on bandpass cross-correlation functions applied to clicks and noise bursts. Hummersone et al. (2010a), on the other hand, conclude from their studies that "a model based on interaural coherence produces the greatest performance gain over the baseline algorithm." Their baseline uses a linear auditory filterbank with the envelope-based inhibitory process of Zurek (1987) to emphasize onsets by suppressing what comes after, followed by ITD and ILD analysis. However, Hummersone et al. (2010b) also achieved good results without the coherence processing, using only the onset-enhancing subtractive inhibition as a precedence model (Zurek's 1987 model), but requiring adaptation of the precedence model parameters to the acoustic environment.

So a coherence weighting may be helpful for realistic spatial sounds as represented by a good peripheral model. Perhaps the complex circuitry of the olivary complex is doing something of this sort.

22.10 Binaural Applications

The power of two microphones over one has recently been exploited by many teams in the PASCAL CHiME speech separation and recognition challenge (Barker et al., 2013) and in the REVERB challenge (Kinoshita et al., 2013). Though many sophisticated signal processing and statistical techniques were developed and evaluated in these works, most of them stopped short of applying much depth of what we know about binaural hearing, such as the importance of the precedence effect.

The power of precedence effect has been confirmed by Smith and Collins (2007), in the context of a pair of microphones on a flat-panel display in a home. They explicity identify onset intervals in which to measure

inter-microphone time delays and estimate talker azimuth.

An early binaural machine hearing application was described by Kaiser and David (1960):

> . . . the binaural processor derives a temporary signal which is used to gate the aural input. In effect, the gating signal leaves the major portions of the preferred talker's speech envelope intact while suppressing sound from other talkers or background noise when these do not overlap with the preferred speech. A preliminary circuit version of such a processor which derives the gating signal by cross-correlation has been built and tested. Subjective measurements in two and three-speaker environments yielded increases of 9 and 5 db, respectively, in signal-to-noise ratio.

This time-domain masking approach to source segregation or enhancement is extended to time–frequency masking in a modern computational auditory scene analysis system, as discussed in Section 23.2.3. A recent promising use of interaural coherence to model precedence effect in better estimating time–frequency masks for enhancing speech in reverberation is reported by Alinaghi et al. (2013). Such techniques are being applied in modern binaural hearing aids, as discussed in Section 28.3.

An example of a successful application of two-microphone directional separation using a precedence effect is presented by Palomäki, Brown, and Wang (2004). The precedence effect model inhibits signals after onsets, in the path to cross-correlators, and the resulting correlation peaks are used in a time–frequency masking approach, leading to improved speech recognition accuracy. And a simple use of a precedence effect with an eight-microphone circular table-top array is provided by Plinge, Hennecke, and Fink (2010). Using spikes detected in rising onset events, they use sparse spike correlations to compute time offsets that robustly determine azimuth to multiple concurrent talkers, even in reverberant environments.

Using multiple microphones and bottom-up processing will not be the end of the story. Blauert (1997) paints a picture of a rich future of binaural applications incorporating higher-level information:

> If we wish to build sophisticated binaural technology equipment for complex tasks, there is no doubt that psychological effects must be taken into account. Let us consider, as an example, a binaural surveillance system for acoustic monitoring of a factory floor. Such a system must know the relevance and meaning of many classes of signals and must pay selective attention to very specific ones, when an abnormal situation has been detected. A system for the evaluation of acoustic qualities of spaces for musical performances must detect and consider a range of different shades of binaural signals, depending on the kind and purpose of the performances. It might even need to take into account the taste of the local audience or that of the most influential local music reviewer. An intelligent binaural hearing aid should know, to a certain extent, which components of the incoming acoustic signals are relevant to its user, for example, track a talker who has just uttered the user's name.

Chapter 23

The Auditory Brain

> ... how do we recognize what one person is saying when others are speaking at the same time (the "cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation? A few of the factors which give mental facility might be the following: (a) The voices come from different directions. (b) Lip-reading, gestures, and the like. (c) Different speaking voices, mean pitches, mean speeds, male and female, and so forth. (d) Accents differing. (e) Transition-probabilities (subject matter, voice dynamics, syntax ...).
>
> — "Some experiments on the recognition of speech, with one and with two ears," Cherry (1953)

> ... the majority of neurons in auditory thalamus and cortex coded well the presence of abstract entities in the sounds without containing much information about their spectro-temporal structure, suggesting that they are sensitive to abstract features in these sounds.
>
> — "Auditory abstraction from spectro-temporal features to coding auditory entities," Chechik and Nelken (2012)

When the brain associates objects, abstract entities, or concepts with sounds, it is extracting meaning from sound. But the process cannot be as simple as a trainable classifier, because the brain must deal with input representing multiple such entities concurrently—the "cocktail party" problem. This, then, is the key function of the auditory brain: analyzing the auditory scene, to jointly decide what sound fragments to pay attention to and what those sound fragments represent. How this happens in the mammalian brain, and how we can model it in machines, remain key problems in the hearing field.

The structure of the auditory brain is complicated, being distributed across many levels and interwoven with somatosensory, visual, motor, and other parts. In this chapter, we survey together the structure and function of the auditory brain, focusing on the how meaning is extracted from complicated sound mixtures. The detailed functions, and their assignment to brain structures, remain rather speculative.

23.1 Scene Analysis: ASA and CASA

In his classic 1990 book *Auditory Scene Analysis: The Perceptual Organization of Sound*, Al Bregman (1990) explores a generalization of what he previously referred to as *auditory stream segregation* (Bregman and Campbell, 1971). The *auditory scene analysis* (ASA) idea is parallel to the concept of visual scene analysis that had been introduced by Duda and Hart (1973) in their *Pattern Classification and Scene Analysis*: how a perceptual system can make sense of complex inputs from cluttered scenes, to extract meaningful descriptions of objects, actions, sources, etc. As Bregman and Pinker (1978) say,

> We see this process of parsing the acoustic information to form coherent streams as analogous to the process of parsing the retinal information to form 'objects' in vision, a process which is currently being studied in artificial intelligence research under the label of 'scene analysis'.

The mechanization of ASA by computational algorithms is known as *computational auditory scene analysis* (CASA). While the roots of scene understanding are older, Bregman provided both the ASA name and the concepts that helped it gel. The term *computational auditory scene analysis* appears to have been used first by Beauvois and Meddis (1991) shortly after Bregman's book came out, and the abbreviation *CASA* was in wide use by 1995.

The basis of ASA is often described in *Gestaltist* terms: sound fragments that are similar, or compatible, or seem to share a common fate, are likely from the same source, and therefore should be analyzed as part of the same whole, or stream. Bregman and others did experiments to elucidate what properties of sound fragments would make them stream together, versus segregate apart into separate streams. Properties such as frequency proximity, common onset and offset, coherent modulation, and harmonicity tend to help parts of a sound mixture join together into a perceptual stream.

Bregman and Pinker (1978) attribute the genesis of the *common-fate* idea to this passage from Helmholtz (1878):

> ... when one musical tone is heard for some time before being joined by the second, and then the second continues after the first has ceased, the separation in sound is facilitated by the succession of time. We have already heard the first musical tone by itself, and hence know immediately that we have to deduct from the compound effect for the effect of this first tone ... When a compound tone commences to sound, all its partial tones commence with the same comparative strength; when it swells, all of them generally swell uniformly; when it ceases, all cease simultaneously. Hence no opportunity is generally given for hearing them separately and independently.

The crux of converting ASA to a computational task is to decide what kinds of sound fragments to use (not necessarily Helmholtz's "partial tones"), how to extract them, and what features, cues, or rules to use to group them into streams. There have been many approaches to these questions, motivated by Bregman's ASA and by available signal representations.

Various forms of stabilized auditory image (SAI) have been popular as starting features for CASA, as investigated by numerous groups in recent decades (Lyon, 1983; Weintraub, 1984; Assmann and Summerfield, 1990; Duda et al., 1990; Mellinger, 1991; Meddis and Hewitt, 1992; Ellis, 1997; Cooke and Ellis, 2001; Slaney, 2005). For example, Duda, Lyon, and Slaney (1990) investigated *correlograms and the separation of sounds* by comparing auditory effects to visual effects in auditory correlograms (SAI movies), for mixtures of synthetic vowels with various modulations. Their key finding was that without modulation, a mixture of three steady vowels just sounds like and looks like a mess, but with modulation, such as pitch shift or vibrato, a vowel can "pop out" of the mixture, in both the auditory and visual percepts. The SAI was shown to be a good representation for observing the comodulation of partial tones that make a vowel hang together and distinguish itself from the mixture, even when those partial tones cannot themselves be identified. Exactly how the SAI is best analyzed to take advantage of this visual percept is still an open question.

The notions of machine hearing and auditory scene analysis have long been tightly coupled, especially through the writings of Bernard Mont-Reynaud and his colleagues at Stanford's Center for Computer Research in Music and Acoustics (CCRMA). His student David Mellinger (1991) connected scene analysis to the extraction of meaning in the form of an interpretation as sources: "Finally *auditory scene analysis* refers to the entire process from the reception of a sound signal through source formation including event formation along the way." This is what the brain does, and what we want machines to do as well. Mont-Reynaud

(1992) gives an inspiring account of the machine hearing research at CCRMA, focused on the scene analysis problem, and crediting their outreach to people in a wide variety of interrelated fields, via the CCRMA Hearing Seminar, for the ideas that they were building on. This tradition at CCRMA continues today, with all aspects of hearing being reviewed in the weekly seminar, led for the last 20 years by Malcolm Slaney, and contributing to progress in machine hearing.

Cooke and Ellis (2001) summarize progress in the field, in terms of a joint focus on machine hearing and auditory scene analysis; they advanced CASA at their respective institutions with an amazing proliferation of good ideas and machine hearing developments, applied primarily to speech.

CASA is a big enough field that it has several books of its own (Rosenthal and Okuno, 1998; Divenyi, 2005; Wang and Brown, 2006). In this chapter, we touch on it only briefly.

23.2 Attention and Stream Segregation

Humans have a hard time decoding several concurrent speech streams, but can more reliably decode one of several concurrent streams if they know which one to pay attention to. For example, if three speech sources are spatially separated by azimuth, knowing which direction to pay attention to can raise a keyword detection accuracy from near one-third (chance) to better than 90% (Kidd et al., 2005). In a two-speaker task with pitch or vocal-tract-length difference, knowing which speaker to pay attention to can have a similarly large effect on detecting words in the attended sentence, sometimes stronger than binaural ITD cues (Darwin and Hukin, 2000).

The problem of how to model attention, as a key part of the field of cognitive psychology, has long been connected with models of the brain and vision and auditory scene analysis; in fact, Bregman credits Ulric Neisser's work on attention as a major influence on his ASA work.

Neisser (1967) describes auditory attention relative to speech streams and preattentive processing in terms of an analysis-by-synthesis theory, contrasting his approach with attention theories of Broadbent (1958), who proposed that irrelevant portions are filtered out, and of Treisman (1964), who proposed that irrelevant or unattended portions are attenuated (weakened), but not removed completely:

> On this hypothesis, to "follow" one conversation in preference to others is to synthesize a series of linguistic units which match it successfully. Irrelevant, unattended streams of speech are neither "filtered out" nor "attenuated"; they fail to enjoy the benefits of analysis-by-synthesis. As a result, they are analyzed only by the passive mechanisms, which might be called "preattentive processes" by analogy with the corresponding stage of vision. Like their visual counterparts, these processes can establish localization, form crude segments, and guide responses to certain simple situations. However, their capacity for detail is strictly limited.

His analysis-by-synthetic strategy is basically a way to use the parts of the incoming sound features that contribute to constructing a meaningful interpretation, while ignoring the rest, rather than removing or attenuating or even modeling them.

Neisser's approach is what many machine hearing systems implicitly do: put the computational and modeling effort into the high-level units of interest, while trying to be tolerant of other nonmodeled sounds that are also present and represented in the extracted features. Doing it more explicitly may be a good idea.

23.2.1 Analysis by Synthesis

In a speech recognition system, a word sequence is typically found by searching for a best path through a network of words. In such a system, the analysis-by-synthesis strategy can be implemented by a *beam search*—a strategy that always tries to extend the path through a moderately narrow beam of options around

a current best hypothesis of the constructed message. Such a system naturally attends to the current message, as opposed to a full-search strategy that lets all possible messages continue to compete until a final output is needed.

In this way, the analysis-by-synthesis method is essentially an alternative to using the feature-based stream separation approach of CASA. For it to work, the evaluation of how well a path through the word network matches the input sound features needs to allow and ignore a considerable amount of "extra" sound in the input.

In terms of our four-layer model, Neisser's strategy would have layer three doing preattentive feature extraction, and layer four would be completely responsible for the auditory scene analysis-by-synthesis. Bregman's ASA moves away from this, doing stream formation preattentively, which would fit better in layer three; Bregman and Campbell (1971) say, "The stream-forming processes described in this paper probably fall into the category of 'preattentive processes' discussed by Neisser." Simply put, these are the alternatives that a model of selective attention must choose between: ignore the extra interfering sounds, or separate them (by filtering, attenuation, or stream assignment). Either way, the system needs to "pay attention" to the part that seems like what it wants to hear.

In a machine hearing system, it is potentially possible to attend to multiple streams at the same time. In this respect, super-human performance on some tasks is likely to be possible. On a constrained multitalker speech recognition task, super-human performance has been demonstrated using a statistical approach trained on speech from the specific talkers involved (Kristjansson et al., 2006), but that approach is not likely to generalize as well as a CASA approach might.

Attention can be directed by side information, such as an instruction of whether to listen to the talker on the left versus on the right, or to the male voice versus the female voice. Or it may be more dynamically controlled, feeding back from the initially discovered position or identity of a source of interest, or from information decoded from the initial part of a message. Sometimes attention may be captured based on the "saliency" of a source, which is why we have terms like "attention-grabbing" for some kinds of sounds. Depending on the needs of an application, different strategies will be needed to control attention. If multiple streams can be separated, they can be analyzed after the fact, with attention to each in turn. In real-time applications such as voice communication and recognition, attentional processes may need to quickly commit to what parts of the signal to follow, and ignore the rest. In a two-microphone game controller, for example, an easy strategy would be to attend to whatever talker is most nearly "straight ahead" of the device.

In humans, attention is thought by some to be a process involving cortical feedback to thalamus, controlling what portions of sensory input get selected for projection to cortex, and what portions get suppressed (King, 1997; Suga et al., 2000; Yu et al., 2004). This suppression is analogous to the masking gains in time–frequency masking approaches to CASA (see Section 23.2.3), but can probably modify ascending information in more elaborate ways than simply varying gains.

As an alternative to the masking interpretation, as Daniel Kersten explains, cortical backprojections (even between cortical regions) may be there to support the analysis-by-synthesis approach (Kersten, 2000):

> It has been argued that the inherent confounding of diverse scene causes in natural patterns, including images, necessitates analysis-by-synthesis through a generative model that tests top-down predictions of the input. One commonly discussed explanation for the pattern of back-projections between cortical areas is that these connections enable the expression of unresolved high-level hypotheses in the language of an earlier level. This expression can then be tested with respect to the incoming data at the earlier level. Thus, domain-specific models in memory can be manipulated to check for fits to the incoming data in ways that are difficult bottom-up.

Mesgarani and Chang (2012) have shown that the response of neurons in cortex (in a secondary auditory cortical field), when a human listens to a mixture of speech sounds, depends on the human's state of attention.

The response can represent one talker or another, depending on what the human is trying to listen for. It seems likely that this cortical response depends on feedback to the MGB, where cues such as pitch maps from the IC can be used to help the system select features more attuned to one talker or the other.

23.2.2 "Pure Audition" versus Top-Down Processing

Slaney (1998) has provided an eloquent critique of using just bottom-up techniques, or *pure audition*, for auditory scene analysis, as well as a critique of a typical goal of CASA being *sound separation* (Slaney, 2005). The alternative is *sound understanding*, that is, a connection of the bottom-up auditory data to top-down meaningful interpretations. A good example of a system that explores a way to make such a connection was provided by Barker, Cooke, and Ellis (2000a). In their system, one or more models of speech are used to explain portions (*coherent source fragments*) of the acoustic signal, as represented in a time–frequency plane (a spectrogram). The fragments used were deliberately as simple as possible: just individual small cells in the time–frequency plane. For this, they introduced the notion of the time–frequency mask, a plane of values indicating whether each element of the time–frequency plane was interpreted as part of a target signal, or of background.

Slaney's distinction parallels the Bregman versus Neisser approaches to attention outlined above. Bregman advocates mostly bottom-up stream formation, while Neisser's analysis-by-synthesis is top-down in the sense that Slaney means.

23.2.3 ASA by Time–Frequency Mask

Auditory scene analysis inherently needs an attention mechanism, to form and track streams and decide which stream to pay attention to. The use of a mask in the time–frequency plane to assign sound fragments to a target stream (versus other streams or a noise background) has become a popular and effective method of CASA. The approach provides enough flexibility to tie a variety of auditory representations to machine learning systems and applications. I (Lyon, 1983) introduced the idea of masking for sound separation in the cochleagram domain this way:

> Following the local, directional interpretation, per-channel time-variable gains are applied to the input cochleagrams to produce output cochleagrams representing different sound streams. These gains change very quickly, typically reacting in under 0.5 msec to a change in correlation peak position caused by an onset from a different source. In the extreme case (locally high SNR), gains of zero and unity may be said to "gate" local sound fragments to the appropriate output stream. Thus, unlike techniques that compute a slowly changing optimal spectral modification of the signal, this model must be viewed as very much a time-domain technique, which takes advantage of both the fine time resolution and the frequency separating properties of the cochlea.

Binary masking or stream assignment became a popular approach in separating concurrent speech sounds. In the first monaural two-speaker CASA system, Weintraub (1984) grouped and assigned sound fragments at a somewhat higher level, assigning time–frequency regions, rather than individual time–frequency cells, to streams: "A group object is a collection (across both frequency and time) of neural events, having similar properties, that are perceived as a unit. It is an intermediate level in the representation of sounds and corresponds to the natural segmentation of the incoming sound into frequency–time regions that have similar properties." Assmann and Summerfield (1990) used a place–time analysis based on an auditory correlogram (SAI), from which they estimated spectra of two vowels at two different pitches, but did not use a mask approach. Later, Meddis and Hewitt (1992) used a binary time–frequency masking approach for recognition of concurrent vowels of different pitches, "... using the decision concerning pitch values to segregate frequency-selective

channels into two mutually exclusive sets of channels, one for each vowel." With this masking approach they were able to make a system that got closer to human performance, including improvement with increasing pitch separation.

Green, Cooke, and Crawford (1995) evaluated the notion of binary masking in a time–frequency plane as a simulation of what a CASA system might be able to achieve, and showed that masking could potentially improve speech recognition in noise to within a few dB of human listener performance. Further work by Hermansky, Tibrewala, and Pavel (1996), Cooke, Morris, and Green (1997), Drygajlo and El-Maliki (1998), and Hu and Wang (2001) developed the concept as a practical approach related to spectral subtraction (Boll, 1979), in the context of speech and speaker recognition. Wang (2005) goes so far as to support binary mask generation as the *goal* of auditory scene analysis.

Kim, Lu, Hu, and Loizou (2009) made an encouraging step in showing the improvement of speech intelligibility in noise for normal-hearing human listeners, which they accomplished by using a time–frequency mask estimated from the noisy signal; their system relied on training on the particular noise types and speakers involved, so was not yet a real solution. Loizou and Kim (2011) went on to analyze the types of errors that worked against intelligibility, and charted a further encouraging direction for optimizing such systems to give not just subjective quality improvements, but also real intelligibility improvements, for normal-hearing subjects. Using perceptual experiments, they showed that overestimating the signal spectrum was more detrimental than under-estimating it, in terms of squared error.

Watts (2010) demonstrated intelligibility and *mean opinion score* improvement, using a binaural (two-microphone) approach, using masking within the domain of a PZFC cochlear model's output. This masking was the basis of the Audience, Inc., speech enhancement technology on mobile phones.

Mask estimation based on binaural cues for separation by direction, and on periodicity cues for separation by talker pitch, have been well explored. It is even possible to estimate masks to separate unvoiced speech fragments from nonspeech interference in monaural signals, based on the statistics of speech energy distributions; Hu and Wang (2008) conclude, "our system captures most of the unvoiced speech without including much interference."

When portions of a sound mixture that are mostly noise are removed, and portions that are mostly signal are kept, the signal-to-noise ratio naturally is improved. But portions of the signal that are set to zero may be important, and the unnatural cue of silence where signal is expected may harm intelligibility more than a masking noise would, so this method of SNR improvement does not always lead to an intelligibility improvement or natural-sounding separated sound. And it does not necessarily lead to a sound spectrum that satisfies the expectations of a speech recognition system. As an alternative to trying to produce a cleaned-up signal, the mask can be interpreted as indicating *erasures*, or missing data, at time–frequency cells that are mostly noise. This approach to using the mask has been explored mostly in application to speech recognition in noise (Cooke et al., 1997; Barker et al., 2000b; Roman et al., 2003; Raj and Stern, 2005).

A number of researchers have shown how fairly sophisticated classifiers can be trained to decide for each time–frequency element whether it fits better with the target stream or with background—an idea that appears to have originated with Seltzer et al. (2000). DeLiang Wang and his colleagues have explored the limiting behavior of this approach, in terms of an *ideal binary mask* and *ideal ratio mask*, which are the masks that optimize the signal-to-noise ratio of a sound stream reconstructed as a spectrogram inversion of the masked time–frequency plane (Brungart et al., 2006; Srinivasan et al., 2006; Li and Wang, 2009). Ideal masks can be computed by an *oracle* that knows the target signal and the interference, as in the original "simulation" of Green et al. (1995); so it is easy to build a set of training data for training a classifier to compute an approximation to the ideal mask from the mixed sound signal.

The mask value at each time–frequency element (or spectrogram pixel) is typically computed independently, either by comparing estimated signal and noise levels, or by a more general classifier that can be given any amount of context as input (Wang et al., 2014). For example, the mask may be estimated based on an

analysis of a binaural auditory image (cross-correlogram) that separates sources by their lateral position (Roman et al., 2003), or by a process that searches for a best overall interpretation of the pitches and locations of multiple sound sources, and that evaluates compatibility of binaural and monaural correlation features with a preferred source in that interpretation (Woodruff and Wang, 2013).

23.3 Stages in the Brain

The auditory nervous system has several more stations between the sense organs and the cortex than the visual system has; the retina projects directly to the thalamus, while the cochlea has intervening processing steps in cochlea nucleus, olivary complex, lateral lemniscus, and inferior colliculus. Though they are much studied, there is relatively little agreement about the computational goals or functions of some of these stages. We expect that their computation of representations of sound to project through the thalamus to the auditory cortex might eventually be understood in terms of the auditory image concept: mapping the signals from the auditory nerve into something more image-like, like what comes from the optic nerve, but extracted from correlations of temporal structure from lower levels.

In the mammalian brain, the *midbrain* is the upper part of the brainstem. The lower part of the brainstem, the *pons* and the *medulla oblongata*, are where the auditory nerve enters at the cochlear nuclei, and where the binaural centers of the olivary complex are. Some functions of these brainstem regions have been discussed in Chapter 20 through Chapter 22. Between the midbrain and the cortex is the *thalamus*, a structure with many sensory and motor functions. The layout of subcortical regions is sketched in Figure 23.1.

The *inferior colliculus* (IC) in the midbrain and the *medial geniculate body* (MGB) in the thalamus are the main auditory stages between the lower levels and cortex. Like all other levels of the auditory nervous system, these areas have matching left-side and right-side parts. That is, they are bilaterally symmetric across the midline. There are crossing connections (commissures) linking the two sides at several different levels.

While these are important centers in the ascending auditory pathways, it would be a mistake to neglect the fact that they also have huge descending connections, from cortex to geniculate especially. Suga et al. (2003) describe the function of these *corticofugal* connections in terms of learning—allowing the system to adapt to ethologically meaningful properties of sounds.

The parts of the brain don't divide up as cleanly and logically as we might wish, and the namings of the various divisions are accordingly complicated. Divisions such as brainstem, pons, midbrain, thalamus, and cortex are not specific to hearing, but appear frequently in the hearing literature. The hearing-related nuclei within these might be described with such terms as auditory midbrain, subthalamic auditory nuclei, and names of the various specific nuclei. For an engineer, it is hard to relate them to functional models. Nevertheless, there is a lot known that bears on our quest to make machines hear; we review only a small part of it here.

23.3.1 Where is the Auditory Image?

If the subcortical auditory nervous system computes "image-like" representations that project to cortex, we would like to know where, and we'd like to have ways to measure and characterize what's going on, and what these auditory images "look like" as a proxy for what sounds "sound like." Auditory images are probably not computed in the thalamus, if the visual nervous system is a good analogy—the thalamus gets images as input, from the retina, and "relays" them to the cortex, probably with some mechanisms to help focus attention on the interesting aspects of the image. Therefore, we presume that auditory image calculation is done below the thalamus. Kuwabara and Suga (1993) concluded that "delay lines and amplitude selectivity are created in subthalamic auditory nuclei" (in echolocating bats). That is, the auditory-image-like neural responses that they find in thalamus come from the midbrain; and maybe lower.

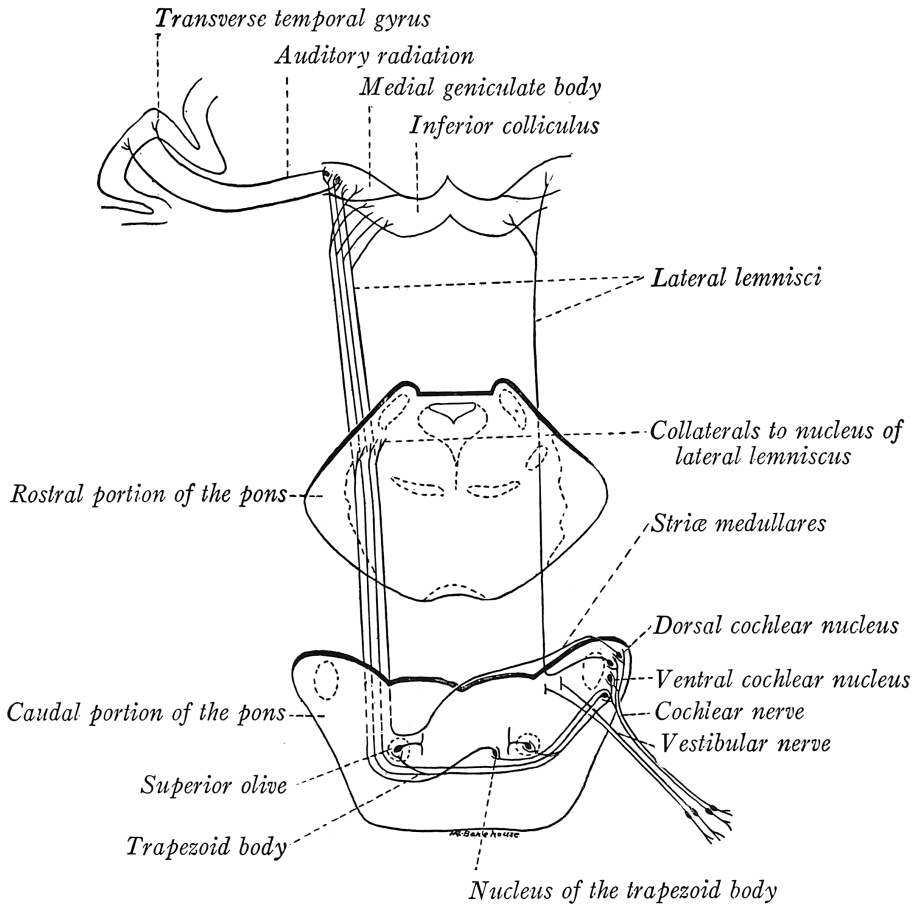


Fig. 233.—Diagram of the auditory pathway. (Based on the researches of Cajal and Kreidl.)

Figure 23.1: The afferent connections in the auditory nervous system, as rendered by Miss M. E. Bakehouse for Ranson (1920). The lower areas, labelled in the "caudal portion of the pons," are near the boundary between the medulla oblongata, the lower part of the brainstem, where the auditory nerve enters the brain at the cochlear nucleus, and the pons, the middle part of the brainstem. The main auditory area of the midbrain, the upper part of the brainstem, is the inferior colliculus (IC), and of the thalamus is the medial geniculate body (MGB). We have expanded our knowledge of the connections and functions and sub-areas since then, but the overall anatomy as known a century ago remains accurate. As the drawing shows, the main afferent projections go from the cochlea to contralateral brain areas—sounds from the left are processed on the right, and vice-versa.

Unfortunately, finding the auditory image has been hard. Finding any coherent auditory processing functionality has been hard—we can't see the functional forest for all the phenomenological trees—so this does not really weigh against the concept, it just means there's work to do. Neurons that respond best to a particular combination of frequency and repetition rate may be particularly hard to identify, and searching for them with sinusoidal amplitude modulation, as many have done, is probably not a good way to find them, since such stimuli evoke a fairly weak pitch percept.

Bendor and Wang (2006) find neurons in lateral Heschl's gyrus (HG), a secondary cortical area at the edge of primary auditory cortex, that respond to a specific pitch, even with missing fundamental, and respond more strongly to a signal with higher pitch salience—that is, a signal that evokes a strong pitch sensation. Other studies suggest an involvement of HG in pitch perception. Schneider et al. (2005) indicate that HG displays differences between groups of humans segregated by certain pitch perception preferences. Wong et al. (2008) find that the left side HG is bigger in people who are better at learning to understand pitch-based languages. Foster and Zatorre (2010) find that this region is bigger on the right side for musicians.

Hall and Plack (2009) suggest that maybe planum temporale (nearby HG) is the place responding to pitch, but also question whether the responses found with iterated ripple noise (IRN) stimuli are really the pitch tuning sought. If one of these really is a pitch map in cortex, it may be computed from an auditory image computed at a lower level, by aggregating across place channels, as in our "pitchogram" of Section 21.11.

Simmons and Simmons (2011) find evidence for common midbrain mechanisms of periodicity-pitch extraction in a wide range of vertebrates ("bats and frogs and animals in between").

23.3.2 Where is CASA?

In our four-layer modular model of machine hearing systems (Chapter 1), where should we try to perform the functions of CASA? Perhaps in the third level: extraction of features from auditory images. But we could make better auditory images if we had separation of events at a lower level, so perhaps we can integrate CASA into the process of forming auditory images.

There is a long history of doing CASA after generation of auditory images, or *correlograms* as they were known in much of this work (Slaney, 2005): my work on binaural separation (Lyon, 1983), Weintraub's two-voice pitch tracking and separation (Weintraub, 1984), Ellis's *wefts* as sound analysis primitives (Ellis, 1997), and many others. On the other hand, there has also been a lot of work on separating sounds in the time–frequency plane, or neural activity patterns (NAPs, or cochleagrams) (Barker et al., 2000b; Cooke and Ellis, 2001).

In all of these approaches, some kind of attention mechanism is needed to decide which part of the mixture to pay attention to. This attention mechanism must be at least partly in, or controlled by, our model's layer four, the layer most closely specialized to the application that defines the kind of meaning that the system is trained to extract.

23.4 Higher Auditory Pathways

How the brain handles different cues, or sound source properties, remains a matter of debate and speculation. One theory is that there are functionally specialized pathways for object identity (including size and message) and object location: the "what" and "where" pathways. Alternative theories point out that when multiple sources are present, this approach brings up a difficult *binding problem*, or keeping the cues for each source, such as pitch, timbre, and location, in their separate pathways, somehow bound together; the cues for the sources are interdependent, so can't be processed far apart (Bizley et al., 2009).

CASA methods that jointly consider pitch and azimuth cues to determine what sound fragments to keep or suppress need to have those cues accessible in one place. At the same time, systems that need to quickly

react to the location of new salient sources would do well to use a specialized "where" pathway, not slowed down by the slower interpretive processes of CASA. The brain can have it both ways, and the specialized brains of owls that catch mice in the dark, or bats that catch moths in the dark, certainly do so.

Responses to sound in secondary cortical areas may be part of the "what" pathway. As in secondary and later visual areas, we may find *grandmother cells*—cells with responses so selective that they fire only when you see (or hear) your grandmother, or similarly specific responses (Gross, 2002). It is notoriously difficult to discover what stimulus such cells are most responsive to. When a sound is found that a brain cell likes, characterizing that cell's tuning in terms of the parameters used to make the sound may be very unsatisfactory. For example, cells characterized in terms of spectro-temporal receptive fields will appear to be tuned to spectro-temporal patterns, even if what they really care about are abstract speech features, or signs of danger, or conspecific neighbors, or other such "what" information carried by sounds.

This difficulty in interpreting what the neurons care about explains why animals with very particular auditory needs, such as owls, bats, and songbirds, sometimes help researchers find faster progress in analyzing the auditory brain (Konishi, 1991; Suga et al., 2000; Jarvis, 2004). Denny (2007) explains that, in bats, where biosonar is usually thought of as a "where" strategy, the auditory image approach is literally applicable to the "what" in identifying prey:

> They form sound pictures that are different qualitatively from light pictures, but are processed as automatically and provide information (but different information) in as much detail. The images formed must incorporate the shapes of all the distinct objects perceived, so that they become segregated in perception, and these shapes must be updated with each pulse—a process that has been likened to computerized tomography. Not bad for a half-gram brain.

The functional organization of the brain is not yet clear. While research on the auditory brain continues, models that localize functions to structures are useful in building functional machine hearing systems. One hypothesized assignment of function to different brain areas has been presented by Watts (2012), and is shown in Figure 23.2.

23.4.1 The Inferior Colliculus

Many functions have been hypothesized for the inferior colliculus (IC). However, as Casseday and Covey (1996) note, "A general statement of the function of the inferior colliculus is lacking, even after more than three decades of electrophysiological investigation." They argue that the mammalian IC does various kinds of ethologically important detection, filtering, demodulating, and converting of fast time-varying input to slower action-oriented output. "Neurons in the inferior colliculus are filters for sounds that require immediate action, such as certain sounds made by prey, predators, or conspecifics." In that sense, the IC may be doing the "extraction of meaning" that we usually think of as a higher-level function. Or, especially in higher animals, it may also be extracting features specific to the needs of more elaborate meaning extraction circuits in the cortex—specialized auditory images, perhaps.

Langner, Albert, and Briede (2002) have reported direct observation of maps with orthogonal frequency and pitch-period axes, in ICC (the central nucleus of IC) in chinchillas. Langner, Dinse, and Godde (2009) have shown a corresponding map in primary auditory cortex, in cats. So we consider the extraction of Licklider-style auditory images to be one of the likely functions of IC. The details, however, remain unclear, and these observations have not yet been well corroborated or refined.

23.4.2 The Medial Geniculate Body

In the visual nervous system, the optic nerve connects to the *lateral geniculate body* (often called *lateral geniculate nucleus*, LGN) of the thalamus, a pathway physically and logically parallel to the auditory connec-

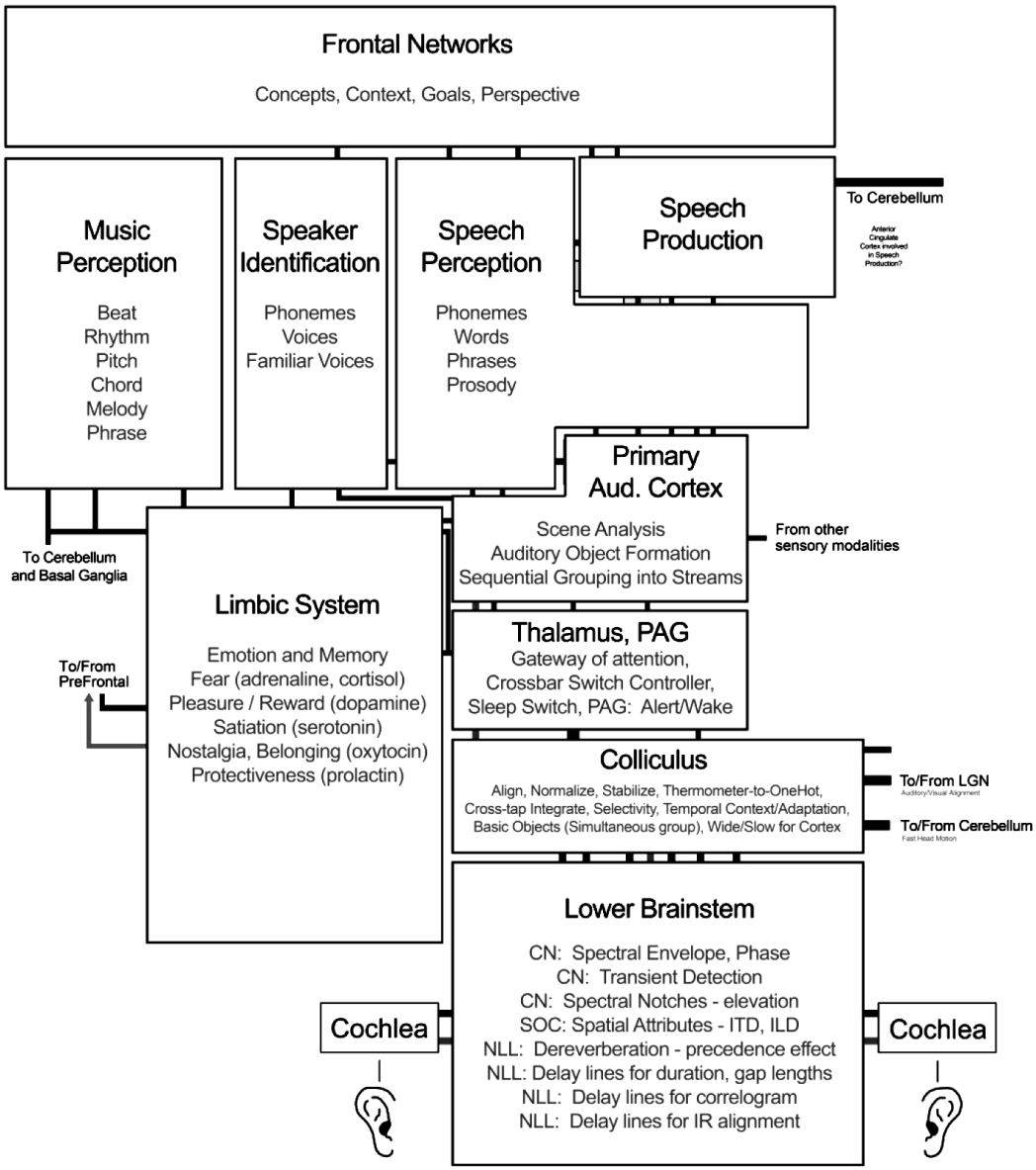


Figure 23.2: The brain function block diagram of Watts (2012) shows a hypothesized assignment of functions to structures. [Figure 1 (Watts, 2012) reproduced with permission of Springer.]

tion through the MGB. Therefore, we can guess that these adjacent thalamic "relay stations" have similar or analogous functions in the visual and auditory systems. Indeed, when the optic nerve in ferrets was forced to connect to the MGB instead of to the LGN, a visual space map was induced in what would ordinarily have been auditory cortex (Roe et al., 1990). The repurposed cortex even develops more-or-less normal visual orientation-selective modules (Sharma et al., 2000), suggesting that the higher-level cortical structures are largely interchangeable.

One function of the thalamic sensory areas is thought to be the mediation of attention. They may also do some temporal decorrelation processing to make the signaling to the cortex more efficient, according to Dong and Atick (1995), who argue that an attentional gain-control mechanism is effected by a varying bandpass filter that can reduce both temporal correlation (by suppressing slow fluctuations) and noise (by suppressing fast fluctuations). Their linearized theory offers a good match to measurements of temporal receptive fields (responses to sinusoidal temporal modulation) of the visual information relayed by LGN neurons to primary visual cortex. In terms of auditory processing, this kind of filtering might be analogous to the RASTA filter discussed in Section 5.9, but with adaptive parameters.

The ventral, dorsal, and medial divisions, or subnuclei, of the MGB have a variety of different cell response types, and several different frequency axes, or tonotopic dimensions. They seem to encode all sorts of binaural and monaural cues, but consistent dimensions of mapping (other than cochlear place) are difficult to find. These regions have a variety of connections and forms, including inputs from other senses, and partially laminar structure.

Chowdhury and Suga (2000) describe how the plasticity of frequency maps in primary auditory cortex is mediated by corticofugal (from cortex outward) feedback to MGB. In echolocating bats in particular, this mechanism allows the bat to adapt its cortical processing to greatly emphasize the frequencies of its bio-sonar vocalizations, as well as to adapt its maps of echo delay times. The discrimination of echo delays in bats is likely equivalent, or at least analogous, to pitch discrimination in other mammals—both are essentially autocorrelation-like processes. The MGB is involved in adapting these maps, but the actual computation is probably at a lower level, such as in IC.

While models of visual information processing in LGN and auditory information processing in MGB are still somewhat primitive, we do at least have a useful conceptual model of this stage as a preprocessor of images, or auditory images, on the way to cortex. The initial generation of auditory images is probably at levels below the MGB, in the brainstem, where there are no counterparts in the visual system.

23.4.3 Auditory Cortex

Primary auditory cortex, like lower auditory areas, is tonotopically organized; that is, it has a dimension with a "best frequency" in correspondence with cochlear place. However, cortical neurons are mostly unresponsive to pure tones, so their responses are studied with more "interesting" stimuli, including speech and various kinds of modulated signals. Neurons are then typically characterized in terms of the spectro-temporal patterns likely to evoke a response, in terms of spectro-temporal receptive fields (STRFs). But recent studies suggest that what the neurons really care about is more complicated than their STRFs would suggest.

In human epileptic patients, with electrode arrays temporarily on their brain surfaces, studies of speech perception have been done while waiting for information on seizures. Using estimated STRFs, the experimenters were able to reconstruct reasonable looking spectrograms of speech from the recordings of cortical neurons in secondary areas such as superior temporal gyrus (Pasley et al., 2012), similar to what they had been able to do in ferret brains (Mesgarani et al., 2008). Subsequent investigation showed that it was possible to reconstruct spectrograms of the attended talker in a two-talker stimulus (Mesgarani and Chang, 2012), suggesting that either the lower level features had been separated, or that the neural signals being picked up coded an interpretation in terms of decoded speech units, not a stimulus feature representation per se. More

recently, Mesgarani et al. (2014) have shown that these response properties in this secondary cortical region are closely aligned to abstract phonetic features.

These studies provide clues about some of the levels of speech-related cortical representations in humans. A corresponding idea in animals, using ethologically relevant signals (cats listening to birds) and studying the brain's response to them, has been explored by Chechik et al. (2006), who found an information reduction, or "funneling" of cues, at higher levels of the brain, and by Chechik and Nelken (2012), who further showed that at the higher levels (thamalus and primary auditory cortex), the responses were much more indicative of ethologically relevant abstract categories than of spectrotemporal sound features.

Abstract categorical information extracted from sound has even been shown to affect responses to related images in visual cortex (Vetter et al., 2014). Thus, we see the cortex as extracting meaning, more than just representing the physical features of sound.

23.4.4 Descending Pathways

The strong feedback from cortex to thalamus may be involved both in dynamic attention and in training preattentive features at the level of thalamus and perhaps even lower. Suga (2008) describes the latter type of function: "The corticofugal system has multiple functions. One of the most important functions is the improvement and adjustment (reorganization) of subcortical auditory signal processing for cortical signal processing."

Winer (2006) paints a more elaborate picture of the complexity we are confronted with in trying to understand the descending pathways:

> These are among the largest pathways in the brain, with descending connections to auditory and nonauditory thalamic, midbrain, and medullary regions. Auditory corticofugal influence thus reaches sites immediately presynaptic to the cortex, sites remote from the cortex, as in periolivary regions that may have a centrifugal role, and to the cochlear nucleus, which could influence early central events in hearing. Other targets include the striatum (possible premotor functions), the amygdala and central gray (prospective limbic and motivational roles), and the pontine nuclei (for precerebellar control). The size, specificity, laminar origins, and morphologic diversity of auditory corticofugal axons is consonant with an interpretation of multiple roles in parallel descending systems.

In our machine hearing models, we can potentially use such feedback in several corresponding ways: as a short-term modulation for attention, as a backpropagation training path to optimize lower-level feature extraction, and maybe more.

23.5 Prospects

Lacking more specific characterization of the functions of IC and MGB, we think of these brain areas as doing auditory image formation and feature extraction, and feature filtering for attention, in support of, and as controlled by, higher-level processes in auditory cortex that extract meaning. There are many more specific proposals for what various areas are doing (Watts, 2012), but not yet an accepted architecture that we can use to refine our machine models. While we wait for progress, we treat the function of these areas under the umbrella of CASA.

From Colin Cherry's 1953 "cocktail party effect" paper to modern systems that exploit multiple binaural and monaural cues to enhance a signal of interest and suppress interference, CASA, the "machine" version of ASA, has been making steady progress. A continuing tight coupling of CASA research with psychophysical

and physiological studies of complex multisource sound processing will likely lead to better machine hearing, along with better understanding of human hearing.

The relatively high-level functionality of CASA is expected to evolve quickly, so we have not tried to summarize the current state of the art in detail. We believe there is significant potential in better integration of techniques such as masking with representations such as auditory images. We hope that the lower-level systems that we teach in this book will facilitate such further advances.

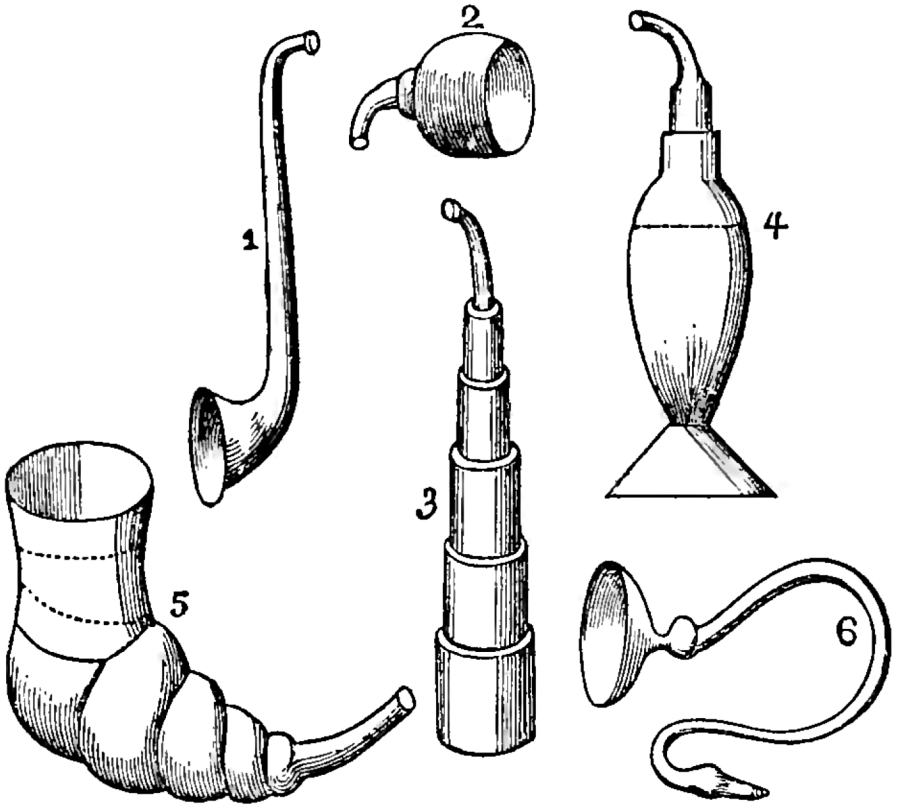
Part V

Learning and Applications

Part V Dedication: Max Mathews

This part is dedicated to the memory of Max Vernon Mathews (1926–2011), the father of computer music. Max had a decades-long focus on applications of computers to hearing and to music analysis, synthesis, and performance. His work on computer speech, music, and hearing started in the late 1950s at Bell Labs (Mathews, 1959, 1961, 1963). I had the opportunity to know Max at Stanford's CCRMA (Center for Computer Research in Music and Acoustics) where he worked for many years. When I taught my Human and Machine Hearing course at Stanford in 2010 (Psych 303, in affiliation with the Mind, Brain, and Computation center), Max came and audited the class once a week, climbing the stairs to the third floor with his hiking sticks. He invited me to his lab and explained the "coupled-form" filter that he was using for music synthesis; I subsequently adopted it as the basis for the digital implementation of my various cochlear filter models, so it figures prominently in earlier parts of the book.

In this part, we discuss the top two layers of our simple framework for machine hearing systems: types of systems that can be trained to address machine hearing applications, and ways that features can be extracted into a form suitable to be presented as inputs to such systems. We discuss several example applications, including ones on which we have published studies, and a survey of some others.



An early application of machine hearing concepts was in the improvement of hearing aids. These improved ear trumpets (Turnbull, 1887) are predecessors to more sophisticated hearing aids.

Chapter 24

Neural Networks for Machine Learning

> In order for a digital neocortex to learn a new skill, it will still require many iterations of education, just as a biological neocortex does, but once a single neocortex somewhere and at some time learns something, it can share that knowledge with every other digital neocortex without delay. We can each have our own private neocortex extenders in the cloud, just as we have our own private stores of personal data today.
>
> — *How to Create a Mind*, Ray Kurzweil (2012)

24.1 Learning from Data

Just as our brains use learning as we develop the ability to interpret the world through patterns arriving on the auditory nerves, so can machines use learning to develop the ability to extract meaning from the sound representations extracted by auditory models.

The inputs that machines learn from are called *data*, and come in many forms. Sometimes we use *supervised learning*: training data associate sounds with answers, and the machine learns a model for that association, so that it will give good answers for novel sound data later. If we have lots of sound data, but no good answers to say what it means, we can still model the data and learn to produce compact meaningful descriptions and predictions of it, using *unsupervised learning*.

In this chapter we focus on supervised learning, and on *artificial neural networks* (ANNs or simply neural networks or neural nets) as a general class of techniques that were originally motivated by theories of how brains work. Neural nets and their descendants are widely employed in classification problems, where the answers are class decisions, and in regression problems, where the answers are continuous functions of the inputs. Our examples focus on classification.

The machine learning (ML) community discovered long ago that learning to get mostly right answers on a training set is not a safe goal for supervised learning. A system that models the training data well may get all the right answers on those, but still fail to generalize to independent testing data. So it is important to have an independent test set, and to use techniques that learn from the training data but generalize well to the independent testing data. Furthermore, trying many different systems and parameters, and picking the one that is best on the testing data is not safe, since that effectively involves the testing data in the optimization process. For this reason, there are various conventional strategies used in training, evaluating, and optimizing ML systems. In this chapter, we won't get into those, but we will use examples with a 50/50 training/testing data split, to give a first-order view of the training and generalization problem.

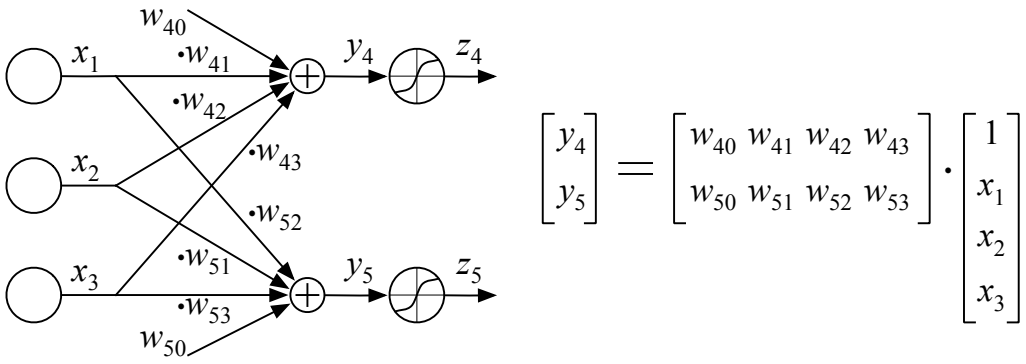


Figure 24.1: A three-input two-output single-layer perceptron with threshold adjustments, as a signal flow graph, and with the linear part as a matrix–vector multiplication. Each linear combination y_j is the dot product of the input vector with a row of the weight matrix: $y_j = \sum_i w_{ji} x_i$ (where $x_0 = 1$ to allow threshold adjustment through w_{j0}). The optional nonlinear part is shown as sigmoidal (s-shaped) nonlinearities. The empty circles on the left are input units, which may sometimes represent the outputs of another perceptron.

24.2 The Perceptron

Invented in the 1950s (Rosenblatt, 1957), the *perceptron* is the prototypical learning machine for pattern classification. It is the first of a wide range of trainable computing architectures known as neural networks. The basic structure of the perceptron is still at the core of many modern machine learning systems (Collobert and Bengio, 2004), though it has been stretched in so many directions that it is sometimes hard to recognize.

The original simple perceptron, or *single-layer perceptron* (SLP, though it is sometimes called a *two-layer perceptron* when inputs are counted as a layer) is little more than a linear mapping from a vector of input features to an answer, for whatever problem it is trained to provide answers for. Simplified to its linear version, it's just a matrix multiplication:

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

where $\mathbf{x}$ is a column vector of feature values, or measurements, representing the input to the machine, describing the problem case to be analyzed; $\mathbf{y}$ is the answer, a scalar or column vector that the machine is trained to provide; and $\mathbf{W}$ is a matrix, of the appropriate dimensions to map $\mathbf{x}$ to $\mathbf{y}$. The output $\mathbf{y}$, or each of its dimensions if it is more than just a scalar, is often interpreted relative to a threshold, or in terms of its sign, such as positive for *Yes* and negative for *No*; that thresholding makes the perceptron nonlinear, and useful as a decision device, or classifier, as we discuss below.

The values in the matrix $\mathbf{W}$ are known as *weights*, since each output is a weighted sum of inputs via the matrix multiplication. The weights are where the learning occurs—a training algorithm tries to set the weights to give good answers, based on examples in a set of training data. The matrix of weights is based on the idea of varying synapse strengths between neurons in a biological brain.

Figure 24.1 shows a perceptron with three input feature dimensions and two outputs. The two-output perceptron is equivalent to a pair of one-output perceptrons with the same set of inputs, but independent weights. The weight matrix is 4x2 because the 3-dimensional input is augmented with a constant 1, multiplying a bias parameter w_{k0} for each output y_k.

One might think that the perceptron is too simple to solve all problems. Indeed, Minsky and Papert (1969) set the field back by more than a decade by proving that perceptrons are unable to learn even some very simple input–output mappings. When people realized how easy it was to work around some of those

limitations, however, the field of neural networks had its heyday in the 1980s and 1990s. More recently, ML methods have gotten more formal and mathematical, but in many cases are still built around perceptron-like operations, whether they say so or not.

24.3 The Training Phase

There are many ways to train a perceptron. Possibly the simplest is to compute the matrix $\mathbf{W}$ that minimizes the mean squared error in mapping training inputs $\mathbf{x}$ to training outputs $\mathbf{y}$, using well-known linear least squares techniques. If our training targets are categories, we first have to convert our categorical targets to numeric values; for example, +1 for one class and −1 for the other. Least-squares may not be quite the right thing to optimize if what we really care about is classification accuracy, but it is often good enough.

For large-scale learning problems, which may mean many thousands of feature dimensions and many millions of training samples (or an effectively infinite pool of potential training data to draw from), it is still possible to do the training by setting up and solving a least-squares problem, because linear least-squares techniques only need second-order sums of products of the input and output values, which can be collected in one pass over the training samples (or over a large finite subset of the training samples). An alternative for large and nonlinear problems is *online training*, where each training example is considered in turn, and the matrix $\mathbf{W}$ is incrementally modified as needed to give a better answer for each example.

Frequently, the method of *stochastic gradient descent* is used, moving through the space of trainable parameters (the weights) in a direction estimated to best reduce an energy function or some such measure of the system's error, as estimated from one or a few training examples. The stochastic gradient descent method was originally developed in the 1960s for training linear networks, for example in the Widrow–Hoff *least mean square* (LMS) method for adaptive filters (Widrow and Hoff, 1960), and has been extended for training all sorts of nonlinear learning machines.

24.4 Nonlinearities at the Output

The linear perceptron's output is often converted to a hard decision, by comparison to a threshold, as mentioned above. If the perceptron is used as a classifier, the goal is to get the output of the linear part of the perceptron above the threshold (or above zero) for patterns of one class, and below for the other. The target outputs are typically represented as 0 and 1, or −1 and +1. The *sign* or *signum* function $\mathrm{sgn}(y)$ is the nonlinearity that maps y to −1 or +1 depending on its sign. Minimizing the mean squared error with respect to such targets is equivalent to minimizing the proportion of classification errors, but linear least squares techniques are not suitable for finding the solution, due to this nonlinearity.

The perceptron training algorithm, proposed by Rosenblatt in the original perceptron work, incrementally modifies the weight matrix when classification errors are encountered in the training set. If the set of training patterns is linearly separable—that is, if there exists a hyperplane that correctly separates the points into classes—the algorithm is guaranteed to converge to a weight matrix that separates the classes with no errors. Consider a single-output perceptron (multi-output perceptrons are treated as several of these in parallel). Treating outputs and training targets as values −1 or +1, the perceptron training rule is very simple: for any input pattern for which the output is not equal to the target, modify each weight in the direction that moves the output toward being correct.

At training step n, if $\mathrm{sgn}(\mathbf{Wx}) \neq t$, then

$$w_j(n+1) = w_j(n) + tx_j$$

Example Problem for Neural Networks

Consider this example problem: we want to classify talkers as male or female, based on a single vowel utterance, using only measurements of the first and second formant frequencies (F_1 and F_2). For data, we use an online database of vowel data from North Texas talkers (Assmann and Katz, 2000).

With data from the ten adult male and ten adult female talkers, using only the averages between initial and final F_1 and F_2 values for 12 different vowels, we construct training and testing sets of two-dimensional features and two-class (one-bit) targets. The first five males and the first five females, repeating each vowel an average of 10 times each, make a training set of about 1200 points; the second half of the talkers make a similar-size testing set.

First, we train a simple perceptron: three trainable weights connect the F_1, F_2, and constant inputs to the decision output. The resulting decision boundary is necessarily a straight line in F_1–F_2 space, as shown in Figure 24.2

It is evident that there is a lot of confusion in the middle of the F_1–F_2 space. Males tend to have longer vocal tracts, and hence lower formant frequencies, than females, but some vowels also have lower F_1 and F_2 than other vowels, so the middle of the feature space has a confusing mixture of clumps of male and female sample points. If we added pitch (F_0) as a feature, the problem would be relatively easy, since pitch alone is enough to distinguish male from female with better than 95% accuracy. We use the F_1–F_2 example because the nonseparable nature of the problem helps to illustrate some of the issues in machine learning. Turner et al. (2009) provide a much more in-depth discussion of the relations between formant frequencies, vowel identity, and talker gender, vocal tract length, and pitch.

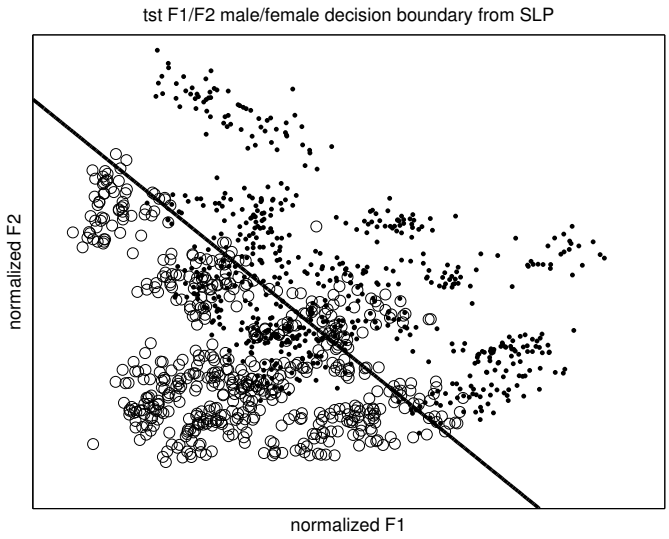


Figure 24.2: A feature-space map showing the decision boundary of a single-layer perceptron (SLP) classifying talker gender from F_1–F_2 data. Data points from the testing set are shown, with males as circles and females as dots (corresponding training points can be seen in Figure 24.6). Near the decision boundary, many classification errors are made: 267 errors of 1200 items in the testing set. This perceptron also makes 298 errors on the training set. The first two formant frequencies are apparently not quite sufficient to distinguish male from female talkers.

Quick Linear Training in MATLAB

In MATLAB, training a linear perceptron is a one-liner, if the training inputs and targets are already gathered up into matrices `x` and `targets`, with a column per training sample:

```
W = x \ targets;  % Least-squares training
y = W * x;  % y should now be close to targets
```

If there exists a matrix that will map all the inputs to all the targets exactly, this will find it, in which case `y` will be equal to `targets`. More generally, MATLAB's matrix division operator will find the matrix that maps `x` to a `y` near `targets`, minimizing the total squared error of `y` relative to `targets`. This formulation is easy to set up and solve as a least-squares problem in systems other than MATLAB, too, of course.

This kind of least-squares training is good for *regression* problems: learning a function that approximately maps input values to the training values. But perceptrons, and their training algorithm, were actually designed for *classification*, where the goal is to minimize a count of misclassifications, not a sum of squared errors. Whether for regression or classification, we typically use nonlinearities in our perceptrons, such that training is not quite this easy. That is, the *loss function* that we are minimizing may not be the total squared error of the linear part of the perceptron operation, so a different method for minimizing the loss needs to be found.

otherwise, that is, if the output is already correct, leave the weights as they are:

$$w_j(n+1) = w_j(n)$$

In this *perceptron rule*, the multiplication of the input x_j by the target t causes the weight update to go in the right direction; if the target is positive, and the input is positive, then the weight update will be positive, which means the output will be more likely to be positive for such patterns in the future. The size of the update is proportional to the input value, so the biggest changes will be made where they will have the biggest effect on the weighted sum.

In generalizations of the perceptron rule, one usually uses a learning-rate factor, to keep the weight changes small compared to the weights. The perceptron performance can oscillate with training, when the data are not linearly separable; but with a declining learning rate, the weights can be made to converge to minimize a loss function that measures how far the misclassified points are from the decision hyperplane (Kashyap, 1970).

Other generalizations will typically replace the target value t, which indicates the direction in which to change the weight, with a factor proportional to the difference between the target and the actual output, and sometimes additional factors, to specify a direction and amount to change. For example, the Widrow–Hoff LMS rule, for linear perceptrons (adaptive linear networks, or *adalines*, as they called them), with learning rate η and output error $t - y_j$ is:

$$w_j(n+1) = w_j(n) + \eta(t - y_j)x_j$$

Other nonlinear functions, such as the popular logistic function, are commonly used at the output of the linear perceptron, as an alternative to the hard threshold (sgn) function. The logistic function:

$$\text{logistic}(y) = \frac{1}{1 + \exp(-y)}$$

maps y to a value between 0 and 1, so that the result can be interpreted as a probability—for example, an estimated probability of a class given the pattern, called a posterior probability. If this output is interpreted

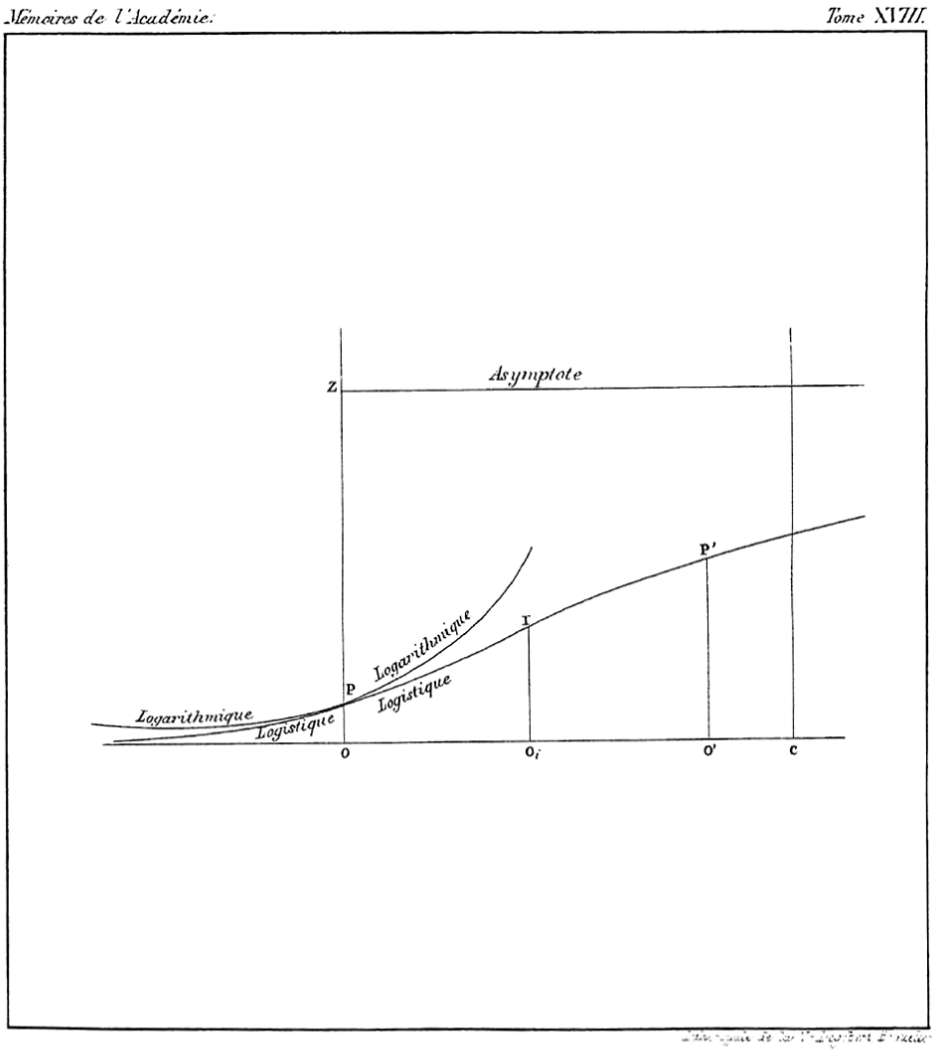


Figure 24.3: The logistic (*logistique*) function, which plays a big role in statistics and in artificial neural networks, was orignally derived as a population growth function (Verhulst, 1845), showing how exponential growth might be moderated as the carrying capacity of a region is approached.

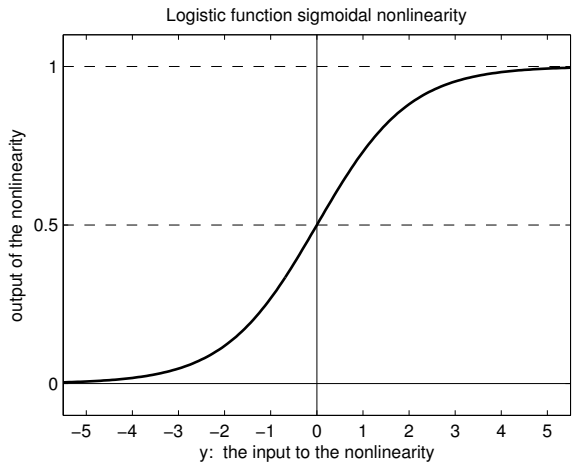


Figure 24.4: A popular nonlinearity at the output of a perceptron is the logistic function, which maps any value to an output between 0 and 1.

as a probability, then the y value, the input to the logistic function, is the log of the odds associated with that probability, or what is sometimes called the *logit*. Therefore, such a nonlinearity can represent a probabilistic model in which the log odds is a linear function of the inputs.

It is easy to do online stochastic gradient descent training with a differentiable nonlinearity at the output, by putting an additional factor into the LMS training rule, as we'll see later when we describe the training of multilayer perceptrons (MLPs).

The hyperbolic tangent (tanh) function is closely related to the logistic function, and is sometimes used instead. It is symmetric about 0, from -1 to $+1$, and has a slope of 1 at the origin. Though the logistic and tanh are completely equivalent in their functional power, the tanh, with its output centered on 0, has been found to lead to faster training convergence in MLPs (Orr and Müller, 1998).

These nonlinearities, and other *sigmoids* (s-shaped curves) also have the property that their output can be made to approach a hard-decision nonlinearity, like that of the original perceptron, by scaling up the weights. Thus, perceptrons using these functions can be made to closely approach anything that can be done with nondifferentiable hard-decision nonlinearities; but they can also do more, such as be trained to estimate probabilities.

24.5 Nonlinearities at the Input

A powerful way to enhance the capability of a single-layer perceptron is to provide it with more input features, as nonlinear combinations of the existing inputs. Tom Cover (1965) pointed out that "A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated."

For example, if we provide all the second-order multiplicative combinations (squares and pairwise products) of the features, then the perceptron's decision boundaries, hyperplanes in the higher-dimensional feature space, correspond to arbitrary quadratic boundaries (such as ellipsoids) in the original feature space. In our two-dimensional example, the original features F_1 and F_2 are augmented by F_1^2, F_2^2, and F_1F_2, resulting in a small improvement as shown in Figure 24.5. With more dimensions, the number of added features grows quadratically.

An advantage of this approach is that the perceptron itself is still linear, so can be trained by simple

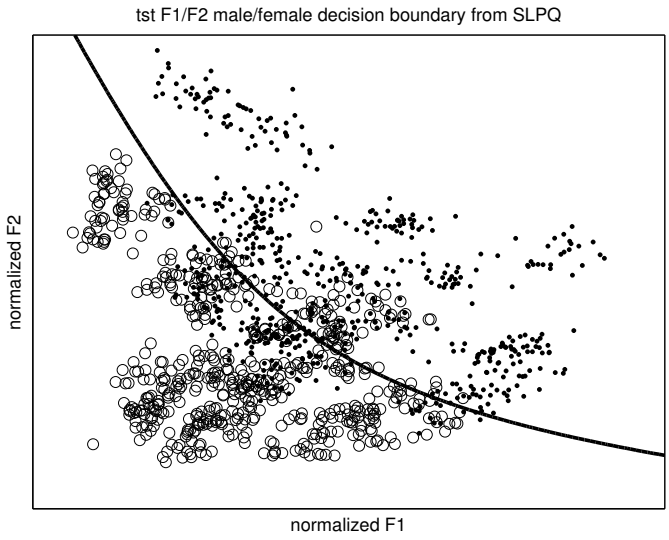


Figure 24.5: A feature-space map showing the decision boundary of a single-layer perceptron classifying talker gender from F_1–F_2 data, when the input feature vector is augmented with three quadratic dimensions (two squares and a cross term). The nonlinearites at the input help, but only a little; test-set errors are reduced to 247, from the 267 of Figure 24.2.

linear least-squares matrix methods. A disadvantage is that the dimensionality of the input can get very high, increasing the cost of weight training, storage, and application.

There is no limit to the number of new features that can be derived via nonlinear functions of the original features—but just because you can doesn't mean you should. Giving the perceptron the power to separate your training samples will often lead to a machine that is overfitted to the training data, and that doesn't generalize well to novel test data. The problem of building a classifier that separates noisy data points correctly into two classes is said to be *ill-posed*. That problem is addressed by other training methods, that include mechanisms to *regularize* the problem. As Poggio, Torre, and Koch (1985) say, "We will use the general term *regularization* for any method used to make an ill-posed problem well-posed." In particular, regularization encourages simpler approximate solutions, often leading to a unique optimum. For the linear classifier with expanded input space, the method of *regularized least squares* is easy to apply and usually effective (Rifkin and Lippert, 2007), though it is more suited to regression than to classification. See Section 24.9 for more on regularization, including an example with eighth-order input terms generated from the two formant-frequency inputs.

24.6 Multiple Layers

If the outputs of a perceptron are provided as inputs to another perceptron that provides the system outputs, the resulting system is called a two-layer perceptron. More generally, the number of layers can be more than two. As long as it contains no feedback loops, an interconnection of perceptrons is known as a *multilayer perceptron* or MLP.

If the layered perceptrons are linear, so is the MLP, in which case it can't do any more than a single-layer linear perceptron could do. But with nonlinearities at the outputs of each perceptron, MLPs become all-powerful, in the sense that they can learn very complicated decision boundaries between classes, given enough layers and enough dimensionality of the layers.

The nonlinearities traditionally used in MLPs are the logistic function and the tanh. These nonlinearities are used partly because they result in a very simple implementation of the gradient descent training algorithm known as *error back-propagation*, or simply *backpropagation*. More recently, the half-wave-rectification or positive-part nonlinearity has become popular for use in deep (many-layer) networks, in which a neuron with this nonlinearity is referred to as a *rectified linear unit* (ReLU).

24.7 Neural Units and Neural Networks

At a slightly lower level than layers and matrices, we can look at individual *neurons* or *units* of a network. Each such unit performs a single weighted sum, based on its weights and its inputs, and then, in the general case of nonlinear units, maps that sum through a fixed nonlinearity.

The original perceptron unit, with its hard thresholded output, is the same as what is known as the McCulloch–Pitts neuron model, or a threshold logic element. The output is a logical value, true or false, 1 or 0, based on whether a weighted sum of inputs is above or below a threshold.

The neural unit commonly used in the MLP uses the logistic function nonlinearity instead. The backpropagation training algorithm that makes multiple layers usable requires that the nonlinearity be differentiable—because backpropagation uses the *chain rule* of differential calculus—so the threshold function is not suitable for use in MLPs.

Networks of these units may include arbitrary connections, from the output of any unit to the input of any other. These networks can be represented as directed graphs, with nodes for the neuron units and edges for output-to-input connections. For simple and multilayer perceptrons, these graphs do not contain any cycles; they are DAGs—directed acyclic graphs. If the graph has cycles, the network is said to be recurrent; it cannot be arranged in layers, and the outputs cannot be computed from the inputs with one pass over the units. Consideration of recurrent networks is beyond our scope here, but they may be very useful as systems for learning dynamic trajectories in feature space, for example for learning sequences in speech or music.

A unit's output is often called its *activation*, and the nonlinearity is called the *activation function*. Inputs are also often represented as unit activations. It can help the uniformity of implementation of the evaluation and training algorithms to store all input, hidden, and output activations in a single form, such as in a single array indexed by unit number.

Depending on the output interpretation desired, sometimes the output-layer units have the activation functions omitted, so the final layer is a linear network. For our discussion here, we'll assume that all hidden and output units are alike and nonlinear.

24.8 Training by Error Back-Propagation

The backpropagation learning rule was discovered and published by several people in different fields in the 1960s and 1970s, but it didn't really affect the field of artificial neural networks (ANNs) until much later. Its popularization by Rumelhart and his colleagues in the 1980s (Rumelhart et al., 1986; Rumelhart and McClelland, 1987) led to an explosion of successful applications of MLPs, and to a proliferation of other variations on the theme of ANNs.

Backpropagation can be applied as an online learning rule, like the perceptron rule of Rosenblatt, and the LMS rule of Widrow and Hoff. These training rules work by making small changes to the weights on every training sample, in a direction that will reduce the error for that sample. Alternatively, changes can be accumulated over batches of training samples, and applied less often. The intended result of the small local changes, or the accumulated batch changes, is that the average error over the training population will be reduced as the weights evolve. Under some very broad conditions that we won't go into, such rules do

Multilayer Perceptron Examples

If we apply a much more powerful neural network, an MLP with two hidden layers, with six neurons per hidden layer, we produce a classifier that does a much better job of separating the training data into male and female regions of the feature space, as shown in Figure 24.6. This map illustrates the power of an MLP to learn complicated decision boundaries from training data. It also illustrates the problem of overfitting: when we test it on the test set, we find it gets even more errors than the simple linear perceptron did!

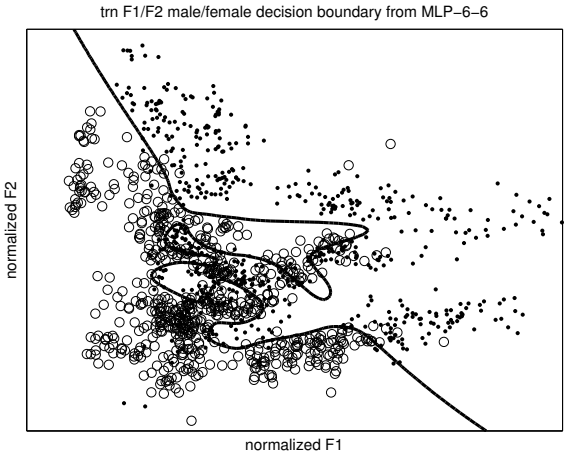


Figure 24.6: A map showing the decision boundary of a "too powerful" MLP, classifying talker gender from F_1–F_2 data. Data points from the training set are shown, using the same symbols as in Figure 24.2 (where the testing data points can be seen). The complicated decision boundary does a fairly good job of separating the training talkers into male and female, making only 186 errors on the training set. But it makes 290 errors on the testing set, which is worse than the simple linear perceptron.

A key problem in machine learning is to find a good compromise between the powerful capability of a trainable system to model the training data, and the need to generalize without over-fitting. One way to do that is to use a network with just enough trainable weights, or just enough modeling power. The network result shown in Figure 24.7 is an example of that approach: on our example problem, a net with only one hidden layer of 5 units makes fewer errors on the test set than do larger and smaller nets.

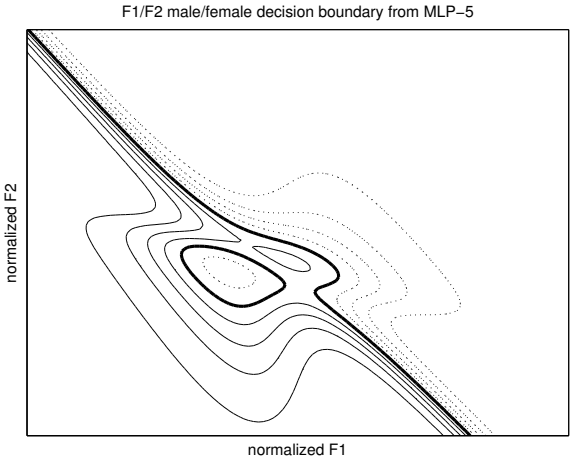


Figure 24.7: A map showing the decision boundary of a smaller MLP, classifying talker gender from F_1–F_2 data, along with contours of estimated class probability, in multiples of 10%. This network makes only 214 errors on the test set.

work very robustly, though sometimes convergence requires a very slow learning rate. However, even if they tend to reduce the error and converge, they don't necessarily converge to the lowest possible error, or global optimum, when the systems are nonlinear.

Like LMS, the backpropagation training algorithm simply changes each weight by an amount proportional to that weight's differential effect on a loss function, such as sum of squared output errors.

Starting at the output, for each unit, indexed by j, we define a differential sensitivity, d_j, proportional to the derivative of the total squared error with respect to that unit's activation z_j. At the output units, these sensitivities are proportional to the errors of the output activations relative to the targets:

$$d_j = z_j - t_j$$

where t_j is the target for unit j, and z_j is its actual output. Next we find the sensitivity of the output error to the pre-nonlinearity sum, the y_j. For the logistic function, due to its special form, this sensitivity is:

$$\frac{\partial d_j}{\partial y_j} = \frac{\partial z_j}{\partial y_j} = z_j(1 - z_j)$$

That simple factor $z_j(1 - z_j)$, and the resulting computational economy, is what makes this particular nonlinearity, and the related tanh function, so popular. In the case of a tanh sigmoid instead of the logistic, with limits -1 and $+1$ instead of 0 and 1, the derivative is of the same simple form: the product of distances from the two limits, $(z_j + 1)(1 - z_j)$.

Sensitivities of the total squared error to weights, and to lower-layer activations, are easily found in terms of this sensitivity to the pre-nonlinearity y_j, using an algorithm that proceeds from the output and propagates sensitivities back through the layers. The details are widely available in books, and in open-source code in every popular programming language (Haykin, 1994; Krogh, 2008).

In summary, for each training sample, a forward pass over all the units is used to find their activations, and then a backward pass over all the units finds their deltas and updates their weights. A table lookup is often used for the nonlinear function, once per neural unit, and the rest is just a handful of multiplications and additions per weight. All of this is repeated, not just for every training sample, but typically for a few or many *epochs* of training on the same training set, with different random orders on each epoch, and with a gradually decreasing learning rate. The weights tend to converge to a local optimum.

Since the random incremental updates tend to cause the weights to wander in a very noisy way away from the local optimum that they are trying to converge to, a "momentum" term is sometimes used to smooth the weights. The momentum feature in typical backpropagation training is no more than a first-order lowpass filter on each weight value. Alternatively, weight-training updates can be batched to reduce noise; adding up all the changes from a batch, using a lower learning rate, is much like running a smoothing filter that applies updates gradually over time. When the network is evaluated for all of the training samples in a batch (which might be all of the training samples, or a random subset, or a group on some fixed number), and then the back-propagated weight updates are computed, some intermediate storage is needed. Storing all the activations for each training sample is feasible when the batches and networks are not too big. The MATLAB function *bbackprop* (a "per-epoch backpropagation training" by Dale Patterson) does a batch per epoch this way. For the examples here, I have modified it to do random subsets and a declining learning rate, which helps it explore the weight space better and settle to a good local optimum. It takes less than a minute to run 30 000 epochs of 1200 samples in typical experiments, randomly choosing about half of the samples in each epoch.

24.9 Cost Functions and Regularization

Weight-training algorithms for neural networks generally correspond to approaches to optimizing a *cost function* on the model and training data. For example, the linear least-squares method corresponds to solving a system of linear equations to minimize a cost that is the sum of squared errors at the outputs, in a model that makes predictions by linear combinations of inputs. Other methods may attempt to approximately optimize the same cost function via a different approach, such as by an online stochastic gradient descent algorithm. Adaline training (Widrow and Hoff, 1960) is an example of a stochastic gradient descent algorithm.

The least-squares cost function as described does not include any direct cost term from the model—the weights—so the weights can easily get very large. A popular regularization strategy is to add a cost term based on the weight values. If we include a cost (or *penalty*) term proportional to the sum of squared weight values, then we can again cast the optimization as a linear least-squares problem and again get easy batch or online training methods. The resulting weights will be smaller, and the error at the output larger, compared to the nonregularized case, as a compromise that depends on the proportions used in the cost function. This approach is known as L2 regularization (since it includes a cost based on the L2 norm of the parameters), and it can be very useful in avoiding over-fitting to the training data, by keeping the weight magnitudes smaller.

The method of *regularized least squares* (RLS) incorporates such an L2 weight cost (Rifkin and Lippert, 2007). It provides an efficient batch weight computation for one-layer classifiers, generalizing the one-line least-squares weight computation given above. In MLP training by backpropagation, an L2 regularization is often incorporated by the method of *weight decay*: adding small weight updates toward zero, proportional to the size of each weight.

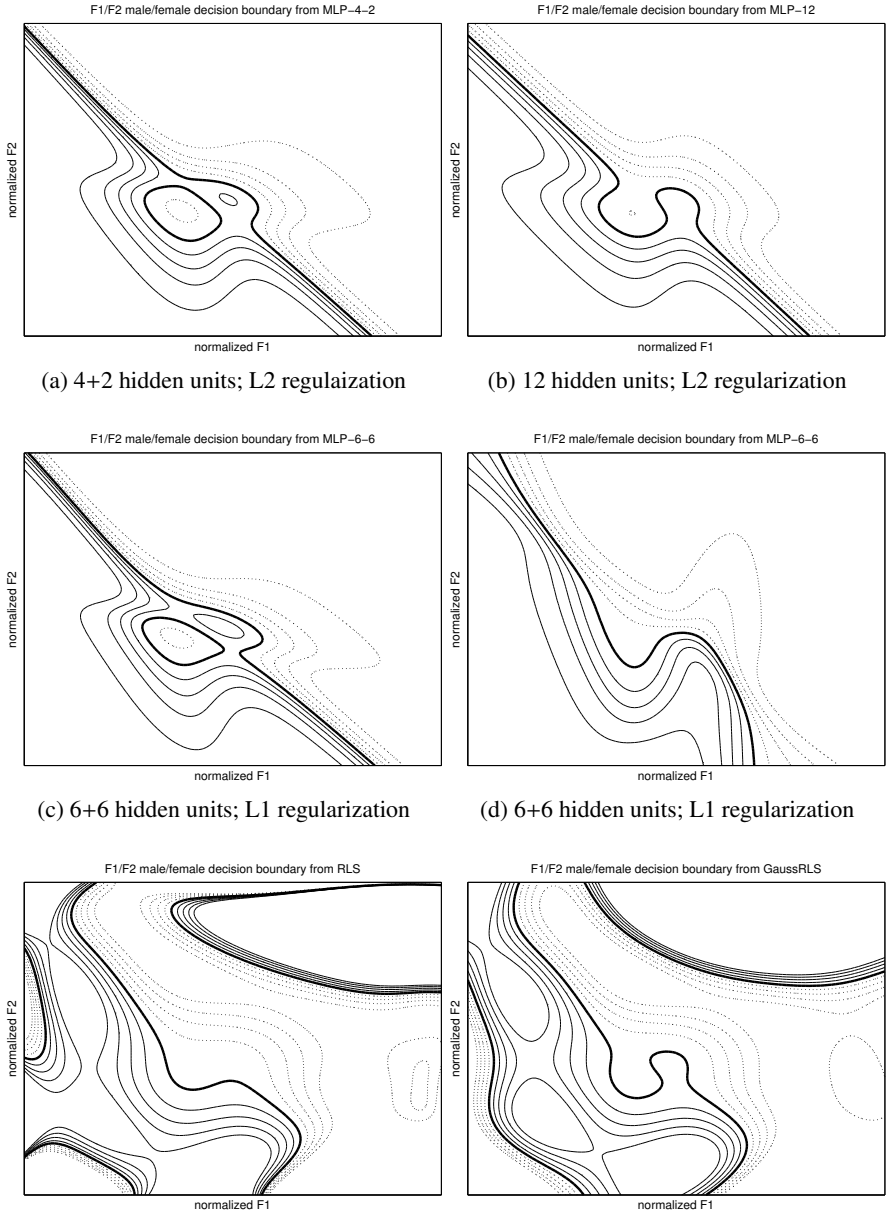
For our example classification problem, powerful classifiers trained with L2 regularization do much better than the nonregularized example. Whether we use the MLP or the RLS with polynomial or other nonlinear input-space expansion, the results are comparable: 206 to 218 errors on the testing set (much better than the 290 errors without regularization).

Another popular cost term is the sum of absolute values of the weights. This L1 regularization doesn't lead to such simple training algorithms, but it tends to push small weights to zero, while allowing some very large weights, compared to L2 regularization. The resulting weight matrices tend to be *sparse*, in the sense that they have only a small minority of nonzero coefficients.

When there is a good reason to think that a sparse weight matrix is a good representation of some underlying reality—for example an expectation that most inputs have no predictive value for most outputs—L1 regularization can be a good way to reach a robust network model. This notion of what's expected is formally captured in statistics as a *prior distribution* of model parameters. Much of the transition from neural networks to statistical learning theory is about how to handle such prior distributions—to get training algorithms that generate systems and models that are more robust by conforming more closely to what's expected. Conversely, results from statistical learning theory are sometimes incorporated back into the neural-network paradigm via modified cost functions.

24.10 Multiclass Classifiers

When an input is to be classified with respect to several classes, rather than a single binary decision, there are several approaches that can be used. A multi-output neural network can be trained to effectively approximate the probability of each class given the input data (including the effect of a prior probability, depending on the training) (Gish, 1990). This multi-output net is about the same, except for sharing internal structure, as a set of independent one-versus-all or class-versus-background binary classifiers. This approach has been shown to work as well as more complicated methods, on a wide range of problems (Rifkin and Klautau, 2004). See



(a) 4+2 hidden units; L2 regulaization

(b) 12 hidden units; L2 regularization

(c) 6+6 hidden units; L1 regularization

(d) 6+6 hidden units; L1 regularization

(e) RLS with 8th-order polynomials

(f) RLS with Gaussian kernel

Figure 24.8: Decision boundaries of different neural nets and least-squares classifiers, all of which yield 206 to 218 errors on the testing set. Panels (a) and (b) are two-hidden-layer and one-hidden-layer nets trained with L2 regularization (weight decay). Panels (c) and (d) are identical two-hidden-layer structures trained with L1 regularization from different random starts; in both cases, enough weights go to zero to reduce them to effectively the structure in panel (a), with only 4 active units in the first hidden layer and 2 in the second. Panels (e) and (f) are examples of the modern regularized-least-squares method operating on a large nonlinearly-expanded input space, using polynomial expansion for (e), and Gaussians at the training points in (f). The RLS methods have formed additional decision regions to accommodate the training outlier points seen in Figure 24.6.

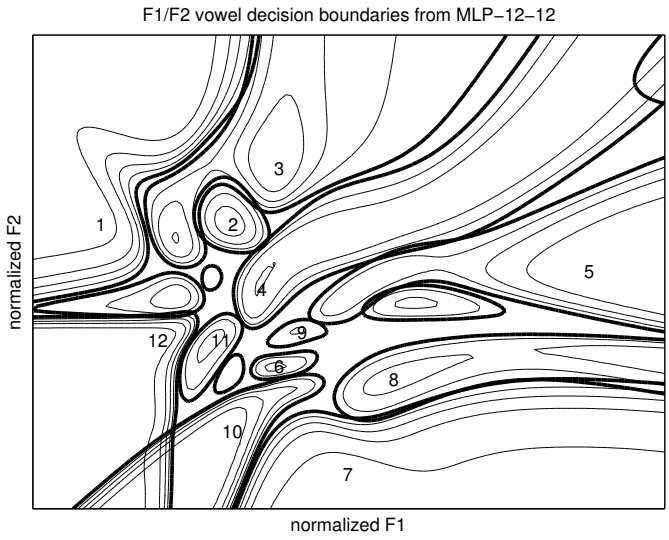


Figure 24.9: A map showing the decision boundaries of a 12-output MLP with 12+12 hidden units, classifying vowel identity from F_1–F_2 data for a mixed-gender training population. Vowels 2, 6, 9, and 12 have two regions each, but only one of each is labeled. Probability contours for 50%, 60%, 70%, 80% and 90% are shown, with the 50% contour darker. The probability estimates are not constrained to add to 1, and in regions with no training data they often add to more than 1, as the crossing contours show. The first-choice vowel classification accuracy with this net is about 50% on the testing set.

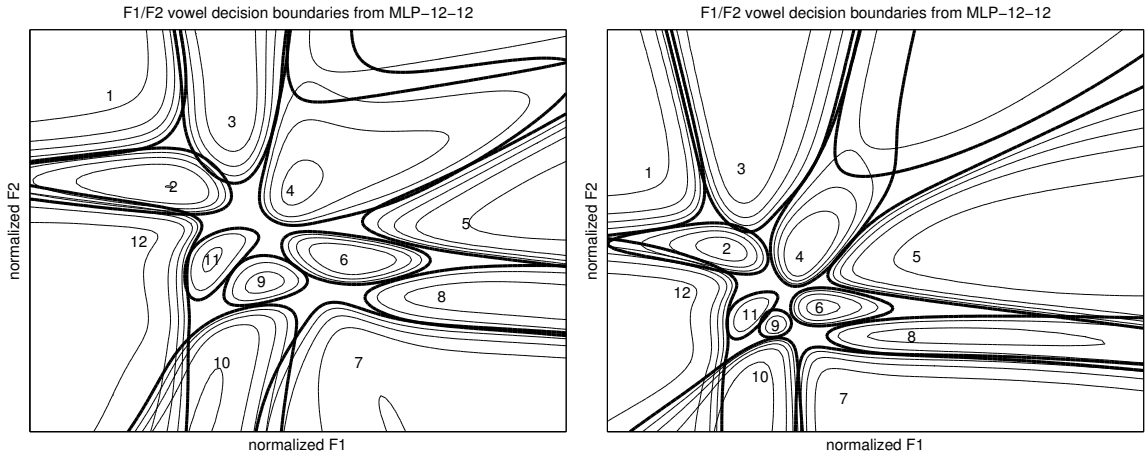
Figure 24.10 for an example of a 12-class vowel classifier using two hidden layers of 12 units each, with L2-regularized backpropagation training.

Sometimes the training of a multiclass network uses a training strategy resembling the original perceptron training rule: perform a weight update only when the net makes a classification error on the training sample (or only when the correct output does not exceed all others by some margin). Alternatively, ordinary regression-like backpropagation can be used, training to push the correct-class output toward 1 and the other class outputs toward -1 (with a tanh sigmoid). These methods are not as different as they sound, since a correct classification with a sufficient margin will push the correct and wrong class outputs to near the extremes of the sigmoid output values, where the derivatives are small enough to cause little weight change. When the net is to be used to estimate posterior probabilities, rather than to make a class decision, other training modifications can be useful, such as updating at a lower rate toward negative targets than toward the correct positive target (Yaeger et al., 1998).

When estimating class probabilities, it can also be useful to provide "negative" examples in training—examples of data patterns not corresponding to any class, or members of an implicit "null" class. In the vowel classification example, a negative training sample might be one that comes from a nonvowel, or noise. If no such examples are available, it may still be useful to synthesize negative examples from a broad or uniform distribution of feature values, as we did for Figure 24.11. Doing so reduces the areas associated with the vowels to regions where there are enough positive examples to overcome the null background.

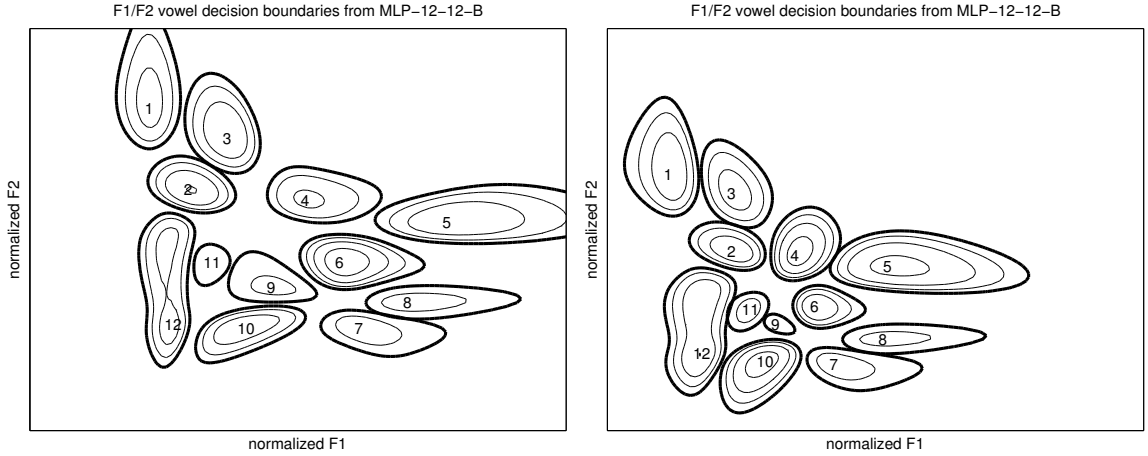
24.11 Neural Network Successes and Failures

For a long time, systems involving trained weight matrices were known as *neural networks*. That has been changing in recent years as more rigorous statistical underpinnings have been incorporated, and the neural



(a) Female vowel regions (b) Male vowel regions

Figure 24.10: Decision boundaries of a neural network that classifies three-dimensional features (F_1, F_2, and talker gender) into 12 vowel classes; on the left, the gender input is low for female, and on the right it is high for male. Vowels are much more separable with the additional input information; the first-choice accuracy goes from about half to about two-thirds with this additional input. A continuous pitch input is similarly helpful, since it tends to correlate with the talker's vocal-tract length at least as well as gender does, but such a 3D feature space is harder to illustrate.



(a) Female vowel regions (b) Male vowel regions

Figure 24.11: Decision boundaries as in Figure 24.10, but for a net trained with additional synthetic training points, uniformly spread over the F_1–F_2 plane, with targets all low to represent a null class, or no vowel. Now the estimated vowel class probabilities usually add to less than one, especially in regions with no vowel training points. Vowels 7 and 8 are the ones corresponding to the English words *hawed* and *hod*, which are highly confusable, often not distinguished by American English speakers, even in North Texas, so their responses overlap and never get to probabilities as high as 0.7. Vowel 9, in *herd*, is also very fuzzy in F_1–F_2 space, as the *r* sound is mostly signaled by a low third formant; it is especially confusable with vowel 11, *hood*.

network field has gradually become the *statistical learning* or *statistical ML* field.

Our treatment of perceptrons and neural networks has been mechanistic, rather than theoretical. This level of treatment has proven to be enough to allow such techniques to be widely adopted and applied to a huge range of applications, often with good success. But limitations and failures of this approach are also commonly seen. To avoid such problems, and to get the most out of this neural-network approach and other ML approaches, a more theoretical treatment, well grounded in the science of statistics, can be very beneficial.

In the 1980s and 1990s there were numerous reports of successes and failures of neural network applications, and a secondary literature of analysis of those outcomes. For example, one study of dozens of applications in which neural network approaches were compared to more conventional statistical approaches (Paliwal and Kumar, 2009) found that the neural nets usually worked better, but that the statistical approaches were usually not tried very seriously or "tuned" as much as the neural nets were. A more theoretical study (Frasconi et al., 1997) sought to find a deeper understanding of which approach would work better under what conditions. In recent years, many studies have focused on the essential equivalence of ML and classical statistical modeling methods. Focusing on such connections has enabled the discovery of new and improved ML methods.

Some other ML systems, such as *support-vector machines* (SVMs) (Cortes and Vapnik, 1995), are essentially the same, at one level, as simple perceptrons: trainable linear classifiers. But they are completely different in their conceptual underpinnings and in their weight-training algorithms. Methods (such as SVMs) that have proven to be robust on many problems are often those that can be cast as solutions to convex optimization problems, so that the training algorithm is guaranteed to converge to a global optimum (Sra et al., 2011). At very large scale (large numbers of training samples and high dimensionality), a range of new and powerful methods have been developed recently; at very large scale, stochastic gradient descent methods are again found to be effective, as they are more practical than trying to solve a global optimization problem using all the training data (Bottou and Bousquet, 2008).

24.12 Statistical Learning Theory

In a statistical learning approach, training data are treated as samples of random variables to be modeled. In a neural network, the weight matrix can be considered to be the parameters of such a statistical model. Connections of neural nets and other ML structures to statistical learning theory have been investigated at various levels (Evgeniou et al., 2000; Dunne, 2007).

Statistical approaches tend to divide into two main camps: *frequentist* versus *Bayesian* (though some argue that all approaches are really Bayesian, just not with very useful priors). Roughly, a frequentist (or maximum-likelihood) approach tries to find the model under which the data are most likely, while the Bayesian approach is to find the model that is most likely, given the data and the prior distribution on model parameters. When there is sufficient training data, and the assumed prior distribution is reasonable, these two methods will give similar results. When training data are scarce, however, the Bayesian approach will do a better job of combining information from the data with information from prior knowledge. Even very weak prior knowledge, such as some idea of how big the coefficients can sensibly be, is enough to get improved results that automatically account for how much training data is available.

The frequentist approach of maximum-likelihood estimation (MLE) is easy to cast as a cost-function minimization problem. Patching up the MLE approach by adding a regularizer term to the cost function is equivalent, in many cases, to using a Bayesian approach where the regularization penalty is proportional to the logartihm of the prior. And Bayesians will tell you that omitting the regularizer term is roughly equivalent to adopting a uniform prior distribution on the parameters. In the statistical regression literature, L2-regularized least-squares regression is known *ridge regression*, and L1-regularized least-squares regression is known as *the lasso*, or *lasso regression*.

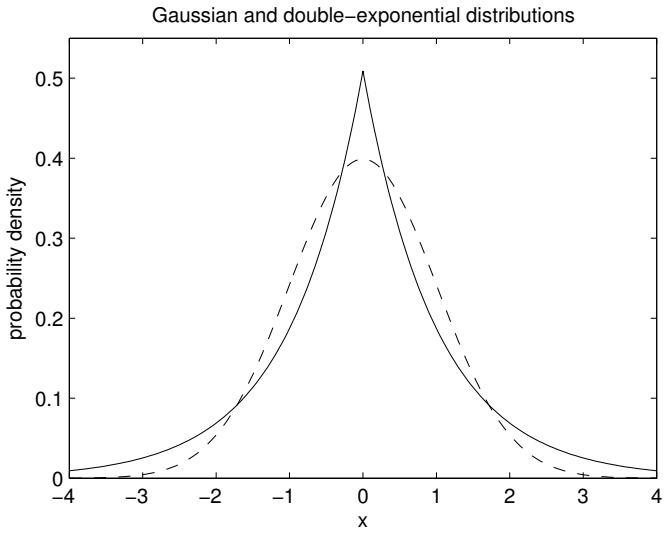


Figure 24.12: Gaussian (dashed curve) and double-exponential (solid curve) prior distributions on parameters, compared here with variances equal to 1, correspond to using L2 and L1 regularizers, respectively. The double-exponential distribution has many more values at zero, fewer small nonzero values, and more large values, compared to the Gaussian.

The fact that these two approaches tend to give similar results when there is plenty of training data does not save us from the curse of dimensionality. When the data dimensionality is high, the number of parameters to be learned is very high, and the amount of training data needed by an MLE method may be enormous. This is the domain in which we often find ourselves in machine hearing problems, where the data input may be something like a 100×100 (10 000-dimensional) auditory image. The assumption of sparsity, or a prior distribution with a sharp peak at and zero and long tails, can be invoked in a Bayesian approach—or an L1 regularizer can be added in a cost-function-optimization approach—to robustly estimate a huge number of parameters from limited training data. Many modern large-scale ML applications take advantage of some such sparsifying regularization technique. Where training data are plentiful, or very efficient training is needed, a simpler L2 regularizer is often preferred.

An applicable result in statistical learning theory is the *"no free lunch" theorem*—that no single learning or modeling method is best on all problems, and that when nothing is known about the nature of the data it doesn't much matter what method is used to model it (Duda et al., 2001). That is, whatever ML tools you have lying around may be "good enough," and probably can't be improved on, until you know more about the data and the problem. To do better, you need to study the data, and choose models that incorporate sensible priors based on insights into the data.

24.13 Summary and Perspective

Neural networks, including particularly MLPs trained by backpropagation, provide a popular and powerful mechanism for machine learning. With two or three layers of trainable weights and sigmoidal nonlinearities, these networks are commonly found to work well, to train a nonlinear mapping of input patterns to desired outputs, as defined by a training set. Training them to convergence is slow, and will often tend to overfit the training data, so other tricks, techniques, and variations have been developed to give better generalization and less training time (Orr and Müller, 1998). Modern work with *deep* networks, with four or more layers of

trainable weights, is often most successful with other nonlinear activation functions, such as *rectifying linear* (half-wave rectification) units (Glorot et al., 2011).

The MLP is particularly applicable when a classification problem is well characterized by a single input pattern of fixed dimensionality, as opposed to a trajectory through time or a structured description of variable dimensionality. For example, the MLP has been very successful in classifying characters, such as optically scanned or handwritten letters or digits, that have already been segmented out as characters. OCR and handwriting recognition systems typically have a trainable neural network classifier at their core, but also extensive other machinery around it to segment the input, to consider alternative segmentations, to incorporate language models, etc. (Yaeger et al., 1998). Similarly in machine hearing applications, the MLP may be a good classifier, but may need additional machinery before it to present patterns representing appropriate portions of an audio input, and machinery afterward to finish the job. For some tasks, we can simplify to the single-pattern paradigm: a whole sound file (for example, a recording of a musical performance of a bird's song) can be summarized by a pattern to be presented to a trainable classifier that goes straight to the answer: a song's genre, or mood, or species, for example.

In spite of its successes, the MLP runs into difficulties and limitations, such as the difficulty of successfully training the lower levels of many-layer networks to find useful features in the input data. Among the interesting modern alternatives are both single-layer networks with large random or structured input spaces, such as support-vector machines (SVMs) (Cortes and Vapnik, 1995), and many-level "deep" networks with the lower layers learning bottom-up on large datasets, unsupervised, instead of via back-propagation from training targets (Hinton et al., 2006).

There are a variety of other modern ML methods that can be applied to problems in hearing. Some can be understood as variations on neural networks, while others look very different, whether the underlying statistical models are similar or different.

Chapter 25

Feature Spaces

> Our sensations are simply effects which are produced in our organs by objective causes; precisely how these effects manifest themselves depends principally and in essence upon the type of apparatus that reacts to the objective causes. What information, then, can the qualities of such sensations give us about the characteristics of the external causes and influences which produce them? Only this: our sensations are signs, not images, of such characteristics. One expects an image to be similar in some respect to the object of which it is an image; in a statue one expects similarity of form, in a drawing similarity of perspective, in a painting similarity of colour. A sign, however, need not be similar in any way to that of which it is a sign. The sole relationship between them is that the same object, appearing under the same conditions, must evoke the same sign; thus different signs always signify different causes or influences.
>
> — "The Facts of Perception," Hermann Helmholtz (1878)

> Rather than represent the entire transformation from the set of input variables to the set of output variables by a single neural network function, there is often great benefit in breaking down the mapping into an initial *pre-processing* stage, followed by the parameterized neural network model itself. ... The use of pre-processing can often greatly improve the performance of a pattern recognition system, and there are several reasons why this may be so.
>
> — *Neural Networks for Pattern Recognition*, C. M. Bishop (1995)

The success of machine learning applications, whether in hearing or otherwise, often depends on a good choice of features to represent their input.

Feature extraction is typically a dimensionality reduction. In a sampled sound waveform, every sample is a dimension, in a vector-space conception of the data. Describing a segment of a sound by a spectrum might reduce the number of dimensions from thousands to a hundred or less. Further dimensionality reductions, designed to collapse most of the information into a few dimensions with low correlations between dimensions, are a popular approach in speech and music representation, as well as in image compression (Wintz, 1972). For example, the final transform that makes MFCCs from mel-frequency spectra, as described in Section 5.7, is typically used to reduce about 40 dimensions to about 12.

On the other hand, some feature extraction techniques increase dimensionality. If one thinks of an SAI frame as a feature vector describing a segment of sound, it is typically a dimensionality *increase*: from hundreds (of waveform samples per SAI frame interval) to tens of thousands of dimensions (SAI pixels per frame). For many machine learning systems, the SAI is therefore not the best input representation. It is better to think of it as analogous to the moving-image input to a machine vision system: a layer or two of feature extraction may still be needed to get to features that a learning system will handle most effectively.

In some cases, we do use feature spaces of very high dimensionality directly as input to learning systems. In *sparse* feature spaces, there are many dimensions, but few nonzero components for each feature vector. In such systems, feature extraction will often involve a dimensionality expansion.

25.1 Feature Engineering

The notion of *feature engineering* has been put forward to describe the process of designing representations that are especially suitable to an application domain (Yu et al., 2010), or especially suitable for input to a class of learning systems. Feature engineering can often be the key to making a good system. The whole machine hearing front end, with its cochlea model and auditory image formation, can be seen as feature engineering for sound processing. Or we can think of those front-end stages as producing an input from which we still need to extract good features for the application, much as the auditory cortex extracts features from what it gets from the midbrain.

On the other hand, feature engineering is sometimes criticized by machine learning experts, who point out that a good learning system should be able to work with whatever raw data is available, as effectively as with carefully crafted features or representations, and so feature engineering may be an unnecessary effort (Hamel and Eck, 2010; Humphrey et al., 2012). More principled and automatic techniques of feature extraction are appealing, but are not necessarily as good as some good feature engineering. As Turian et al. (2009) say:

> Compared to manual feature engineering, improved models are appealing because they are less task-specific. ... with logistic regression, adding quadratic filters was almost as effective as manual feature engineering.

Feature engineering can be viewed as a way to capture the knowledge of domain experts. In this respect, feature engineering is probably more leveraged, and more successful, than the older artificial-intelligence approach of knowledge engineering, which tried to capture knowledge much more explicitly, typically in the form of rules (Sebastiani, 2002). Let us accept our job as machine hearing engineers to be the designing of mappings of sound representations into forms that are well suited to machine learning systems that can address our intended applications. The feature engineering doesn't need to solve the problem—it just needs to make a good interface between what we know about hearing and what we know about the learning system that can address the task.

For example, representations such as *chromagrams* and *intervalgrams*, which we introduce in Chapter 27, are extracted from the *pitchograms* that we introduced in Chapter 21, in a pipeline as shown in Figure 25.1. These feature representations are engineered to capture what we know about the role of musical pitch in melody, to enable matching of songs on the basis of melody, and are engineered to work into a particular learning system that uses approximate-nearest-neighbor search via locality-sensitive-hash indexing (Indyk and Motwani, 1998).

Consider, as another example problem, the analysis of hyena laughs. Mathevon et al. (2010) used a 13-dimensional hand-crafted feature set, including the mean, min, max and coefficient of variability of fundamental frequency (pitch), duration, and several measures of spectral shape and variability. The extracted acoustic features were shown to provide a lot of information about the animal's sex, age, dominance, and individual identity—but could a high-dimensional, more "raw" set of auditory features have done even better? That kind of question is hard to answer without doing the experiment; in this case, the relatively small amount of training data available might tip the decision toward the hand-crafted features. Alternatively, an appropriately crafted regularization strategy might make the raw features work as well or better, even when data are limited.

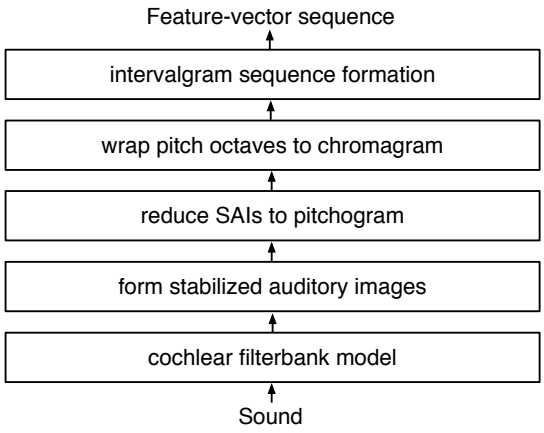


Figure 25.1: This multistage feature extraction pipeline for melody matching, described in detail in Chapter 27, maps into the first three levels of our *four-layer model* of Figure 1.5. Some of the stages here could alternatively be pushed into the uppermost layer, the machine learning system, with the possbiility that they would be end-to-end optimized for the application.

25.2 Automatic Feature Optimization by Deep Networks

The recent resurgence of artificial neural networks with multiple nonlinear layers has led to proposals that such networks should be good candidates to replace conventional feature extraction processes. Since each layer of such a network makes linear combinations of its inputs, followed by nonlinear compression and dimensionality-reducing "pooling" or aggregation of various sorts, a layer can potentially learn to do the same thing that a typical layer of hand-crafted feature extraction is doing, but with the linear combining weights optimized to the data. Humphrey et al. (2012) point out that in the field of music information retrieval, popular features for beat and tempo tracking, melody matching, etc., can be cast in the form of deep networks, so it is conceivable that a deep-net learning system could discover the same features, or better features, for such tasks.

Some automatic feature learning approaches use a Fourier power spectrum as their raw input representation of sound (Hamel and Eck, 2010; Sainath et al., 2013), though some even go as low as waveforms (Hoshen et al., 2015; Palaz et al., 2015). These systems all work well, given enough training data, but getting good performance from them requires some hand-crafted network architecture, including logarithmic or other compression layers. Essentially, they are filterbank-based sound analyzers that use backpropagation to adjust the filterbank parameters.

When we think about how to apply trainable networks in the four-layer modular structure of Figure 1.5, we have a range of opportunities. Probably the simplest, consistent with the modularization as described, is to do a moderate amount of engineered feature extraction at layer 3 (application-dependent feature extraction), and keep the learning in layer 4 (meaning extraction by machine learning). Alternatively, we can push the learning down into layer 3, or even down into layer 2 (brainstem – stabilized auditory image extraction) if we want to try to optimize those layers to what the problem's training set is telling us.

Getting such a deep network to learn the kind of structure that is engineered into the SAI is likely to be difficult and expensive, but may be worth trying. Whether automatically discovering the statistical regularities in real signals will lead to sound representations as good as what the auditory system generates at low levels—or better—is an interesting question that is being explored through techniques of sparse coding (Smith and Lewicki, 2006; Karklin et al., 2012), among other approaches.

Humphrey and Bello (2012) showed that they could get state-of-the-art performance on a chord recognition task from a deep net with just a spectrogram as input, as opposed to any chord-specific feature engineering. On the other hand, they also showed that using knowledge of music signals to generate extended training data to recognize transposition invariance and gain invariance led to better generalization and test accuracy—essentially undercutting the idea that a learning system might do as well without engineering domain knowledge into it.

In a recent comparative study of methods for acoustic scene classification (Barchiesi et al., 2015), the best system used recurrence quantification analysis (RQA) features, operating on MFCC sequences (Roma et al., 2013). RQA temporal pattern features had originally been engineered for the representation and comparison of electrocardiograms (Zbilut and Webber, 2006). This complicated nonlinear characterization of time patterns apparently captured something that the deep networks of the other systems were not able to learn on their own.

25.3 Bandpass Power and Quadratic Features

The popular approach of representing sound by the envelope or short-time power of various bandpass-filtered versions of the sound is closely related to generalized methods that we call *quadratic features*.

At the level of waveforms, quadratic features include the power (smoothed or aggregated mean square) of the output of a linear function of the input, such as a bandpass filter. They also include smoothed products of pairs of input feature dimensions, such as autocorrelation coefficients, which are smoothed products of waveform samples separated by a lag. These two classes of quadratic features are equivalent, in the sense that there is a linear relationship between the spaces, since power spectra are Fourier transforms of autocorrelation functions. In the space of quadratic features we can include other nonlinear combinations of inputs, such as the points in a stabilized auditory image that are generated by triggered temporal integration, which can be regarded as a quadratic correlation between an input signal and a nonlinearly sparsified trigger signal, as discussed in Chapter 21. Like linear features, quadratic features are often rectified; this additional nonlinearity breaks the equivalence of different quadratic feature sets such as spectra and autocorrelation coefficients.

Various machine learning techniques have been used to directly learn waveform filters to capture information about sound (Lee et al., 2000; Smith and Lewicki, 2006; Jaitly and Hinton, 2011). These techniques generally produce bandpass filters that resemble, at least approximately, cochlear filters. By paying attention to power or amplitude, and ignoring phase, they are learning quadratic features. In some cases they also learn dual-bandpass filters that are conjectured to be matched to vowels in the speech patterns used for training (Jaitly and Hinton, 2011). In all of these methods, the filter learning depends on a nonlinearity of some kind, not necessarily quadratic, to rectify the filter output or to choose a sparsification based on the outputs with highest power.

While using machine learning to design waveform filters has been shown to be a workable idea, the real power of this approach is that it can be applied at higher levels in the stack of auditory representations, to compute SAIs and beyond, and to more general input spaces, including visual (Rust et al., 2005) and text (Turian et al., 2009).

Andén and Mallat (2011, 2014) have described an architecture for cascading layers of bandpass quadratic feature extraction in the form of wavelet transforms with power detection and aggregation, or *wavelet modulus operators*. The neural net architectures for visual object classification by Poggio and his collaborators is analogous, in the sense of being layers of linear and *pooling* operators (Riesenhuber and Poggio, 2000; Serre et al., 2005; Bouvrie et al., 2009). In both cases, the use of quadratic features or pooling (essentially averaging of higher powers than squares) allows the representation of structure at large scales, with some degree of position invariance, while still being very selective for local patterns.

25.4 Quadratic Features of Cochlear Filterbank Outputs

McDermott and Simoncelli (2011) have explored a system of quadratic feature extraction for the representation of "sound textures." After a fairly conventional cochlear filterbank, with outputs representing compressed envelope versus time for each frequency channel, they use a secondary analysis through another filterbank on each channel, to analyze modulations. The features they extract are then mostly time averages of products of those outputs, including variance of each modulation frequency channel, and the cross-correlations of different modulation frequencies within a cochlear frequency channel, as well as cross-correlations between different cochlear frequency channels for a given modulation frequency. In addition, the variance and some other moments of the original cochlear frequency channels are included. With 32 cochlear channels, and 20 modulation channels each, they compute a total of 1515 statistics, more than 90% of which are quadratic features of the cochlear filterbank output (the others are first, third, and fourth moments of the cochlear channel outputs). Many more quadratic features could have been computed, but these were enough to capture most of the texture distinctions they investigated (though they were not good at capturing harmonic and rhythmic properties).

In our SAI approach, each point in an SAI can be computed as a quadratic feature of the cochlear filterbank output, if the correlation method of Section 21.3 is used. Compared to the approach of McDermott and Simoncelli (2011), there is a conceptual equivalence of the SAI to the outputs of the modulation filterbank layer (because an autocorrelation analysis is related to a power-spectrum analysis via a Fourier transform), but there are important differences. Most importantly, there is a difference of bandwidth in the interface between the cochlear fitlerbank and the second analyzer layer. For the SAI, we keep fine time structure consistent with the signals on the auditory nerve (more than 1 kHz bandwidth for the half-wave-rectified signals), whereas in the envelope-based model the bandwidth is restricted to about the filter-channel bandwidth, since envelopes are computed via Hilbert transforms. In addition, the techniques differ in effective time and frequency analysis resolution and output dimensionality, and in what secondary quadratic features are used. With the SAI, we frequently compute a summary SAI, or time-lag marginal, which is excellent at representing pitch or harmonicity, in contrast to the texture features that are not good for pitch and harmonicity. It is likely that the approaches can be combined to extend the results of McDermott and Simoncelli (2011) to a broader class of sounds.

Since the bandpass filtering and rectification in the cochlea, and the triggered temporal integration of auditory image formation in the brainstem, may be interpreted as two layers of quadratic/pooling feature extraction, it makes sense to think about using the same class of operations for subsequent layers, for extracting features about sound texture, pitch, tempo, rhythm, or whatever we need for an application. When we don't know what the application needs, machine learning techniques can be applied to automatically find good features within the same framework.

Such features in the auditory and visual nervous systems are also known as *stimulus-energy features* (Rajan and Bialek, 2013).

25.5 Nonlinearities and Gain Control in Feature Extraction

Kayser et al. (2003) did experiments in learning an appropriate neural-network nonlinearity for stably representing features of natural visual stimuli, and found that many of their neurons converged on a quadratic, or energy-detection, function. Training on Gaussian noise did not lead to similar nonlinearities; the non-Gaussian distribution of image features—or sound features—is necessary to justify such nonlinear processing.

Wang and Shamma (1994) proposed a system of auditory feature extraction based on feed-forward divisive normalization—a form of gain control, similar to what Heeger (1991, 1992) had proposed in vision. They noted that, "A common sequence of operations in the early stages of most sensory systems is a multiscale

transform followed by a compressive nonlinearity," and based their compression on a gain controlled by a broader spectral power estimate than the signal being normalized, in the context of a cochlear model. Schwartz and Simoncelli (2001) further developed and abstracted this concept, basing the normalization gain on a neighborhood of rectified linear feature filters such as oriented edge filters in vision or waveform filters in hearing:

> Signals are initially decomposed using a bank of linear filters. Each filter response is then rectified and divided by a weighted sum of rectified responses of neighboring filters. We show that this decomposition, with parameters optimized for the statistics of a generic ensemble of natural images or sounds, provides a good characterization of the nonlinear response properties of typical neurons in primary visual cortex or auditory nerve, respectively. These results suggest that nonlinear response properties of sensory neurons are not an accident of biological implementation, but have an important functional role.

They used square-law rectifiers in this work, but later used half-wave rectifiers (Wainwright, Schwartz, and Simoncelli, 2002). Subsequently, Atencio et al. (2008) have shown that neural responses (at least in auditory cortex) can be even better modeled by combining the two rectifier types, through a process that converges on the best first and second filters ("maximally informative dimensions," MIDs), each dimension having its own nonlinearity, to fit neural responses. The first MID in the spectro-temporal domain usually converges on a halfwave-like nonlinearity, and the second, to tune up the approximation of a given neuron, usually on something like a squaring nonlinearity. Both nonlinearities saturate, which may be inherent to the nonlinearities or may be the effect of a divisive normalization or other inhibitory or AGC effect. At lower levels, they found that a single MID, or a spike-triggered average, with half-wave rectification captures the neural responses adequately.

It is also likely that lateral inhibition or competition between feature detectors operates at all levels, causing saturation and dynamic range compression, along with sharpening—perhaps even a "winner-take-all" effect in extreme cases. Lateral inhibition may be the cause of some of the saturation seen in the MIDs.

Whether feature detection filters and nonlinearities are learned or constructed, these studies motivate and reinforce the idea that some kind of rectification and gain control is useful, and that a static quadratic nonlinearity may be too simple a model. These principles appear to be applicable at multiple levels of processing; we already incorporate them in the cochlear model.

25.6 Neurally Inspired Feature Extraction

Kouh and Poggio (2008) argue that a single class of canonical neural operator is sufficient to realize most of the types of responses found in the cortex and used in brain models for feature extraction, including sigmoid-like response from divisive inhibition (gain control), tuned Gaussian-like responses matched to specific input patterns, max-like pooling or winner-take-all responses, etc. Since these operations can then be put into alignment with neural circuits, models of visual object recognition and other cortical functions can be evolved in parallel with understanding of cortical structure and function.

In primary visual cortex (V1), "complex cells" typically respond to oriented or moving patterns, with spatial frequency preference and strong orientation preference, but without phase sensitivity. These responses are typically modeled by an "energy model," the sum of quadratic features of filters of different phases, or the max response of filters, pooled over a small region. Subsequent responses in secondary and later visual areas, modeled using similar structures, are tuned to more complicated visual patterns, or object parts. So far in the auditory system, our knowledge of cortical function and our corresponding models are rather shallow. But based on progress in the visual system, we can expect to see progress, if we come up with the right sound

stimuli to reveal the preferences of several levels of auditory "complex cells." It is possible that features learned automatically by our machine learning systems, on tasks in speech, music, and environmental sound, may point the way toward what auditory primitives to use in such experiments.

25.7 Sparsification and Winner-Take-All Features

A sparse vector is one with most components equal to zero. Such vectors can be efficiently represented by reporting only the locations and values of the nonzero elements. Operations on such representations can be very efficient, as discussed in Section 26.2.1. The opposite of sparse is dense; a dense vector has mostly nonzero elements. Sparsification is the conversion of a dense vector to a sparse approximation. Sparsification often works via an additive model, finding sparse combinations of kernels, or dictionary elements, that add up to a good approximation of an input signal (Smith and Lewicki, 2006; Karklin et al., 2012).

In an extreme case, a dense vector can be sparsified to just its largest element. This operation is called winner-take-all (WTA) coding. The coding may be just the index of the largest element, or may include the element's value as well.

In an early biomimetic vision device, I implemented a winner-take-all system, using lateral inhibition between analog light-sensing cells, in a feedback arrangement, to make a pattern sensor for an optical mouse (Lyon, 1981, 2014):

> What we would like is a way to get an interesting digital bitmap image reliably. A way to do this is to implement a form of "inhibition" between cells, so that after some cell outputs have gone high, all others are held low and the picture is stable from then on. This is somewhat analogous to the lateral inhibition in the retina of most biological vision systems (Békésy, 1967). It has the desirable effect of producing sensible images, almost independent of light level.

Lazzaro et al. (1989) developed an analog VLSI circuit for emphasizing the largest input, and making the other elements nearly zero. Essentially, the circuit exponentiates the inputs and then normalizes. The same method has developed into a popular output layer for neural networks: *softmax*, introduced by Bridle (1990). These methods approach a hard WTA at high scale factors into the exponential.

There are many variations and adaptations of WTA coding, all relying on the observation that most of the information in a dense vector is often captured by the few largest elements. For example, in a time waveform, a sparse set of onset events can capture most of the information about pitch and rhythm. Besides our application in stabilized auditory images, such features have been investigated as a step toward higher-level music analysis (Bello et al., 2005; Collins, 2005). The detection of onsets is typically done by a WTA-like peak picking on a pre-processed time waveform.

More recently, Yagnik et al. (2011) have applied WTA as a technique for approximating rank distances between vectors, and as a locality-sensitive hash code for finding approximate nearest neighbors based on such rank distances. The technique encodes the index of the highest element in each of a large number or random subset of dimensions of the original dense vector. These techniques have proved to be very powerful in vision (Dean et al., 2013) and other applications.

This WTA method of creating codes or "words" from dense vectors is closely related to using vector quantization (VQ) to generate codeword indices in a large number of codebooks on subsets or transformations of the original dimensions, as we do in Section 26.1.3. But by defining its quantization regions based mainly on the largest dimensions, the WTA methods pays more attention to outliers, as opposed to assuming a Gaussian-like distribution as L2 methods do.

25.8 Which Approach Will Win?

There are so many approaches to feature extraction, or constructing feature spaces, that this chapter has only scratched the surface. Whatever approach is taken for any given application and learning system, we expect to see a continuing tension between carefully crafted features and automatically learned features. It is hard to see how either end of this range of alternatives will ever dominate the other. Settings with larger training resources will support more learned features, while data-poor settings will require more careful feature engineering.

Chapter 26

Sound Search

> This task aims at identifying the pictures relevant to a few word query, within a large picture collection. Solving such a problem is of particular interest from a user perspective since most people are used to efficiently access large textual corpora through text querying and would like to benefit from a similar interface to search collections of pictures.
>
> — "A discriminative kernel-based model to rank images from text queries," Grangier and Bengio (2008)

This chapter is adapted from "Sound retrieval and ranking using auditory sparse-code representations" by Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C. Walters, and Gal Chechik (Lyon et al., 2010b).

Our first-reported large-scale application of the machine hearing approach is a sound search system (Lyon et al., 2010b) based directly on the PAMIR image search system described by Grangier and Bengio (2008). These are a form of "document ranking and retrieval from text queries," for image and sound documents.

While considerable effort has been devoted to speech and music recognition and indexing, the wide range of sounds that people—and machines—may encounter in their everyday life has been far less studied. Such sounds cover a wide variety of objects, actions, events, and communications: from natural ambient sounds, through animal and human vocalizations, to artificial sounds that are abundant in today's environment.

Building an artificial system that processes and classifies many types of sounds poses two major challenges. First, we need to develop efficient algorithms that can learn to classify or rank a large set of different sound categories. Recent developments in machine learning, and particularly progress in large-scale methods (Bottou et al., 2007), provide several efficient algorithms for this task. Second, and sometimes more challenging, we need to develop a representation of sounds that captures the full range of auditory features that humans use to discriminate and identify different sounds, so that machines have a chance to do so as well. Unfortunately, our current understanding of how the plethora of naturally encountered sounds should be represented is still very limited.

To evaluate and compare auditory representations, we use a real-world task of content-based ranking and retrieval of sound documents, given text queries. In this application, a user enters a textual search query, and in response is presented with an ordered list of sound documents, ranked by relevance to the query. For instance, a user typing "dog" will receive an ordered set of files, where the top ones should contain sounds of barking dogs. Importantly, ordering the sound documents is based solely on acoustic content: no text annotations or other metadata are used at retrieval time in this task. Rather, at training time, a set of annotated sound documents (sound files with textual tags) is used, allowing the system to learn to match the acoustic features of a dog bark to the text tag "dog," and similarly for a large set of potential sound-related text queries. In this way, a small labeled set can be used to enable content-based retrieval from a much larger, unlabeled set.

Several previous studies have addressed the problem of content-based sound retrieval, focusing mostly on the machine learning and information retrieval aspects of that task, using standard acoustic representations (Whitman and Rifkin, 2002; Slaney, 2002; Barrington et al., 2007; Turnbull et al., 2008; Chechik et al., 2008). Here we focus on the complementary problem, of finding a good representation of sounds using the given learning algorithm.

The study particularly evaluates a representation of sounds that is based on models of the mammalian auditory system, which we conjectured would be better than typical short-time spectral representations on this task. Unlike many commonly used representations, the auditory representation emphasizes fine timing relations rather than spectral analysis. We found that the auditory representation outperformed standard MFCC features. The following sections describe the auditory representation, the learning approach, and the experiments and results.

26.1 Modeling Sounds

We assume that a feature space that captures aspects of pitch, loudness, and timbre is what is needed to support indexing of sounds. Studies of statistical feature spaces for timbre have suggested that a mel-cepstral space works well (Terasawa et al., 2006); other studies use more complex spectral dynamics features extracted from spectra (Shamma, 2003). The auditory-image based representations of Dinther and Patterson (2006) were designed to separate different aspects of size of speakers and musical instruments: pulse rate and resonance scale, treating resonance scale as an aspect of timbre. Here we consider the auditory image model as a general sound feature, rather than using a feature space previously optimized for speech and music.

We focus on a class of representations based on sparse coding of stabilized auditory images like those described in Chapter 21, and compare these representations to mel-frequency cepstral coefficients (MFCCs), such as described in Section 5.7. The motivation for using auditory models follows from the observation that the auditory system is very effective at identifying many sounds, and this may be partially attributed to the acoustic features that are extracted at the early stages of auditory processing.

We extract features with a four-step process, illustrated in Figure 26.1: (1) a nonlinear cascade filterbank with half-wave rectified output; (2) strobed temporal integration, which yields a *stabilized auditory image* (SAI); (3) sparse coding of patches of the SAI; (4) aggregate all frame features to represent the full audio document.

The first two steps, filterbank and auditory image formation, are firmly rooted in auditory physiology and psychoacoustics (Lyon, 1990; Popper and Fay, 1992; Patterson, 2000). These correspond to the first two levels of our general four-layer system structure of Figure 1.5. The third processing step, sparse coding, is in accordance with some properties of neural coding (Olshausen and Field, 2004), and has significant computational benefits that allow us to train large-scale models. The fourth step takes a "bag of features" approach which is common in machine vision and information retrieval. The remainder of this section describes these steps in detail. These two steps together make up the third level, feature extraction, in our four-layer system model. The fourth system layer is the PAMIR machine learning model that is trained to do the task using these features.

26.1.1 Cochlear Model Filterbank

The first processing step is a cascade filterbank inspired by cochlear dynamics, known as the pole–zero filter cascade or PZFC (Lyon, 1998). It is much like the CARFAC model, in that it uses a cascade and produces a bank of bandpass-filtered, half-wave rectified output signals, but it is an earlier generation of modeling. The differences include:

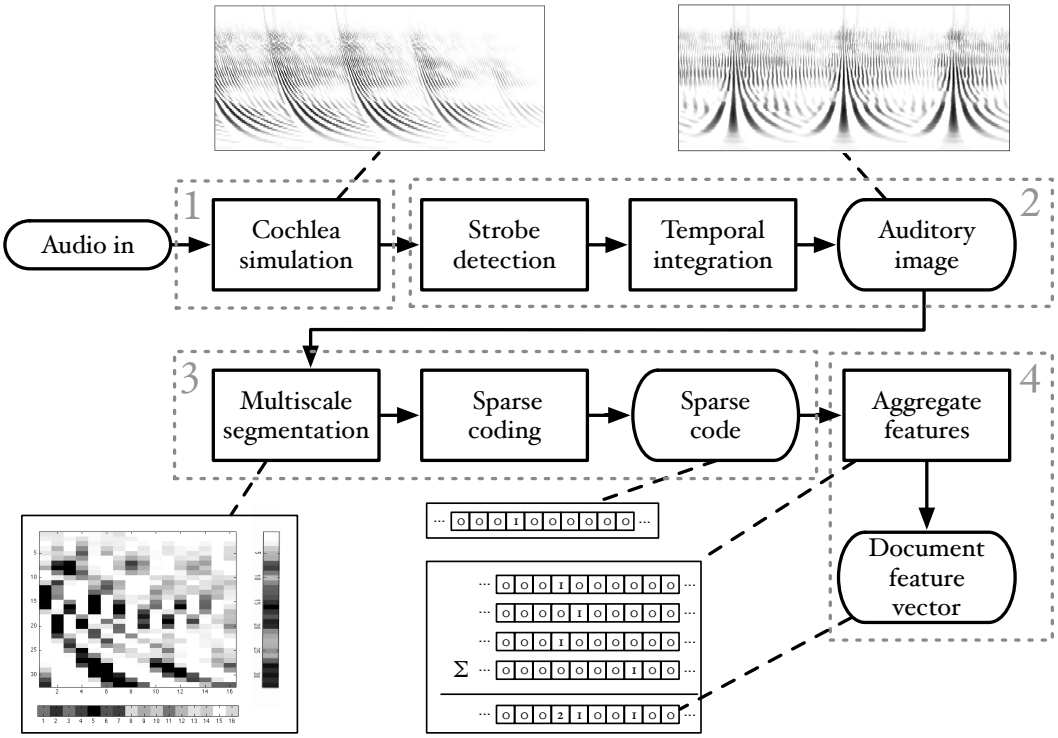


Figure 26.1: Generating sparse codes from an "audio document" using an auditory front end, in four steps: (1) cochlea simulation; (2) stabilized auditory image creation; (3) sparse coding of multiscale patches; (4) aggregation into a "bag of features" representation of the entire audio document. Steps 3 and 4 here correspond to the feature extraction layer in our four-layer system structure. From the point of view of the fourth layer, a PAMIR-based learning and retrieval system, this entire diagram represents a front end providing abstract sparse features for audio document characterization.

1. Like the PZFC auditory filter model in Chapter 13, the filter stages have movable poles but fixed zeros, as opposed to the CARFAC where the poles and zeros move together.

2. The implementation uses the direct-form digital filter stage of Figure 8.19, rather than a coupled-form filter, requiring more coefficient recalculation to make AGC updates to the filter damping.

3. The output rectification is a simple half-wave rectifier; there is no explicit IHC or OHC model.

Like the CARFAC, the PZFC in this project uses a four-stage coupled AGC feeding back to control pole damping.

26.1.2 Auditory Image Stabilization by Strobed Temporal Integration

The second processing step, *strobed temporal integration*, is based on a model of how humans use fine time structure in the perception of sounds, rather than purely on the known physiology of the auditory system (Patterson and Holdsworth, 1996). Using strobed temporal integration, this step "stabilizes" the signal, in the same way that the trigger mechanism in an oscilloscope makes a stable picture from an ongoing time-domain waveform.

The result of this processing is a series of two-dimensional frames of real-valued data, a "movie" of SAI frames. Each frame is indexed by cochlear channel number on the vertical axis and lags relative to identified strobe times on the horizontal axis, as described in Chapter 21.

The SAI that we use here has its zero-lag line at the center of the time-interval axis and is truncated at plus and minus ±26.6ms, which will represent signals that repeat more than 38 times per second, corresponding approximately to the lower frequency limit (maximum period) of human pitch perception.

With a lag dimension extending to 26.6ms, sounds with a repetition rate of above about 38Hz will lead to a stable vertical ridge in the image—a fair approximation to the limits of human pitch perception.

26.1.3 Sparse Coding of an SAI

The third processing step transforms the content of each SAI frame into a sparse code that captures common local patterns. Sparse codes have become prevalent in the characterization of neural sensory systems (Olshausen and Field, 2004). A sparse code is a high-dimensional vector $a \in \mathbb{R}^d$ that contains mostly zeros, and only n nonzero entries, $n \ll d$. As such it provides a powerful representation that can capture complex structures in data, while providing computational efficiency.

In a previous work (Chechik et al., 2008) we compared sound ranking systems that use dense and sparse features. The main conclusion from this comparison was that sparse representations obtain a comparable level of retrieval precision (proportion of retrieved sounds relevant to the query), but achieve it with only a fraction of the time needed for training. For instance, training on a dataset of 3431 files took only 3 hours instead of the 960 hours (40 days) needed for training a Gaussian-mixture model. The reason for the improved computational efficiency is (as we discuss below) that the learning approach we have chosen has computational complexity that depends on the number of nonzero values n, rather than the full dimensionality d. Building on these results, we focused the experiments with auditory features on sparse codes only.

The sparse code is based on identifying the typical patterns in the set of SAIs, and representing a sequence of frames by a histogram of the patterns that appear in it. This histogram is usually sparse, since each sound typically only contains a relatively small number of patterns. This *bag-of-patterns* representation is similar to the common use of a *bag-of-words* representation of text documents, or *bag-of-visual-terms* representation sometimes used in machine vision. However, unlike machine vision problems in which images are somewhat translation invariant—similar patterns could be found at different parts of an image—the SAI is indexed by frequency and lag time. As a result, different positions in the SAI correspond to auditory objects that are

perceptually different. To handle this, instead of looking for global patterns across the whole SAI frame, we search for more local patterns at different parts of the SAI. More specifically, the sparse coding step has two sub-steps: first, define a set of overlapping rectangular patches that cover each SAI frame; second, code each local region using its own sparse-encoder.

For selecting the rectangular local patches, we systematically tried several approaches and tested the precision obtained with each approach in the sound ranking task, as described below. We also tested a few approaches for representing the content of each rectangle, in a compact way. The details of these rectangle selecting procedures are described in the next section.

In the second sub-step, we represent all the vectors that represent the rectangular areas in an SAI using sparse codes. We tested two sparse coding approaches: *vector quantization* (VQ) (Gersho and Gray, 1992) and *matching pursuit* (MP) (Bergeaud and Mallat, 1995; Mallat and Zhang, 1993). In VQ, a dense feature vector is approximated by the closest vector from a codebook (in Euclidean sense). Once the best match has been chosen, the representation can be encoded as a sparse code vector, with a length equal to the size of the codebook, that consists of all zeros, except for a single "one" at the index position of the chosen codeword.

In MP, each vector (representing a rectangular patch, or box, of the SAI) is projected onto the codebook vectors, the largest projection is selected, the signed scalar value of that projection is added to the sparse vector representation (in the appropriate index position), and the vector valued projection is subtracted from the original vector, producing a residual vector. The process is then repeated until the magnitude of the largest projection becomes smaller than a given threshold. For both MP and VQ we learn individual codebooks tailored to represent the rectangles at each specific position in the SAI. The codebook is learned from the full set of rectangles in the data using a standard k-means algorithm. This yields a codebook that is tailored to VQ but that also works with MP. We tested several different codebook sizes (numbers of k-means clusters, or patterns, per rectangle), as described in Table 26.1.

Once each rectangle has been converted into a sparse code (using VQ or MP) these codes are concatenated into one very-high-dimensional sparse-code vector, representing the entire SAI frame. With the default parameter set, using vector quantization, a codebook size of 256 was used for each of the 49 rectangles, leading to a feature vector of length $49 \times 256 = 12\,544$, with 49 nonzero entries.

At each frame time, this feature vector of mostly zeros, with ones (in the VQ case) or amplitude coefficients (in the MP case) at a sparse set of locations, can be thought of as a histogram of feature occurrences in the frame. To represent an entire sound file, we combine the sparse vectors representing histograms of individual frames into a unified histogram—equivalent to simply adding up all the frame feature vectors. In the interpretation as a histogram, it shows how frequently each abstract feature occurs in the sound file. The resulting histogram vector is still largely sparse and is used to represent the sound file to the learning system described in the following section.

The process described in this section involves multiple parameters. In our experiments, we varied these parameters and tested how they affect the precision of sound ranking. More details are given in Section 26.4.

26.1.4 Rectangle Selection

To represent each SAI using a sparse code, we first defined a set of local rectangular patches that covered the SAI. These rectangles are used to identify patterns that are local to some part of the SAI, and they have different sizes in order to capture information at multiple scales. This approach is modeled on methods that have been used for image retrieval, involving various kinds of multiscale local patterns as image feature vectors.

We have experimented with several schemes for selecting local rectangles, and the specific method that we used is based on defining a series of rectangles whose sizes are repeatedly doubled. For instance, we defined baseline rectangles of size 16×32, then multiplied each dimension by powers of two, up to the largest

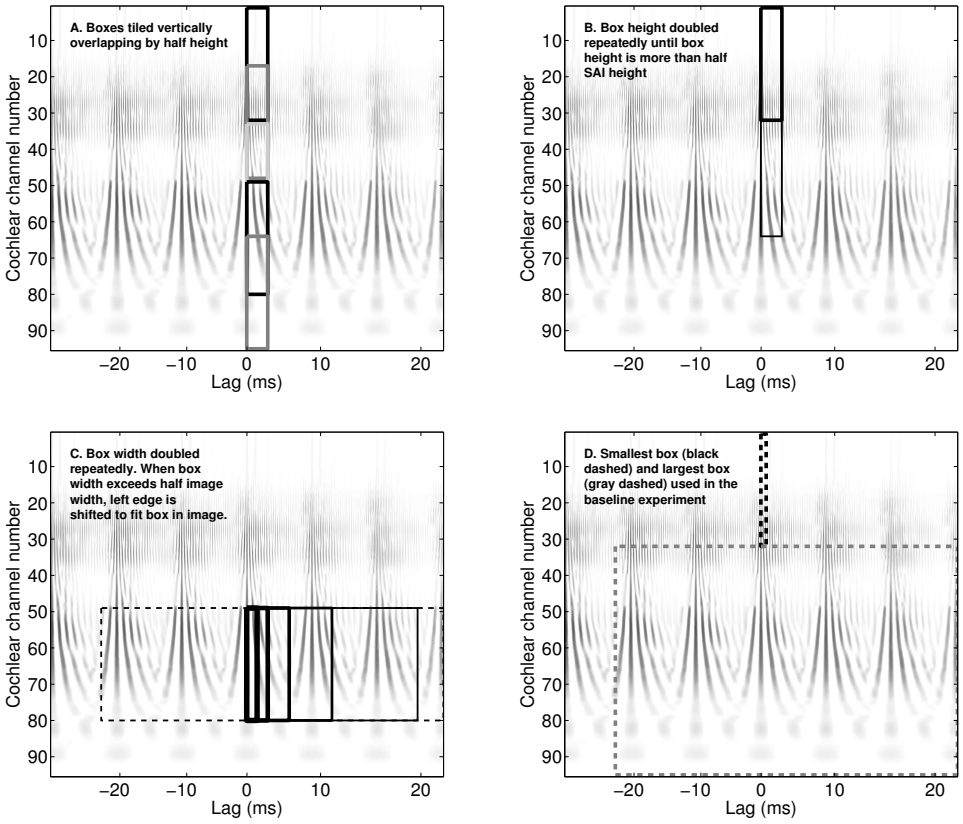


Figure 26.2: Defining a set of local rectangle regions in the SAI: rectangles are chosen to have different sizes, to capture multiscale patterns. In the default set of parameters we used, the smallest rectangle is 16 samples in the lag dimension and 32 channels high, and the largest is 1024 samples by 64 channels.

size that fits in an SAI frame.

In one group of experiments we varied the details of the box-cutting step, using processes we called "up" and "down," depending on whether we start with small rectangles and do some numbers of doublings (up), or start with full-size boxes and do some number of halvings (down). In our baseline we use rectangles of size 16×32 and larger, each dimension being multiplied by powers of two, "up" to the largest size that fits in an SAI frame. We varied the base size of the rectangle, starting from the sizes 8×16 and 32×64. We also restricted the number of sizes, by limiting the numbers of doublings of each dimension. This restriction serves to exclude the global features that are taken across a large part of the auditory image frame. In the "down" separate series of experiments, we instead started from a rectangle size equal to the dimensions of the SAI frame, working downwards by repeatedly cutting the horizontal and vertical dimensions in half. This set excludes features that are very local in the auditory image, depending on when we stop. The complete set of experimental parameters are shown in Table 26.1.

While the codebook sizes remained fixed at 256, the total number of feature dimensions varied, proportional to the number of boxes used.

Given the set of rectangles, we create dense features from each rectangle. The image inside the rectangle is downsampled (by region averaging) to the size of the smallest box (16×32 with the default parameters). The effect of this rescaling is that large rectangles are viewed at a coarser resolution. To further reduce the dimensionality of the data we compute the horizontal and vertical marginals (the average values for each col-

umn and row in the rectangle), and concatenate the two vectors into a single real-valued vector per rectangle. In the default case, this approach leaves a 16 + 32 = 48 element dense feature vector for each rectangle.

This multiscale feature-extraction approach is a way to reduce the very-high-dimensional SAI space to a set of lower-dimensionality local features at different scales, as a step toward making sparse features. Different box sizes and shapes capture both the large-scale image structure, corresponding to pitch and temporal coherence, and the microstructure corresponding to the resonances following each pulse. Wide boxes capture long-term temporal patterns; a smaller height on these restricts the temporal pattern features to a localized frequency region and captures local spectral shape. Tall boxes capture overall spectral shape; smaller widths on these include different scales of temporal pattern with the spectral pattern. Intermediate sizes and shapes capture a variety of localized features, such that even when multiple sounds are present, some of the features corresponding to regions of the SAI dominated by one sound or the other will often still show a recognizable pattern. The use of the marginals of each box reduces the dimensionality into the following sparse-code extraction step, while preserving much of the important information about spectral and temporal structure; even with this reduction, the dimensionality into the sparse code extractors is fairly high, for example 48 dimensions with the default parameters.

26.2 Ranking Sounds Given Text Queries

Practical uses of such a system include searching for sound files or specific moments in the sound track of a movie. For instance, a user may be interested in finding vocalizations of monkeys to be included in a presentation about the rainforest, or in locating the specific scene in a video where a breaking glass can be heard. A similar task is "musical query-by-description," in which a relation is learned between audio documents and words (Whitman and Rifkin, 2002).

We solve the ranking task in two steps. In the first step, sound documents are represented as sparse vectors, following the procedure described above. In the second step, we train a machine learning system to rank the documents using the extracted features.

In a previous study (Chechik et al., 2008), we evaluated different machine learning methods for the second step, while the first step was achieved using standard MFCC features. The methods that we evaluated were Gaussian mixture models (GMM), support vector machines (SVM), and the passive–aggressive model for image retrieval (PAMIR). While all three models achieved similar precision at the ranking task, PAMIR was significantly faster, and the only one that scaled efficiently to large data sets. It is therefore suitable for handling large collections of sounds, such as indexing a large fraction of the sound documents on the World Wide Web. For this reason, in this study we use the PAMIR method as a learning algorithm. The remainder of this section describes the PAMIR learning algorithm (Grangier and Bengio, 2008), recast from the image application to the audio application.

26.2.1 PAMIR for Audio Documents

Consider a text query represented by a sparse vector $q \in \mathbb{R}^{d_q}$ where d_q is the number of possible words that can be used in queries (the query vocabulary size). Also consider a set of audio documents $A \subset \mathbb{R}^{d_a}$, where each audio document is represented as a feature vector, $a \in \mathbb{R}^{d_a}$, and d_a is the dimensionality of the audio feature vector. Let $R(q) \subset A$ be the set of audio documents in A that are relevant to the query q. A ranking system provides a scoring function $S(q, a)$ that allows ranking of all documents $a \in A$ for any given query q. An ideal scoring function would rank all the documents $a \in A$ that are relevant to q ahead of the irrelevant ones:

$$S(q, a^+) > S(q, a^-) \quad \forall\, a^+ \in R(q),\ a^- \in \overline{R}(q)$$

where $\overline{R}(q)$ is the set of audio documents that are not relevant to q.

PAMIR uses a bilinear parametric score:

$$S_{\mathbf{W}}(q, a) = q^T \mathbf{W} a$$

where $\mathbf{W} \in \mathbb{R}^{d_q \times d_a}$ is the trained weight matrix. The matrix $\mathbf{W}$ can be viewed as a linear mapping from audio features to query words. The product $\mathbf{W}a$ can be viewed as a "bag of words" description of the audio document, and the dot product of this bag of words with the query words q gives the score.

To learn the matrix $\mathbf{W}$, we use an algorithm based on the passive–aggressive (PA) family of learning algorithms introduced by Crammer et al. (2006), and inspired by the ranking-SVM training algorithm of Joachims (2002). Here we consider a variant that uses triplets (q_i, a_i^+, a_i^-), each consisting of a text query and two audio documents: one that is relevant to the query, $a_i^+ \in R(q_i)$, and one that is not, $a_i^- \in \bar{R}(q_i)$.

The learning goal for each such triplet is to tune the parameters $\mathbf{W}$ of the scoring function such that the relevant document achieves a score that is larger than the irrelevant one, with a safety margin:

$$S_{\mathbf{W}}(q_i, a_i^+) > S_{\mathbf{W}}(q_i, a_i^-) + 1 \quad \forall \left(q_i, a_i^+, a_i^-\right)$$

To achieve this goal in a "soft margin" sense (Cortes and Vapnik, 1995), we define the hinge loss function summed over all training triplets:

$$L_{\mathbf{W}} = \sum_{(q_i, a_i^+, a_i^-)} l_{\mathbf{W}}(q_i, a_i^+, a_i^-)$$
$$l_{\mathbf{W}}(q_i, a_i^+, a_i^-) = \max\left(0, 1 + S_{\mathbf{W}}(q_i, a_i^-) - S_{\mathbf{W}}(q_i, a_i^+)\right)$$

The sum in $L_{\mathbf{W}}$ is over a set that is typically too large to be enumerated, but we can use an online algorithm, generating training triplets at random, that nevertheless converges to its minimum. We first initialize $\mathbf{W}$ to 0, then follow a sequence of optimization iterations. At each training iteration i, we randomly select a triplet (q_i, a_i^+, a_i^-), and solve the following convex optimization problem:

$$\mathbf{W}_i = \operatorname*{argmin}_{\mathbf{W}} \left\{ \frac{1}{2} \|\mathbf{W} - \mathbf{W}_{i-1}\|^2 + C l_{\mathbf{W}}(q_i, a_i^+, a_i^-) \right\}$$

where $\|\cdot\|^2$, the squared Frobenius norm, is the sum of squares of the matrix coefficients. At each iteration i, optimizing $\mathbf{W}_i$ achieves a trade-off between remaining close to the previous parameters $\mathbf{W}_{i-1}$ and minimizing the loss on the current triplet $l_{\mathbf{W}}(q_i, a_i^+, a_i^-)$. The *aggressiveness* parameter C controls this trade-off.

The problem in this equation can be solved analytically and yields a very simple and efficient parameter update rule (Grangier and Bengio, 2008). To simplify the derivation, think of the query vector q_i as a *set*—mostly zeros, with ones to indicate the query terms, and the audio document feature vector as a *bag*—mostly zeros, with integers to indicate how many times each feature occurred. The update rule is the same for arbitrary vectors, such as unit-norm audio feature vectors, which are a good way to normalize bags over different document lengths.

The training algorithm spends most of its time doing score evaluations. The matrix multiplications, $q^T \mathbf{W} a$, are fairly simple and fast. Since q is a sparse column vector, $q^T \mathbf{W}$ pulls out the rows of $\mathbf{W}$ associated with query terms. The dot products of those rows with a^T are then added up. When a is sparse, these dot products are fast. At each training iteration i, we start by calculating the loss on a training triplet, based on this fast

score computation using the previous iteration's weight matrix:

$$\begin{aligned} l_i = l_{\mathbf{W}}(q_i, a_i^+, a_i^-) &= \max\left(0, 1 - q_i^T \mathbf{W}_{i-1} a_i^+ + q_i^T \mathbf{W}_{i-1} a_i^-\right) \\ &= \max\left(0, 1 - (q_i^T \mathbf{W}_{i-1})(a_i^+ - a_i^-)\right) \end{aligned}$$

This loss calculation involves dot products of query-term rows with two audio feature vectors, or with their difference. When the matrix $\mathbf{W}$ is partially trained, most triplets will have zero loss, and nothing more will be done. So even though the score evaluation is very fast, it is where most of the work goes, looking for triplets with positive loss that can be used to aggressively drive the learning, and being passive otherwise.

If the loss is positive, we update the relevant weight matrix rows in the direction of the difference vector $(a_i^+ - a_i^-)^T$, because that's the direction that gives the most loss reduction per matrix change:

$$\mathbf{V}_i = q_i (a_i^+ - a_i^-)^T$$

$$\mathbf{W}_i = \mathbf{W}_{i-1} + \tau_i \mathbf{V}_i$$

where τ_i is the update rate for this iteration, and $q_i(a_i^+ - a_i^-)^T$ is the outer product that puts the difference row into the $\mathbf{V}_i$ rows corresponding to query terms. For example, the row for the query term *dog* is updated to better match the audio features of a dog-relevant sound, and to be less like the audio features of an irrelevant sound that scored too high.

The update rate for step i is determined by converting the matrix argmin to an update-rate argmin, using the relation $\|\mathbf{W} - \mathbf{W}_{i-1}\|^2 = \tau^2 \|\mathbf{V}_i\|^2$ and the definition of the hinge loss:

$$\begin{aligned} \tau_i &= \operatorname*{argmin}_{\tau} \left\{ \frac{1}{2}\tau^2 \|\mathbf{V}_i\|^2 + C \max\left(0, l_i - \tau \|\mathbf{V}_i\|^2\right) \right\} \\ &= \min\left\{ C, \frac{l_i}{\|\mathbf{V}_i\|^2} \right\} \end{aligned}$$

This rate is either just enough change to drive the loss to zero, or part way there as limited by the aggressiveness parameter C.

There is no work to do for rows not corresponding to query terms, and no work to do for columns where the audio bag of features has no count, so the updates proceed very quickly, even when $\mathbf{W}$ has many millions of elements. The update is about as expensive as the loss evaluation, but is done much less often, once the model has partially converged. At the start, however, when the matrix starts near all zeros, every training triple gives a loss near +1, so there's more work to be done in the initial learning phase. Assuming normalized audio feature vectors, each update will remove on the order of C of the loss for each term present in a query (because $\|\mathbf{V}_i\|^2$ will typically be on the order of 1 if $\|a_i^+\|^2 = \|a_i^-\|^2 = 1$ and there are few query terms). Thus if $C = 0.1$, it takes on the order of 10 iterations per query term to get to where a significant number of updates are skipped due to zero loss. After that fairly quick start, the hinge loss makes the training more discriminative, working mostly on confusable pairs.

As in an SVM (Cortes and Vapnik, 1995), the weight rows are incrementally constructed as linear combinations of training feature vectors, in this case by accumulating changes proportional to $a_i^+ - a_i^-$. The vectors used are the ones in triples that have positive loss (or did have, at some point in the training process). If weight decay is used (multiplying $\mathbf{W}$ by $1 - \epsilon$ every so often, for some small ϵ), the earliest vectors used will be washed out, and the weight rows will converge to be linear combinations of *support vectors*, the training vectors that sometimes contribute positive loss. Without weight decay, the optimization problem as we defined it does not penalize for large weights; the learning process is not as well regularized as it could be, and the resulting model may not generalize well.

The hyperparameter C can be set using cross validation. For a stopping criterion, it is a common practice to continuously trace the performance of the trained model on an independent validation set, and stop training when this performance no longer improves. Early stopping is one form of regularization, and helps to prevent overfitting the model matrix $\mathbf{W}$ to the training set. Other regularization strategies can also be added. For example, an L_1 weight shrinkage scheme can be used, making $\mathbf{W}$ sparse as a side effect, which can sometimes be an advantage (Duchi et al., 2008; Shen and Dietterich, 2009). Brute-force L_1 weight shrinkage can be expensive if done at every training step, but there are effective methods to speed it up, either exactly or approximately (Koh et al., 2007; van den Berg and Friedlander, 2008; Duchi et al., 2008).

Sampling a triplet can be done efficiently. First make lists of audio documents that are relevant for each text query. Given a text query, sample uniformly among all the relevant documents. To sample an irrelevant audio document a^-, repeatedly sample an audio document from the set of all audio documents until finding one that is not relevant to the given query. Since the dataset has significantly more irrelevant documents than relevant documents for any query, an irrelevant audio document is usually found on the first try.

26.3 Experiments

We evaluated the representations in the ranking task using a large set of audio recordings that cover a wide variety of sounds. We compared the sparse-coded SAI with sparse-coded MFCC features. In this section we describe the dataset and the experimental setup.

26.3.1 The Dataset

We collected a dataset of 8638 sound effects from multiple sources (3855 from commercially available sound effect collections, and 4783 from a variety of web sites). Most of the sounds contain only a single "auditory object," representing a "prototypical" sample of a sound category. Most are a few seconds long but there are a few that extend to several minutes.

We manually labeled all of the sound effects by listening to them and typing in a handful of tags for each sound. This was used for adding tags to existing tags (for example, from *www.findsounds.com*) and to tag the nonlabeled files from other sources. When labeling, the original file name was displayed, so the labeling decision was influenced by the description given by the original author of the sound effect. We restricted our tags to a somewhat limited set of terms. We also added high level tags to each file. For instance, files with tags such as "rain," "thunder," and "wind" were also given the tags "ambient" and "nature." Files tagged "cat," "dog," and "monkey" were augmented with tags of "mammal" and "animal." These higher level terms assist in retrieval by inducing structure over the label space. All terms are stemmed, using the Porter stemmer for English. After stemming, we are left with a vocabulary of 3268 unique tags. The sound documents have an average of 3.2 tags each.

26.3.2 The Experimental Setup

We used three-way cross validation to estimate performance of the learned ranker. Specifically, we split the set of audio documents in three equal parts, using two-thirds for training and the remaining third for testing. Training and testing was repeated for all three splits of the data, such that we obtained an estimate of the performance on all the documents. We removed queries that had fewer than five documents in either the training set or the test set, and removed the corresponding documents if these contained no other tag.

We used a second level of cross validation to determine the values of the hyperparameters: the aggressiveness parameter C, and the number of training iterations. In general, performance was good as long as C was not too high, and lower C values required longer training. We selected a value of $C = 0.1$, which was also

found to work well in other applications (Grangier and Bengio, 2008), and 10M iterations. The system is not very sensitive to the values of these parameters.

To quantify the quality of the ranking obtained by the learned model, we used the precision within the top k audio documents from the test set as ranked for each query.

26.3.3 Sparse Coding Parameters

The transformation of SAI frames into sparse codes has several parameters that can be varied. We defined a default parameter set and then performed experiments in which one or a few parameters were varied from this default set.

The default parameters cut the SAI into rectangles starting with the smallest size of 16 lags by 32 channels, leading to a total of 49 rectangles. All the rectangles were reduced to 48 marginal values each, and for each box a codebook of size 256, for a total of $49 \times 256 = 12\,544$ feature dimensions, as described in Section 26.1.3.

Table 26.1: Parameters used for the SAI experiments

Parameter set	Smallest box	Total boxes	Means per box	VQ MP	Box cutting
Default "baseline"	32×16	49	256	VQ	Up
Codebook sizes	32×16	49	4, 16, 64, 256, 512, 1024, 2048, 3000, 4000 6000 8000	VQ	Up
Matching pursuit	32×16	49	4, 16, 64, 256, 1024, 2048, 3000	MP	Up
Box sizes (down)	16×8 32×16 64×32	1, 8, 33, 44, 66 8, 12, 20, 24 1, 2, 3, 4, 5, 6	256	VQ	Down
Box sizes (up)	16×8 32×16 64×32	32, 54, 72, 90, 108 5, 14, 28, 35, 42 2, 4, 6, 10, 12	256	VQ	Up

Using this default experiment as a baseline for comparisons, we systematically varied several parameters and studied their effect on the retrieval precision. We modified the rectangle sets used for sparse segmentation, limiting the number of rectangles used, starting with small and working toward large ("up"), and vice versa ("down"). Further variants used different codebook sizes in the sparse coding (using both VQ and MP).

26.3.4 Comparisons with MFCC

We used mel-frequency cepstral coefficients (MFCC), augmented by their first and second time-differences as additional features of each frame ("delta" and "delta-delta" MFCCs, as commonly used in speech recognition), and turned these dense features into a sparse code by using VQ or MP on the full MFCC vector. We set the MFCC parameters based on a configuration that was optimized for speech, and further systematically varied three parameters of the MFCCs: the number of cepstral coefficients (traditionally 13 for speech), the length of each frame's spectral-analysis Hamming window (traditionally 25 ms), and the number of codebooks used to sparsify the MFCC of each frame. Optimal performance was obtained with a codebook of size 5000, 40ms frames and 40 cepstral coefficients (see Section 26.4). This configuration corresponds to much higher frequency resolution than the typical MFCC features used for speech.

26.4 Results

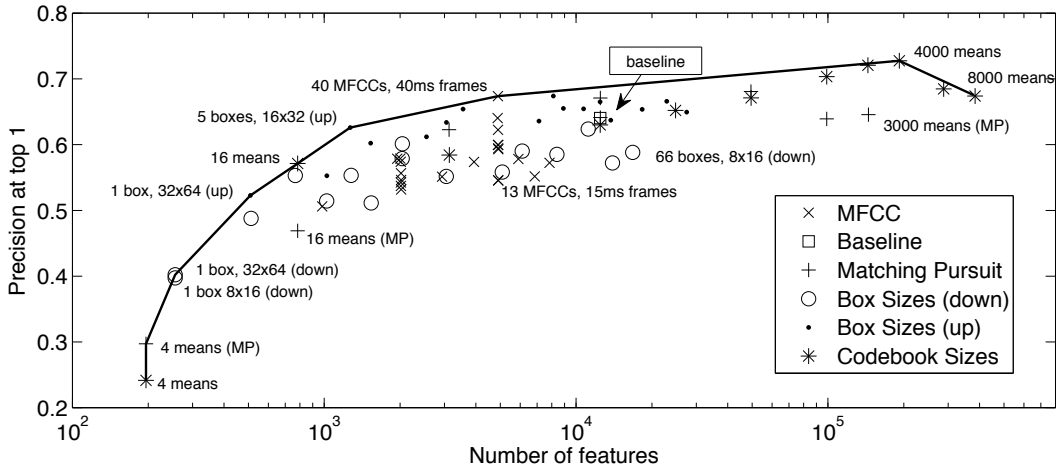


Figure 26.3: Ranking precision at the top-1 sound plotted against feature count, for all experiments. Selected experiment names are plotted on the figure near each point. The different experiment sets are denoted by different markers. The convex hull joining the best-performing points is plotted as a solid line.

Besides the damping parameters mentioned above, we tested the effect of various parameters of the SAI feature extraction procedure on the test-set precision. Figure 26.3 plots the precision of the top ranked sound file against the length of the sparse feature vector, for all our experiments. Each set of experiments has their own marker. For instance, the series of stars show precision for a set of experiments where the number of means (size of the codebook) is varied. The rest of the parameters do not change from one star to the other, and were set at the default parameters defined in Section 26.1.3.

top-k	SAI	MFCC	Percent error reduction
1	27	33	18%
2	39	44	12%
5	60	62	4%
10	72	74	3%
20	81	84	4%

Table 26.2: Comparison of error at top-k for best SAI and MFCC configurations (error defined as one minus precision).

Interestingly, performance saturates with a very large number of features—about 10^5, resulting from using 4000 codewords per codebook, and a total of 49 codebooks—and declines for even larger codebooks. The best parameter configuration achieved 73% precision at the top ranked sound file, which was significantly better than the best MFCC result which achieved 67%. This reflects about 18% smaller false retrieval (from 33% to 27% error). SAI features also achieve better precision-at-top-k consistently for all values of k, although with lower relative precision improvement (Table 26.2). It should be stressed however that the parameters that we found (and the auditory model architecture in general) are not guaranteed to be "optimal," and it is possible that further refinement could further improve the retrieval precision.

Table 26.3 shows three sample queries together with the top 5 test sound documents returned by the best SAI-based and MFCC-based trained systems. The three queries shown are a very small sample of the data,

but show that both systems behave reasonably well, most often returning good documents, or at least documents that appear not-too-far from the expected answer. For instance, the SAI-based system returns document "water-dripping" for the query "gulp," which, while being wrong, is not far from the mark. Similarly, the document "45-Crowd-Applause" is returned by the MFCC-based system for query "applause-audience," despite not being labeled as relevant for that query.

Query	SAI file (labels)	MFCC file (labels)
tarzan	Tarzan-2 (tarzan, yell) tarzan2 (tarzan, yell) 203 (tarzan) wolf (mammal, wolves, wolf) morse (mors, code)	TARZAN (tarzan, yell) 175orgs (steam, whistle) mosquito-2 (mosquito) evil-witch-laugh (witch, laugh, evil) Man-Screams (horror, scream, man)
applause audience	27-Applause-from-audience 30-Applause-from-audience golf50 (golf) firecracker 53-ApplauseLargeAudienceSFX	26-Applause-from-audience phaser1 (trek, phaser, star) fanfare2 (fanfar, trumpet) 45-Crowd-Applause (crowd, applause) golf50
gulp	tite-flamn (hit, drum, roll) water-dripping (water, drip) Monster-growling (horror, monster, growl) Pouring (pour, soda)	GULPS (gulp, drink) drink (gulp, drink) california-myotis-search (blip) jaguar-1 (bigcat, jaguar, mammal)

Table 26.3: Top documents obtained for queries that performed very differently between the SAI and MFCC feature based systems.

The performance that we calculated was based on textual tags, which are often noisy and incomplete. In particular, people may use different terms to describe very similar concepts. Also, the same sound may be described across different aspects. For instance a music piece may be described by the playing instrument ("piano") or the mood it conveys ("soothing") or the name of the piece. This multilabel problem is common in content based retrieval, being shared by image search engines, for example. It typically leads to a pessimistic estimate of precision, or relevance of the retrieved items.

Table 26.4 shows queries that consistently "confused" our system and caused retrieval of sounds with a different label. For each pair of queries q_1 and q_2 we measure confusion, by counting the number of sound files that were ranked within the top-k files for query q_1, but not for q_2 even though q_2 was identical to their labels. For example, there were 7 sound files that were labeled *evil laugh* but were not ranked within the top k documents for the query *evil laugh*, and at the same time ranked highly for *laugh*.

As can be seen from the table, the repeated "mistakes" are often due to labeling inconsistencies: when a sound labeled *laugh* is retrieved for a query *evil laugh*, the system counts it as a mistake, even though this is likely to be a relevant match. In general we find that confused queries are often semantically similar to the sound label, hence the errors made by the ranking systems actually reflect the fact that the sound files have partial or inconsistent labeling. This demonstrates a strength of content-based sound ranking: it can identify relevant sounds even if their textual labels are incomplete, wrong or maybe even maliciously spammed.

Performance within each series was found to be roughly monotonic with the total number of feature dimensions, but with the "up" series doing a little better than the "down" series. While this could be taken as evidence that the finest scales are most useful, we later discovered a better explanation: rectangles bigger than the minimum-size ones were subject to a code bug, such that they would not have had the correct SAI

Query, label		SAI + MFCC errors
clock-tick	cuckoo	8
door knock	door	8
evil laugh	laugh	7
laugh witch	laugh	7
bell-bicycle	bell	7
bee-insect	insect	7

Table 26.4: Error analysis. Queries that were repeatedly confused for another query. All pairs of true-label and confused labels with total count above seven are listed.

resampling—they were "collapsed" into lag 0 such that the time-lag marginal was lost, and the rectangles only retained the local spectral shape information. Therefore none of the experiments took advantage of the temporal structure at lags longer than 16 samples, and many used no time structure at all. Many of the experiments would likely have done better if the larger rectangles were correctly reduced.

26.5 Conclusions and Followup

In this study, auditory features are shown to work well with sparse coding in a task involving a wide range of sounds and labels, as opposed to speech or music. PAMIR is an effective machine learning approach to study a range of variations of sparsified sound feature representations.

The findings support the hypothesis that a machine hearing front end based on emulating the function of the cochlea and converting the fine time structure at its output to a stabilized auditory image is effective at representing general sounds. We convert the auditory model outputs to a very-high-dimensional sparse code, and let the learning algorithm identify the features that are most discriminative.

The auditory model representation we use does not take into account long-term temporal relationships, like the "bag of words" approach common in text document retrieval. The SAI features capture a short window of time, up to about 50 ms for the biggest box features (but less than 1 ms for the rectangles properly used), which is comparable to but a bit less than that captured by MFCC with double deltas. Longer correlations are likely to be also useful for sound retrieval, and it remains an interesting research direction to study how they should be estimated and used.

Since our system currently uses only features from short windows, we envision future work to incorporate more dynamics of the sound over longer times, either as a *bag-of-patterns* using patterns that represent more temporal context, or through other methods.

In a follow-on to the main study, we showed that the gap between MFCC and auditory features is even larger when the sounds being ranked are in the context of other interfering sounds (Lyon et al., 2011).

Chapter 27

Musical Melody Matching

> I hope my critics will excuse me if I conclude from the opposite nature of their objections that I have struck out nearly the right path. As to my Theory of Consonance, I must claim it to be a mere systematisation of *observed facts* (with the exception of the functions of the *cochlea* of the ear, which is moreover an hypothesis that may be entirely dispensed with). But I consider it a mistake to make the Theory of Consonance the essential foundation of the Theory of Music, and, I had thought that this opinion was clearly enough expressed in my book. The essential basis of Music is *Melody*.
>
> — *On the Sensations of Tone*, Hermann Ludwig F. Helmholtz (1870)

This chapter draws on material from the 2012 paper "The Intervalgram: An Audio Feature for Large-scale Melody Recognition" by Thomas C. Walters, David A. Ross, and Richard F. Lyon (Walters et al., 2013).

In this chapter, we review a system for representing the melodic content of short pieces of audio using a novel chroma-based representation known as the "intervalgram," which is a summary of the local pattern of musical intervals in a segment of music. We introduced chroma as pitch within an octave in Section 4.7. The intervalgram is based on a chroma representation derived from the pitchogram, or temporal profile of the stabilized auditory image. Each intervalgram frame is made locally key invariant by means of a "soft" pitch transposition to a local reference. Intervalgrams are generated for a piece of music using multiple overlapping windows. These sets of intervalgrams are used as the basis of a system for detection of identical melodies across a database of music. Using a dynamic-programming-like approach for comparisons between a reference and the song database, performance was evaluated on the dataset. A first test of an intervalgram-based system on this dataset yields a precision at top-1 of 53.8%, with a precision–recall curve that shows very high precision up to moderate recall, suggesting that the intervalgram is adept at identifying the easier-to-match cover songs in the dataset with high robustness. The intervalgram is designed to support locality-sensitive hashing, such that an index lookup from each single intervalgram feature has a moderate probability of retrieving a match, with relatively few false matches. With this indexing approach, a large reference database can be quickly pruned before more detailed matching, as in previous content-identification systems.

We are interested in solving the problem of cover song detection at very large scale. In particular, given a piece of audio, we wish to identify another piece of audio representing the same melody, from a potentially very large reference set. Though our approach aims at the large-scale problem, the representation developed is compared in this paper on a small-scale problem for which other results are available.

There can be many differences between performances with identical melodies. The performer may sing or play the melody at a different speed, in a different key or on a different instrument. However, these changes in performance do not, in general, prevent a human from identifying the same melody, or pattern of notes.

Thus, given a performance of a piece of music, we wish to find a representation that is to the largest extent possible invariant to such changes in instrumentation, key, and tempo.

Serrà Julià (2011) gives a thorough overview of the existing work in the field of melody identification, and breaks down the problem of creating a system for identifying versions of a musical composition into a number of discrete steps. To go from audio signals for pieces of music to a similarity measure, the proposed process is:

- Feature extraction
- Key invariance (invariance to transposition)
- Tempo invariance (invariance to a faster or slower performance)
- Structure invariance (invariance to changes in long-term structure of a piece of music)
- Similarity computation.

In this study, we concentrate on the first three of these steps: the extraction of an audio feature for a signal, the problem of invariance to pitch shift of the melody (both locally and globally) and the problem of invariance to changes in tempo between performances of a piece of music.

For the first stage, we present a system for generating a pitch representation from an audio signal, using the stabilized auditory image (SAI) as an alternative to more typical spectrogram-based approaches. Key invariance is achieved locally (per feature), rather than globally (per song). Individual intervalgrams are key normalized relative to a reference chroma vector, but no guarantees are made that the reference chroma vector will be identical across consecutive features. This local pitch invariance allows for a feature that can track poor-quality performances in which, for example, a singer changes key gradually over the course of a song. It also allows the feature to be calculated in a streaming fashion, without having to wait to process all the audio for a song before making a decision on transposition. Other approaches to this problem have included shift-invariant transforms (Marolt, 2008), the use of all possible transpositions (Ellis and Cotton, 2007) or finding the best transposition as a function of time in a symbolic system (Tsai et al., 2008). Finally, tempo invariance is achieved by the use of variable-length time bins to summarize both local and longer-term structure. This approach is in contrast to other systems (Ellis and Cotton, 2007; Marolt, 2008) which use explicit beat tracking to achieve tempo invariance.

While the features are designed for use in a large-scale retrieval system when coupled with a hashing technique (Baluja and Covell, 2008), in this study we test the baseline performance of the features by using a Euclidean distance measure. A dynamic-programming alignment is performed to find the smallest-cost path through the map of distances between a probe song and a reference song; partial costs, averaged over good paths of reasonable duration, are used to compute a similarity score for each probe–reference pair.

We evaluate performance of the intervalgam (using both SAI-based chroma and spectrogram-based chroma) using the dataset (Ellis and Cotton, 2007). This is a set of 160 songs, in 80 pairs that share an underlying composition. There is no explicit notion of a "cover" versus an "original" in this set, just an "A" version and a "B" version of a given composition, randomly selected. While it is a small corpus, several researchers have made use of this dataset for development of audio features, and report results on it. Ellis and Cotton (2007) report performance in terms of absolute classification accuracy for the LabRosa 2006 and 2007 music information retrieval evaluation exchange (MIREX) competition, and these results are extended by, amongst others, Ravuri and Ellis (2010), who present detection error tradeoff curves for a number of systems.

Since we are ultimately interested in the use of the intervalgram in a large-scale system, it is worth briefly considering the requirements of such a system. In order to perform completely automated detection of cover songs from a large reference collection, it is necessary to tune a system to have extremely low false hit rate on

each reference. For such a system, we are interested less in high absolute recall and more in finding the best possible recall given a very low threshold for false positives. Such systems have previously been reported for nearly-exact-match content identification (Baluja and Covell, 2008). The intervalgram has been developed for and tested with a similar large-scale back end based on indexing, but there is no large accessible data set on which performance can be reported. It is hard to estimate recall on such undocumented data sets, but the system identifies a large number of covers even when tuned for less than 1% false matches.

27.1 Algorithm

27.1.1 The Stabilized Auditory Image

The stabilized auditory image (SAI), described in detail in Chapter 21, is a correlogram-like representation of the output of an auditory filterbank. In this implementation, a 64-channel pole-zero filter cascade (Lyon et al., 2010b; Lyon, 2011a) is used. The output of the filterbank is half-wave rectified and a process of "strobe detection" is carried out. In this process, large peaks in the waveform in each channel are identified. Each rectified bandpass waveform is then cross-correlated with a sparsified version of itself which is zero everywhere except at the identified strobe points. This process of "strobed temporal integration" (Patterson et al., 1992; Walters, 2011) is very similar to performing autocorrelation in each channel, but is considerably cheaper to compute due to the sparsity of points in the strobe signal. The upper panels of Figure 27.1 show a waveform (upper panel) and stabilized auditory image (middle panel) for a sung note. The pitch of the voice is visible as a series of vertical ridges at lags corresponding to multiples of the repetition period of the waveform, and the formant structure is visible in the pattern of horizontal resonances following each large pulse.

27.1.2 Chroma From the Auditory Image

To generate a chroma representation from the SAI, the "temporal profile" is first computed by summing over the frequency dimension; this gives a single vector of values which correspond to the strength of temporally-repeating patterns in the waveform at different lags. The temporal profile gives a representation of the time intervals associated with strong temporal repetition rates, or possible pitches, in the incoming waveform. This SAI temporal profile closely models human pitch perception (Ives and Patterson, 2008); for example, in the case of stimuli with a missing fundamental, there may be no energy in the spectrogram at the frequency of the pitch perceived by a human, but the temporal profile will show a peak at the time interval associated with the missing fundamental.

The lower panel of Figure 27.1 shows the temporal profile of the stabilized auditory image for a sung vowel. The pitch is visible as a set of strong peaks at lags corresponding to integer multiples of the pulse rate of the waveform. Figure 27.2 shows a series of temporal profiles stacked in time, a "pitchogram," for a piece of music with a strong singing voice in the foreground. The dark areas correspond to lags associated with strong repetition rates in the signal, and the evolving melody is visible as a sequence of horizontal stripes corresponding to notes; for example in the first second of the clip there are four strong notes, followed by a break of around 1 second during which there are some weaker note onsets.

The temporal profile is then processed to map lag values to pitch chromas in a set of discrete bins, to yield a representation as chroma vectors, also known as "pitch class profiles" (PCPs) (Serrà Julià, 2011). Other systems generate chroma vectors from spectral slices, in contrast to our use of temporal-profile slices. In our standard implementation, we use 32 pitch bins per octave. Having more bins than the standard 12 semitones in the Western scale allows the final feature to accurately track the pitch in recordings where the performer is either mistuned or changes key gradually over the course of the performance; it also enables more accurate

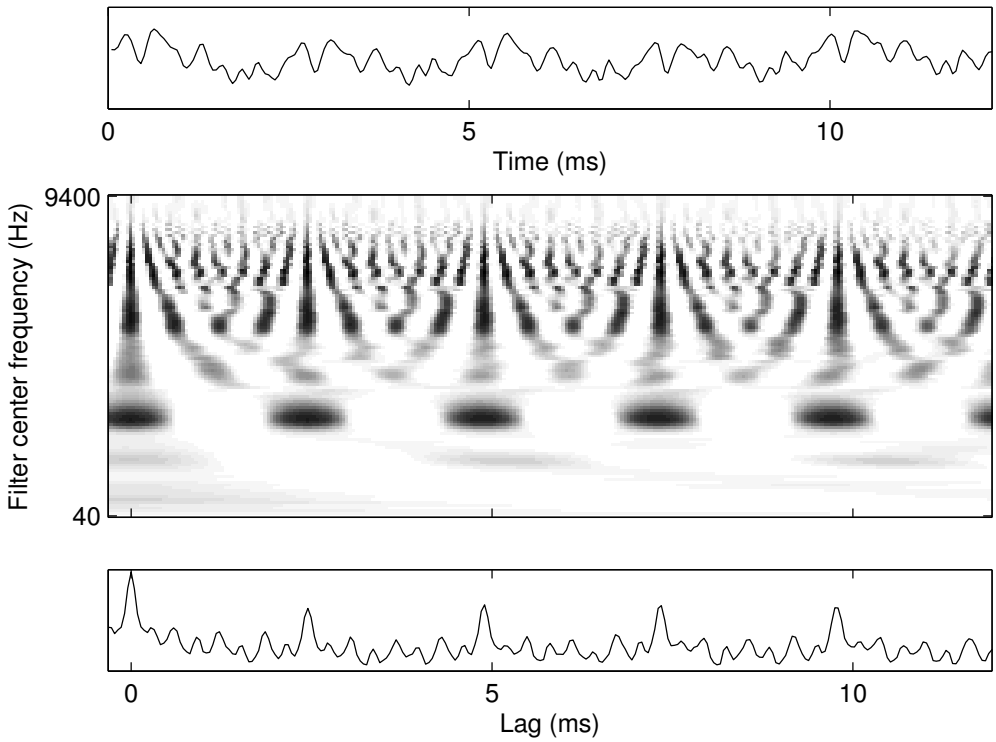


Figure 27.1: Waveform (top panel), stabilized auditory image (SAI, middle panel), and SAI temporal profile (bottom panel) for a human voice singing a note.

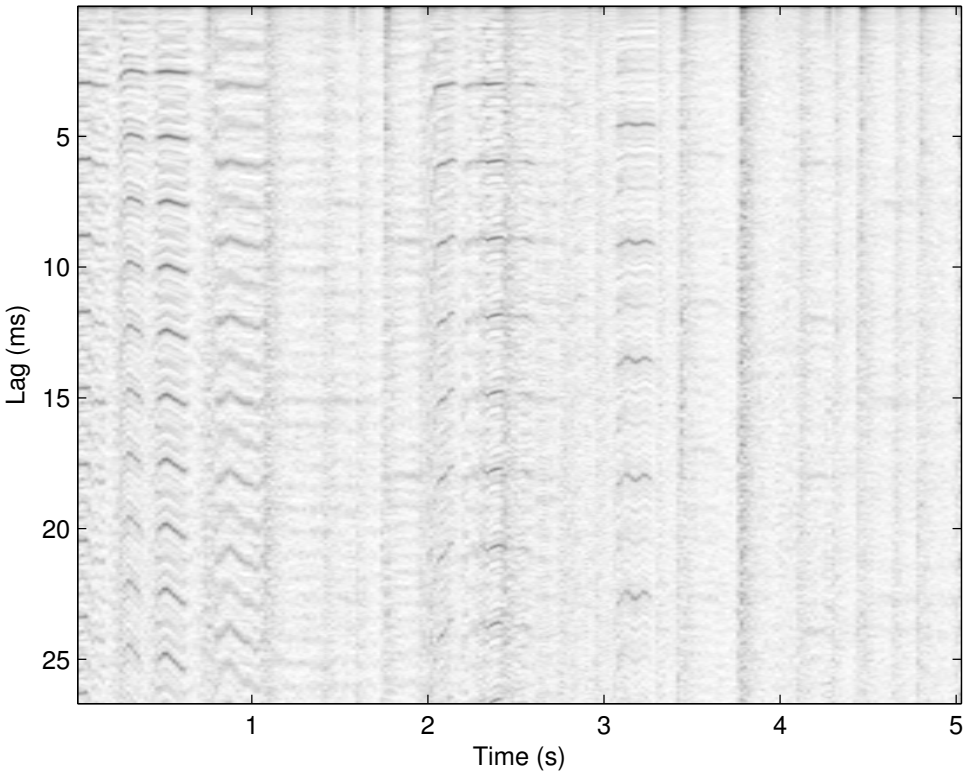


Figure 27.2: A pitchogram created by stacking a number of SAI temporal profiles in time. The lag dimension of the auditory image is now on the vertical axis. Dark ridges are associated with strong repetition rates in the signal.

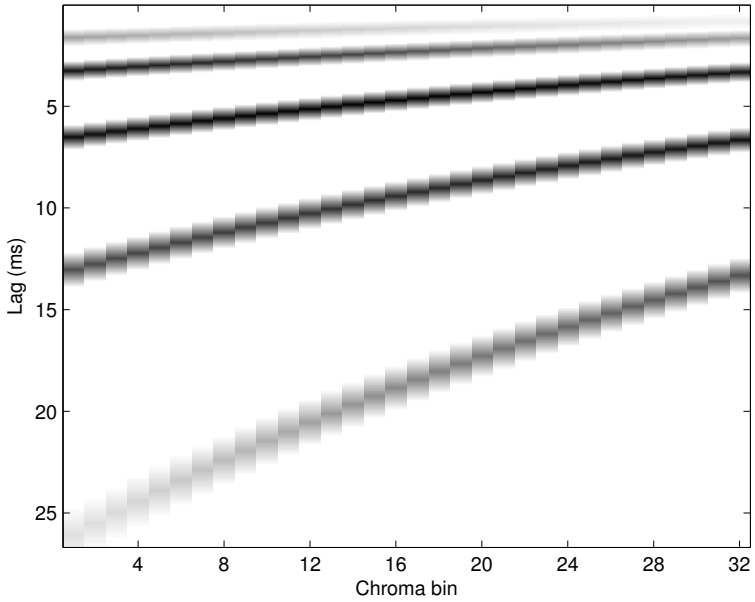


Figure 27.3: Weighting matrix to map from the time-lag axis of the SAI into 32 chroma bins

tracking of pitch sweeps, vibrato, and other nonquantized changes in pitch. Additionally, using an integer power of two for the dimensions of the final representation lends itself to easy use of a wavelet decomposition for hashing, which is discussed below. The chroma bin assignment is done using a weighting matrix, by which the temporal profile is multiplied to map individual samples from the lag dimension of the temporal profile into chroma bins. The weighting matrix is designed to map the linear time-interval axis to a wrapped logarithmic note pitch axis, and to provide a smooth transition between chroma bins. An example weighting matrix is shown in Figure 27.3. The chroma vectors for the same piece of music as in Figure 27.2 are shown in Figure 27.4.

27.1.3 Chroma from the Spectrogram

In addition to the SAI-based chroma representation described above, a more typical spectrogram-based chroma representation was tested as the basis for the intervalgram. In this case, chroma vectors were generated using the `chromagram_E` function distributed with the covers80 dataset (Ellis and Cotton, 2007), with a modified step size to generate chroma vectors at the rate of 50 per second, and 32 pitch bins per octave for compatibility with the SAI-based features above. This function uses a Gaussian weighting function to map FFT bins to chroma, and weights the entire spectrum with a Gaussian weighting function to emphasize octaves in the middle of the range of musical pitches.

27.1.4 Intervalgram Generation

A stream of chroma vectors is generated at a rate of 50 per second. From this chromagram, a stream of intervalgrams is constructed at the rate of around 4 per second. The intervalgram is a two-dimensional feature with dimensions of chroma and time offset; however, depending on the exact design the time-offset axis may be nonlinear.

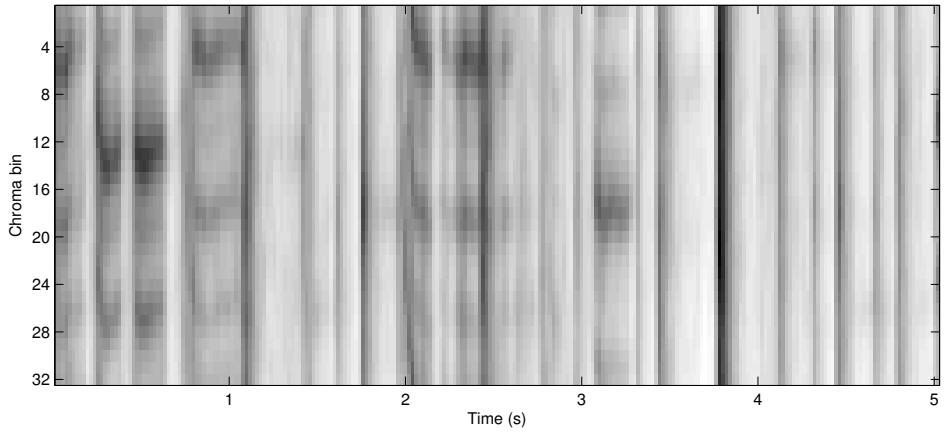


Figure 27.4: Chroma vectors generated from the pitchogram vectors shown in Figure 27.2.

For each time-offset bin in the intervalgram, groups of individual chroma vectors are averaged together to summarize the chroma in various time windows, before and after a central reference time. It takes several contiguous notes to effectively discern the structure of a melody, and for any given melody the stream of notes may be played at a range of speeds. In order to take into account both short- and longer-term structure in the melody, a variable-length time-averaging process is used to provide a fine-grained view of the local melody structure, and simultaneously give a coarser view of longer timescales. Small absolute time offsets use narrow time bin widths, while larger absolute offsets use larger bin widths; this nonlinear stretching of time helps to accommodate a moderate amount of tempo variation. Figure 27.5 shows how chroma vectors are averaged together to make the intervalgram.

In the examples below, the widths of the bins increase from the center of the intervalgram, and are proportional to the sum of a forward and reverse exponential $w_b = f\left(w_f^p + w_f^{-p}\right)$, where p is an integer between 0 and 15 (for the positive bins) and between 0 and –15 (for the negative bins), f is half the central bin width, and w_f is the width factor that determines the rate at which the bin width increases as a function of distance from the center of the intervalgram.

In the best-performing implementation, the temporal axis of the intervalgram is 32 bins wide. The central two slices along the time axis of the intervalgram are the average of 18 chroma vectors each (360 ms each), moving away from the centre of the intervalgram, the outer temporal bins summarize longer time-scales before and after the central time. The number of chroma vectors averaged in each bin increases up to 99 (1.98 s) in the outermost bins, leading to a total temporal span of 26 seconds for each intervalgram.

A "reference" chroma vector is also generated from the stream of incoming chroma vectors at the same rate as the intervalgrams. The reference chroma vector is computed by averaging together nine adjacent chroma vectors using a triangular window. The temporal center of the reference chroma vector corresponds to the temporal center of the intervalgram. In order to achieve local pitch invariance, this reference vector is then circularly cross-correlated with each of the surrounding intervalgram bins. This cross-correlation process implements a "soft" normalization of the surrounding chroma vectors to a prominent pitch or pitches in the reference chroma vector. Given a single pitch peak in the reference chroma vector, the process corresponds exactly to a simple transposition of all chroma vectors to be relative to the single pitch peak. In the case where there are multiple strong peaks in the reference chroma vector, the process corresponds to a simultaneous shifting to multiple reference pitches, followed by a weighted average based on the individual pitch strengths. This process leads to a blurry and more ambiguous interval representation but, crucially, never leads to a hard decision being made about the "correct" pitch of the melody at any point. Making only "soft" decisions at

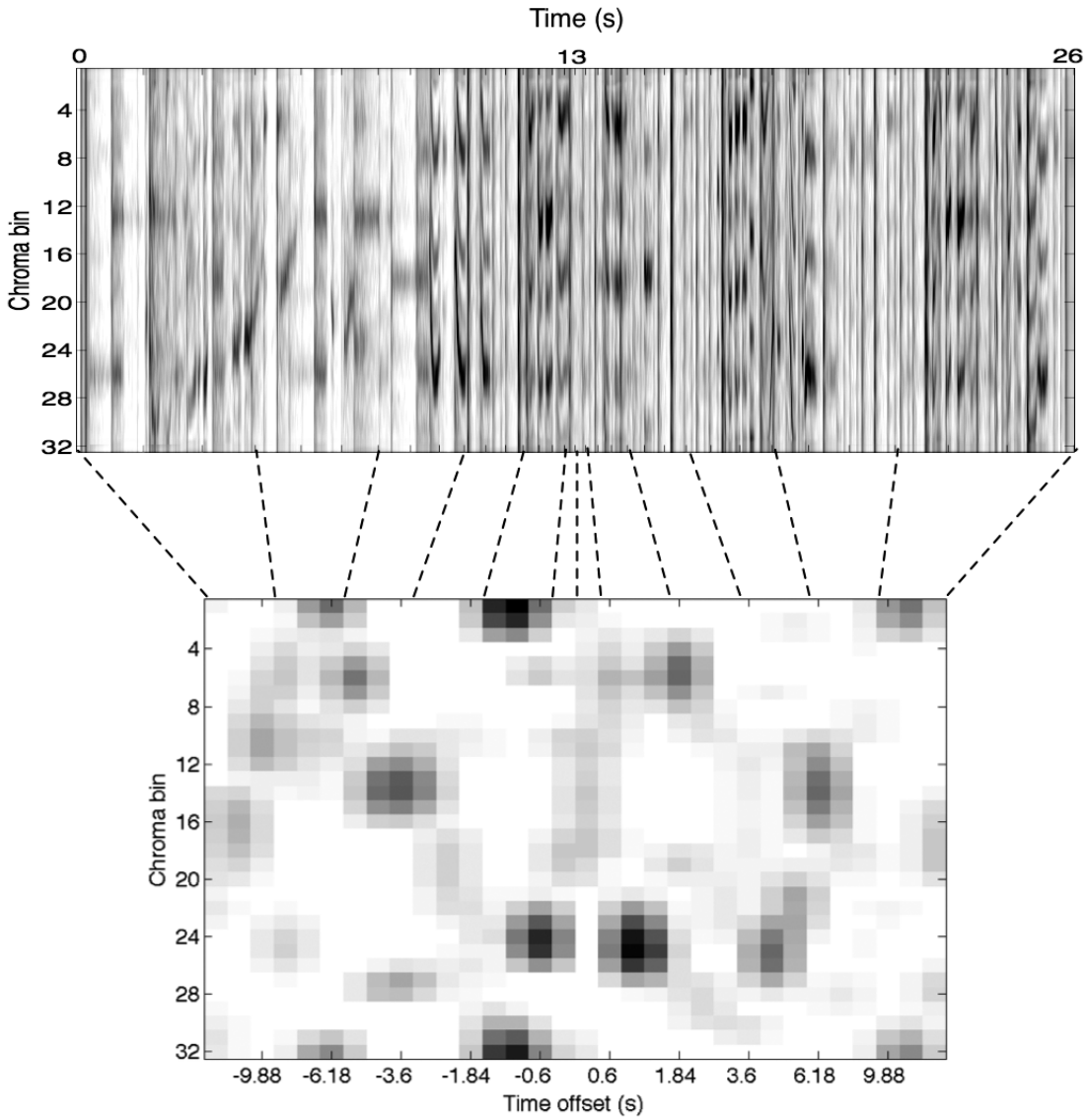


Figure 27.5: The intervalgram is generated from the chromagram using variable-width time bins and cross-correlation with a reference chroma vector to normalize chroma within the individual intervalgram.

each stage means that there is less need for either heuristics or tuning of parameters in building the system. With standard parameters the intervalgram is a 32 by 32 pixel feature vector generated at the rate of one every 240 ms and spanning a 26.4 second window. Since there are many overlapping intervalgrams generated, there are many different pitch reference slices used, some making crisp intervalgrams, and some making fuzzy intervalgrams.

27.1.5 Similarity Scoring

Dynamic programming is a standard approach for aligning two audio representations, and has been used for version identification by many authors; Serrà Julià (2011) provides a representative list of example implementations. To compare sets of features from two recordings, each feature vector from the probe recording is compared to each feature vector from the reference recording, using some distance measure, for example Euclidean distance, correlation, or Hamming distance over a locality-sensitive hash of the feature. This comparison yields a distance matrix with samples from the probe on one axis and samples from the reference on the other. We then find a minimum-cost path through this matrix using a dynamic programming algorithm that is configured to allow jumping over poorly-matching pairs. Starting at the corner corresponding to the beginning of the two recordings the path can continue by jumping forward a certain number of pixels in both the horizontal and vertical dimensions. The total cost for any particular jump is a function of the similarity of the two samples to be jumped to, the cost of the jump direction and the cost of the jump distance. If two versions are exactly time-aligned, we would expect that the minimum-cost path through the distance matrix would be a straight line along the leading diagonal. Since we expect the probe and reference to be roughly aligned, the cost of a diagonal jump is set to be smaller than the cost of an off-diagonal jump.

The minimum and maximum allowed jump lengths in samples can be selected to allow the algorithm to find similar intervalgrams that are more sparsely distributed, interleaved with poorly matching ones, and to constrain the maximum and minimum deviation from the leading diagonal. Values that work well are a minimum jump of 3 and maximum of 4, with a cost factor equal to the longer of the jump dimensions (so a move of 3 steps in the reference and 4 in the probe costs as much as 4,4 even though it uses up less reference time, while jumps of 3,3 and 4,4 along the diagonal can be freely intermixed without affecting the score as long as enough good matching pairs are found to jump between). These lengths, along with the cost penalty for an off-diagonal jump and the difference in cost for long jumps over short jumps, are parameters of the algorithm. Figure 27.6 shows a distance matrix for a probe–reference pair.

In the following section we test the performance of the raw intervalgrams, combined with the dynamic programming approach described above, in finding similarity between cover songs.

27.2 Experiments

We tested performance of the similarity-scoring system based on the intervalgram, as described above, using the standard paradigm for the covers80 dataset, which is to compute a distance matrix for all query songs against all reference songs, and report the percentage of query songs for which the correct reference song has the highest similarity score.

Intervalgrams were computed from the SAI using the parameters outlined in Table 27.1, and scoring of probe–reference pairs was performed using the dynamic programming approach described above. Figure 27.7 shows the matrix of scores for the comparison of each probe with all reference tracks. Darker pixels denote lower score, and lighter pixels denote higher scores. The white crosses show the highest-scoring reference for a given probe. 43 of the 80 probe tracks in the covers80 dataset were correctly matched to their associated reference track leading to a score of 53.8% on the dataset. For comparison, Ellis and Cotton (2007) reports a

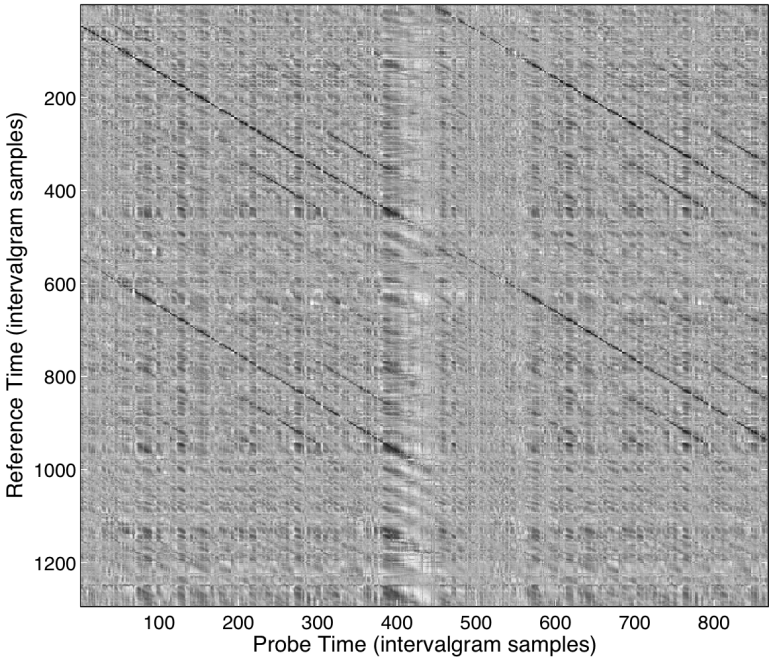


Figure 27.6: Example distance matrix for a pair of songs that share an underlying melody. The darker pixels show the regions where the intervalgrams match closely.

Parameter	Value
Chromagram step size (ms)	20
Chroma bins per octave	32
Total intervalgram width (s)	26.44
Intervalgram step size (ms)	240
Reference chroma vector width (chroma vectors)	4

Table 27.1: Parameters of the best intervalgram for cover song matching

score of 42.5% for their MIREX'06 entry, and 67.5% for their MIREX'07 entry (the latter had the advantage of using covers80 as a development set, so is less directly comparable).

In addition to the SAI-based chroma features, spectrogram-based chroma features were computed from all tracks in the dataset. These features used 32 chroma bins, and were computed at 50 frames per second, to provide a drop-in replacement for the SAI-based features. Intervalgrams were computed from these features using the parameters in Table 27.1.

In order to generate detection error tradeoff curves for the dataset, the scores matrix from Figure 27.7 was dynamically thresholded to determine the number of true and false positives for a given threshold level. The results were compared against the reference system supplied with the covers80 dataset, which is essentially the same as the system entered by LabRosa for the MIREX'06 competition, as documented by Ellis and Cotton (2007). Figure 27.8 shows precision–recall curves of their MIREX'06 entry and of our intervalgram-based system, both with SAI chroma features and spectrogram chroma features. Comparing curves from Ravuri and Ellis (2010), performance of the intervalgram-based systems is seen to consistently lie between that of the LabRosa 2006 and 2007 systems.

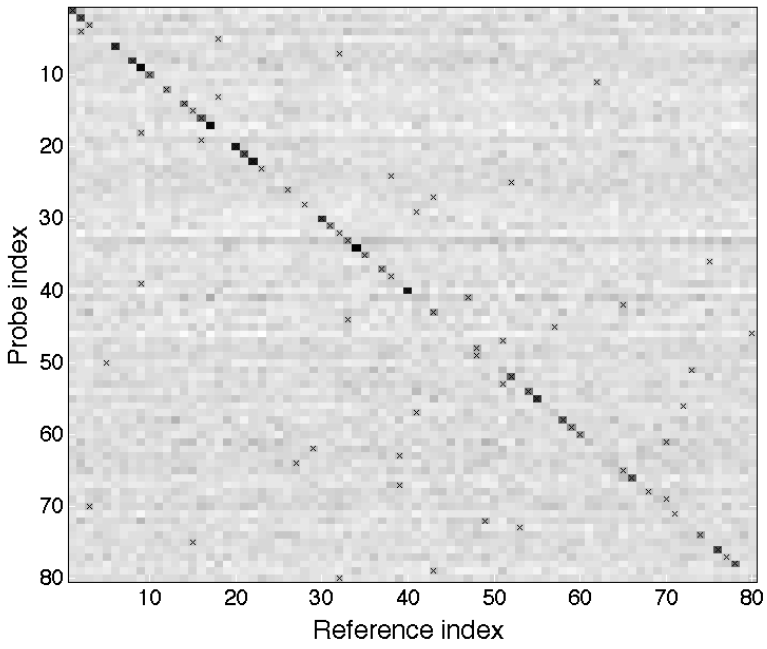


Figure 27.7: Scores matrix for comparing all probes and references in the dataset. Darker pixels denote higher scores, indicating a more likely match. Black crosses denote the best-matching reference for each probe.

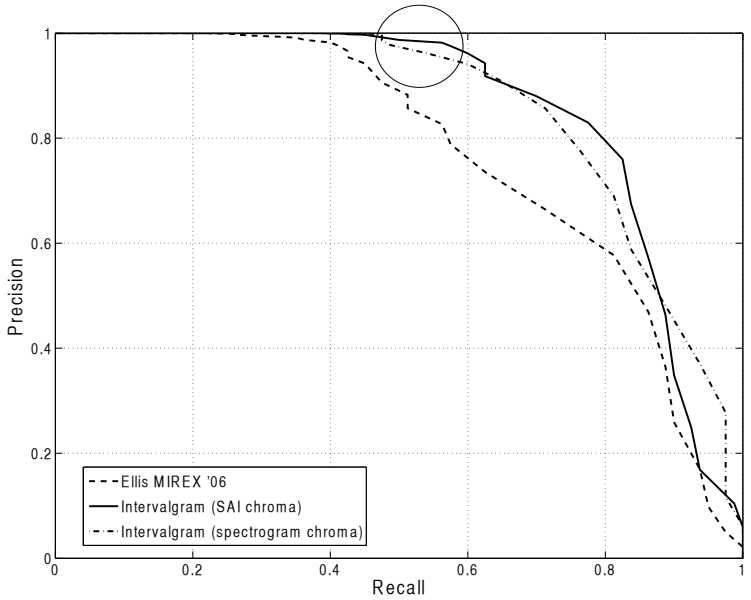


Figure 27.8: Precision–recall curves for the intervalgram-based melody-matching system described in this chapter, and the LabROSA MIREX'06 entry (Ellis and Cotton, 2007). Precision is one minus the probability of falsely matching a song as a cover, while recall is the probability of correctly identifying a cover song. In the high-precision region, near 50% recall and above 95% precision, shown circled, the SAI-based features lead to about half as many false matches as the spectrogram-based features.

Of particular interest is the performance of the features at high precision. The SAI-based intervalgram can achieve 47.5% recall at 99% precision, whereas the Ellis MIREX'06 system achieves 35% recall at 99% precision. These early results suggest that the intervalgram shows good robustness to interference. The intervalgram also stands up well to testing on larger, internal, datasets in combination with hashing techniques, as discussed below.

27.3 Discussion

We have introduced a new chroma-based feature for summarizing musical melodies, which does not require either beat tracking or exhaustive search for transposition invariance, and have demonstrated a good baseline performance on a standard dataset. However, we developed the intervalgram representation to be a suitable candidate for large-scale, highly robust cover-song detection. In the following sections we discuss some approaches to the application of the intervalgram in such a system.

27.3.1 SAI and Spectrogram-Based Chroma

There was no great difference in performance between intervalgrams generated using the temporal profile of the SAI and intervalgrams generated using a spectrogram-based chroma feature. However, there are some small differences in different regions of the precision–recall curve. Recall at high precision is very similar for both forms of chroma features; as precision is allowed to fall, the SAI-based features lead to slightly higher recall for a given precision, but the trend is reversed in the lower-precision end of the curve. This may suggest that there would be a benefit in combining both SAI-based and spectrogram-based chroma into a feature which makes use of both. There is some evidence to suggest that the temporal profile of the SAI may be robust to stimuli in which the pitch is ambiguous (Ives and Patterson, 2008), but this result may be less relevant in the context of music.

Another chromagram feature developed about the same time as ours (Müller and Ewert, 2010), based on a more narrowband spectral analysis with cepstral liftering, is said to perform very well on the melody matching task.

27.3.2 Scaling Up

In order to perform melody recognition on a large database of content, it is necessary to find a cheaper and more efficient way of matching a probe song against many references. The brute-force approach of computing a full distance map for the probe against every possible reference scales as the product of the number of probes and the number of references; thus a system which makes it cheap to find a set of matching segments in all references for a given probe would be of great value. Bertin-Mahieux and Ellis (2011) presented a system using hashed chroma landmarks as keys for a linear-time database lookup. Their system showed promise, and demonstrated a possible approach to large-scale cover-song detection but the reported performance numbers would not make for a practically viable system. While landmark or "interest point" detection has been extremely successful in the context of exact audio matching in noise (Wang, 2003) its effectiveness in such applications is largely due to the strong invariance in the location of strong peaks in the spectrogram. For cover version identification, the variability between musical performances, both in timing and in pitch, means that descriptors summarizing small constellations of interest points (that is, patterns of interest-point positions) will necessarily be less discriminative than descriptors summarizing more complete features over a long time span. With this in mind, we now explore some options for generating compact hashes of full intervalgrams for indexing and retrieval purposes.

27.3.3 Hashing of the Intervalgram

Using the process outlined above, 32×32 pixel intervalgrams are generated from a signal at the rate of one per 240 ms. To effectively find alternative performances of a melody in a large-scale database, it must be possible to do efficient lookup to find sequences of potentially matching intervalgrams. The use of locality-sensitive-hashing (LSH) techniques over long-timescale features for music information retrieval has previously been investigated and found to be effective for large datasets (Casey et al., 2008a). Various techniques based on locality-sensitive hashing (LSH) may be employed to generate a set of compact hashes which summarize the intervalgram, and which can be used as keys to look up likely matches in a key-value lookup system.

An effective technique for summarizing small images with a combination of wavelet analysis and min-hash was presented by Baluja and Covell (2008) in the context of hashing spectrograms for exact audio matching. A similar system of wavelet decomposition was previously applied to image analysis (Jacobs et al., 1995). The system described by Baluja and Covell (2008) has been adapted to produce a compact locality-sensitive hash of the intervalgram. The 32×32 intervalgram is decomposed into a set of wavelet coefficients using a Haar kernel, and the top t wavelet coefficients with the highest magnitude values retained. If the value t is chosen to be much smaller than the total number of pixels in the image, the most prominent structure of the intervalgram will be maintained, with a loss of some detail.

Compared to exact-match audio identification, this system is much more challenging, since the individual hash codes are noisier and less discriminative. The indexing stage necessarily has many false hits when it is tuned to get any reasonable recall, so there are still many (at least thousands out of a reference set of millions) of potential matches to score in detail before deciding whether there is a match.

27.4 Summary and Conclusions

The intervalgram is a pitch-shift-independent feature for melody-recognition tasks. Like other features for melody recognition, it is based on chroma features, but in our work the chroma representation is derived from the temporal profile of a stabilized auditory image, rather than from a spectrogram. To achieve pitch-shift invariance, individual intervalgrams are shifted relative to a reference chroma vector, but no global shift invariance is used. Finally, to achieve some degree of tempo-invariance, variable-width time-offset bins are used to capture both local and longer-term features.

In this study, the performance of the intervalgram was tested by using dynamic-programming techniques to find the cheapest path through similarity matrices comparing a cover song to all references in the dataset. Intervalgrams, followed by dynamic-programming alignment and scoring, gave a precision at top-1 of 53.8%. This performance value and the associated precision–recall curve lie between the performance of the Ellis 2006 and Ellis 2007 MIREX entries (the latter of which was developed using the covers80 dataset).

The intervalgram has shown itself to be a promising feature for melody recognition. It has good performance characteristics for high-precision matching with a low false-positive rate. Furthermore the algorithm is fairly simple and fully feed-forward, with no need for beat tracking or computation of global statistics. This means that it can be run in a streaming fashion, requiring only buffering of enough data to produce the first intervalgram before a stream of intervalgrams can be generated. This feature could make it suitable for applications like query-by-example in which absolute latency is an important factor.

The intervalgram representation lends itself well to large-scale application when coupled with locality-sensitive hashing techniques such as wavelet-decomposition followed by min-hash. The high precision allows for querying of very large music databases with a low false-positive rate.

Chapter 28

Other Applications

> Computational modeling of the auditory periphery has become an integral part of hearing and speech research in recent years. This reflects the importance of computers and computational models as a research tool for experimenting flexibly in the domain of complex auditory phenomena. Both our general understanding and the fragmental knowledge of details known from hearing research can be reconstructed and tested in the form of functional models.
>
> — "Auditory models for speech processing," Matti Karjalainen (1987)

We have covered a few specific application examples in detail in previous chapters, to illustrate some of the ways that an auditory-image front end can be connected to a higher-level application system. In this chapter, we very briefly survey some of the other areas where front ends based on models of hearing have been used to advantage, and where further advances are to be expected.

28.1 Auditory Physiology and Psychoacoustics

As the chapter-opening quote by Karjalainen suggests, problems in the physiology and psychophysics of hearing, such as those introduced in Chapter 4, can be addressed most fruitfully in the context of computational models. Simple models, such as the Fourier spectrum view of sound, have been useful historically, but they run into limits and lead to questions that can only be addressed in the context of more realistic models. Progress in understanding details of psychoacoustic effects such as loudness, masking, pitch, timbre, etc. have come about gradually as these phenomena have been studied in the context of increasingly detailed models of auditory physiology, especially including cochlear function.

An example is the study of auditory filter models for explaining simultaneous masking, as discussed in Chapter 13. Early auditory filter models were simple functional bandpass concepts (rectangular, Gaussian, and simple resonance filters). Later models incorporated more accurate and flexible filter shape descriptions and level-dependent parameter changes. Most recently, we showed that filter models derived from active wave propagation concepts lead to better fits to the data with fewer adjustable parameters (Lyon, 2011b,a). In this sense, using more knowledge of the physiology has led to better explanation of the psychophysics, and thereby a reinforcement of models of both.

Stabilized auditory images are a promising tool for visualizing a range of psychoacoustic phenomena that are hard to explain purely in the time domain or the frequency domain, for example, for the time-asymmetric ramped-versus-damped effects described in Section 21.6. They are also useful illustrations of models of the physiology, for example when used to illustrate the hypothetical maps in various levels of binaural and monaural neural processing. Thus these lower levels of the machine hearing stack may be increasingly widely used, to study how the auditory nervous system might make sense of stimuli used in psychoacoustic experiments.

For example, stabilized auditory images have found use in studies of the representation of pitch and melody in the brain (Griffiths et al., 2001; Patterson et al., 2002).

28.2 Audio Coding and Compression

Digital entertainment audio is usually stored via specialized compression or coding algorithms that seek to minimize the amount of data required while keeping the resulting distortion inaudible. The MPEG (MP3) and Dolby AC-3 compression systems are particularly interesting in that they standardize the decoder, but leave the encoder free to decide how inaudible different distortions may be. In particular, the encoder optionally incorporates a psychoacoustic model of masking, at the implementer's discretion, to help decide how to allocate information to different frequency bands in each analysis frame, for example by choosing the quantization step size in each band. This approach was introduced by Johnston (1988).

These systems typically use analysis/synthesis filterbanks with critical sampling; (that is, just enough samples of the filterbank outputs to match the original number of waveform samples) and an approximate "perfect reconstruction" property (that is, filtering the samples and adding up the results is an excellent alias-free approximation of the original waveform) (Todd et al., 1994; Painter and Spanias, 2000). The "loss" due to compression comes only from quantizing the samples, and the result of each quantization error is the addition of a small "blip," a windowed sinusoid, to the decoded sound. These errors are reasonably localized in time and frequency, and a psychoacoustic model of masking can reasonably predict where they are relative to a masked threshold for such blips. Therefore, better psychoacoustic models lead to better rate–distortion tradeoffs—if distortion is interpreted perceptually—which is why these are referred to as perceptual coders.

It is possible that a better cochlear model would lead to a more effective model of masking for this application, and hence better compression for a given quality level, either for an existing coding standard or a new one.

In audio production, there are other functions that can benefit from a psychoacoustic model running alongside the main process. For example, in audio steganography, metadata such as rights management information is embedded by adding a low-level coded waveform; using an auditory model, the coded waveform can be adjusted to be just below the masked threshold, such that the effect is inaudible, yet keeping the information robust to the kinds of distortions that a good-quality compression system might add (Boney et al., 1996).

Cochlear models are sometimes used directly for audio coding, rather than just "on the side" for estimating masking of distortions. In this application, an inversion or resynthesis algorithm matched to the analysis algorithm is needed. Several researchers have investigated cochlear model inversion algorithms for this and related applications (Irino and Kawahara, 1993; Slaney et al., 1994; Hukin and Damper, 1989; Yang et al., 1992; Kubin and Kleijn, 1999).

28.3 Hearing Aids and Cochlear Implants

Before the telephone, Alexander Graham Bell worked on aids for the deaf. He developed a simple audiometer to assess the degree of hearing impairment, and experimented with the electrical transduction of sound, leading to the invention of the telephone that he is most known for. He started out with very little electrical knowledge or experience, but took to heart Joseph Henry's advice to "Get it!" (Casson, 1910). By studying human and machine hearing from a scientific and engineering perspective, Bell made impressive technical progress.

Helping the hearing impaired has been the application that has driven much progress in hearing research. From early passive ear trumpets, to the self-contained vacuum-tube hearing aids of the 1940s, to modern tiny high-gain aids that make decisions about what is the signal of interest and present it in a way that makes up for the listener's particular hearing deficit, progress has come from diverse directions.

Before the 1960s, a common hearing deficit was a *conductive* loss: inefficient transfer of sound energy through the outer or middle ear, such that not enough energy was delivered into the cochlea. For this deficit, a simple amplifier was a suitable solution. However, the problem of conductive hearing loss has since been mostly solved, at least in developed countries, by advances in middle-ear surgery. Mostly what remains to be treated by hearing aids is what is generally referred to as *sensorineural* hearing loss (Levitt, 2004): energy gets into the cochlea, but the cochlea doesn't perform as well as it should.

It had been discovered earlier (Fowler, 1936) that most of this class of hearing problems came with a new twist: a limited dynamic range, known as *abnormal growth of loudness* or *loudness recruitment*. A person with a sensorineural loss might experience the complete range of loudness, from very quiet to very loud, when the input level changes over only a moderate range, such as 50 dB to 80 dB SPL. The problem is that the cochlea's normal active compressive behavior is not working well, and the result is as if the input sound level range has been expanded. As Villchur (1973) explained, "A deaf person with recruitment perceives sound as though listening through a volume expander followed by an attenuator, the expansion ratio and attenuation being typically frequency dependent." Treating such a loss with amplification can help the listener hear weak sounds, but then moderate and loud sounds will be perceived as much too loud.

Therefore, for the new main target of sensorineural hearing loss of cochlear origin (mostly caused by partial loss of outer-hair-cell function), the general goal of hearing aids came to be to make up for the cochlea's loss of signal-dependent gain. Weak sounds need to be amplified more than strong sounds, to bring the wide dynamic range of sounds in the world into the narrow range that the impaired cochlea can interpret. The impaired cochlea is typically closer to "passive," with damage to outer hair cells causing its normal active and compressive gain function to be greatly reduced, at least in some frequency ranges. Therefore, compressive hearing aids of one form or another eventually became the norm.

The early compressive aids typically combined amplification and limiting, providing a high gain tempered only by a maximum comfortable output level. They were still not very satisfactory with moderately loud sounds, often causing unpleasant distortion or a net loudness loss at high levels. Aids with an adjustable soft AGC-like compression were found to be much more satisfactory, and could be fitted to different types of hearing loss (Killion and Fikret-Pasa, 1993).

To deal with frequency dependence, multiband compression was proposed (Villchur, 1973). With just two bands, the typical pattern of relatively normal low-frequency hearing and impaired high-frequency hearing could be helped much better than with a single-channel system. A single-channel system could also be made frequency dependent, as in the *treble-increases-at-low-level* (TILL) processing developed by Killion and Villchur (1993).

Part of the success of hearing aid development in the 1980s and 1990s was based on the idea that if listening through a hearing aid sounds bad to a person with normal hearing, it's probably not a good solution for the person with impaired hearing. Villchur (1973) had introduced the idea of testing of hearing-aid processing on normal-hearing subjects in the presence of a masking noise, since the masked threshold and masked loudness growth are a good simulation of a hearing loss with recruitment. He found:

> Speech may or may not sound the same in real deafness as it did in the simulated conditions, but for deaf subjects with equivalent hearing characteristics the same acoustical speech elements will remain below threshold after amplification. It is therefore reasonable to expect that the intelligibility of amplified, unprocessed speech for such subjects will be at least as bad as it was here for normals. The results of this experiment suggest that for deaf subjects with recruitment, amplification that brings the weaker acoustical elements of speech to audible levels without making the high-amplitude elements uncomfortably loud is a necessary although possibly insufficient condition for good speech discrimination.

Villchur (1974) went on to better simulate hearing impairment for normal subjects through an expansive

signal processing. His work led him to a conclusion about a better way to design hearing aids: "the benefit of using both compression and postcompression equalization in a hearing aid designed to compensate recruitment is likely to be considerably greater than the arithmetic sum of the separate, limited benefits of each process."

This approach led to some other experimenters listening to their own hearing-aid signal processing, and to a focus on "high-fidelity" hearing aids to avoid the kinds of distortions common in hearing aids of that time. Killion and Tillman (1982) suggested that "An essential building block for any high-fidelity hearing aid is an amplifier–transducer–coupling combination that does not audibly degrade the sound, that is, provides high-fidelity sound reproduction as judged by someone with normal hearing." Killion (1997) later summarized the slow progress using this approach:

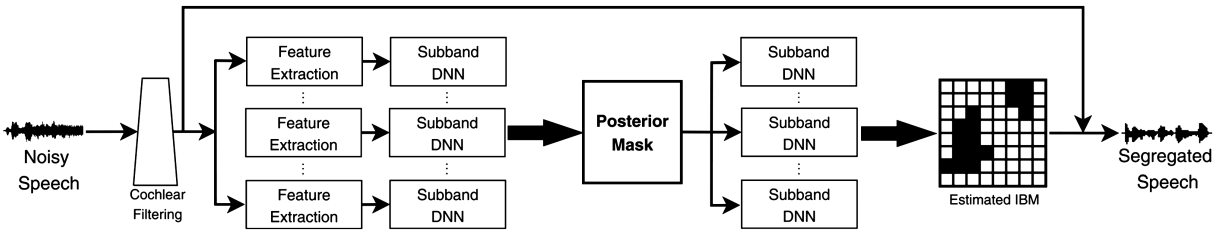
> Some time ago the distortion, narrow bandwidth, irregular response, and inappropriately adjusted (or available) frequency response of hearing aids garbled or muffled so much speech information that hearing in noise was almost impossible. ... By 1990, we had reached the point where hearing-impaired individuals routinely heard better in noise *with* the better hearing aids available; sometimes 5–10 dB better in low-level noise and no worse even in high-level noise.

By the 1990s, better hearing aids used either wideband compression (Killion's K-Amp technology) or multiband compression (the ReSound hearing aids), though many studies in the 1970s and 1980s had reported mixed results on the efficacy of these strategies. As Plomp (1988) pointed out, the problem with many of the early attempts was that they tried too hard to bring signals back to a normal level, adapting their gain too quickly in time, or too independently in different frequency channels, with the result that important contrasts were reduced too much, along both the time and frequency dimensions.

Killion and Villchur (1993) were eventually able to report a net win on intelligibility of speech in noise by hearing-impaired listeners with appropriately fitted high-fidelity compressive aids. They reported improvements in signal-to-noise ratio threshold of 3–9 dB with binaural aids, compared to a net loss at high levels for linear peak-clipping hearing aids. This noise tolerance improvement was accompanied by a big improvement in subjective quality and comfort, so this kind of hearing aid would actually be used more, especially in noisy situations. A typical user might have previously just removed or turned off their aids due to the poor sound quality in loud settings.

The number of bands to use in a multiband compressor has been a constant source of argument in the hearing aid field. It was easy to demonstrate an advantage with two bands, but harder with more. Part of the problem was that the compression in each band was usually dependent only on the energy in that band, rather than coupled to neighboring bands. Schneider and Brennan (1997) note that, "the compression in a given band can be controlled by the signal level in that band. ... multichannel compression schemes inherently tend to 'flatten' the spectrum of the output signal." This excessive flattening is because their channels are not coupled in a way that would be a good model of cochlear processing.

Many others had the same interpretation of multiband as independent band compressors. To some extent, the spectral flattening problem of independent channel gains was alleviated by separating the compression into several stages, not all with independent channels, or by compressing different spectral-shape basis functions (principal components or low-order polynomials) by different amounts (Levitt, 2004). Combining a slow wideband AGC with less aggressive *syllabic*-time-scale compression in each channel was one design pattern that became successful after positive results from studies in the late 1980s (Levitt, 2004). This multi-time-constant approach resembles how we model AGC in the cochlea. The CARFAC model's fastest AGC time scales may not be appropriate in a hearing aid, as they capture effects at the inner hair cells that are still operating in the hearing-impaired listener. Adding a longer-time-scale wideband AGC makes sense, to model the reflexive gain adjustment of the middle ear.



Schematic diagram of the current speech-segregation system. DNN = deep neural network, IBM = ideal binary mask.

Figure 28.1: The hearing aid architecture of Healy et al. (2013), using a cochlear filterbank and binary masking, has been shown to yield a net intelligibility improvement for hearing-impaired subjects in noisy situations. [Figure 1 (Healy et al., 2013) reproduced with permission of AIP Publishing.]

Another approach to solving the problem was presented by Asano et al. (1991): analyze many bands, but then use that result to design a compensation filter whose response is smoothly changing in both the time and frequency domains. Their *compensating loudness by analyzing input-signal digital hearing aid* (CLAIDHA) algorithm was later evaluated in combination with a hearing-impairment simulator by normal listeners, and found to outperform a more conventional multiband compression filterbank hearing aid (Chung et al., 1996).

Yund and Buckles (1995) point out that using many compression bands usually doesn't hurt, but also doesn't help much beyond eight bands; they hypothesize that the sort of "cross-coupled" compression proposed by White (1986) might help by enhancing contrast between bands. Kates (1993b) makes a similar point in showing how a hearing aid's gain-versus-frequency profile can adapt to the sound to make an "optimal" remapping toward the response expected from a normal cochlea, using a simplified cochlear model with coupling from an octave below to a half octave above each frequency channel to control the gains. These proposals are essentially the application of concepts from cochlear modeling (along the lines of our CARFAC model's coupled AGC) to hearing aids: restoring realistic frequency-dependent nonlinear compression for a listener whose cochlea has gone passive.

Schneider and Brennan (1997) took a step toward a more coupled approach, in combining independent channel gain controls with an overall gain control, but the effect of this compromise on preserving contrast between channels was thereby not dependent on the frequency separation of the channels. Hamacher et al. (2005) describes this approach as state of the art in hearing aids. Only recently have researchers developed hearing-aid strategies that couple the channels in a psychophysically realistic multiband compressor—described as incorporating cross-frequency *suppression* effects (Rasetshwane et al., 2014).

A properly adjusted many-band compression hearing aid should never perform worse than a linear amplification hearing aid, or an aid with fewer bands, since it should be adjustable to be close to the behavior of the simpler device if that's the best fit; yet worse performance of multiband aids is still sometimes reported (Kates, 2010). Better adjustment and better cross-channel coupling should at least fix this anomaly.

Irino and Patterson (2006b) describe an efficient gammachirp-based analysis/synthesis filterbank for hearing aids and other applications, incorporating a "fast-acting level control circuit." Our CARFAC could be used similarly, for quickly adapting the gain versus frequency, with appropriate coupling, though we have not worked out a good resynthesis filter structure yet.

Leveraging the possibility of adaptive frequency-dependent gain, modern hearing aids sometimes incorporate higher levels of machine-hearing-like functionality, attempting to identify and amplify the signal of interest, while suppressing noise. Kollmeier et al. (1993) combined multiband compression with binaural processing that enhances sound from straight in front of the listener and suppresses sound from other directions, including reverberation. This combination of strategies, essentially a CASA approach (see Chapter 23),

was found to provide substantial improvements in speech intelligibility for most of the hearing-impaired listeners that it was tested on. Wittkop et al. (1997) combined these strategies with algorithms that used the modulation patterns of speech syllables and the fundamental frequency patterns of speech to more reliably decide what parts of the sound to amplify. They predicted that "a combination of these algorithms appears promising for future 'intelligent' digital hearing aids." Rohdenburg et al. (2008) extended the binaural approach to hearing aids that are able to automatically focus their attention in different directions. They used a 6-microphone configuration for better selectivity. Most recently, a time–frequency mask estimated from a monaural signal has been shown capable of yielding good intelligibility gains for hearing-impaired listeners in noisy sound mixtures (Healy et al., 2013); Figure 28.1 shows their system, based on a cochlear-model filterbank.

It is hard to know exactly which techniques have found their way into which hearing aids, but most hearing aid makers do claim a variety of directional, speech enhancement, and noise reduction features in their high-end digital products.

Levitt (2007) recounts how some of these ideas made their way into digital hearing aids, mostly based on efforts from outside the hearing-aid industry:

> It is significant to note that the companies that led the way in implementing digital technology in hearing aids [Nicolet, AT&T/ReSound, and 3M] were not traditional hearing aid companies but rather major industrial companies with a history of innovative research and development. These companies also introduced new ideas and methods that had a lasting impact on the field.

Edwards (2007) reminds us that the hearing approach is not necessarily aligned with work done in other fields, such as those that work specifically on speech and music:

> Most audio industries have very specific types of sound that they process. The telecommunications industry usually processes speech at high signal-to-noise ratios; the music industry processes only voice and musical instruments, often separated into individual tracks; the teleconference industry processes only sounds that exist in conference rooms, such as speech and air conditioning noise.
>
> Hearing aids, however, have to be able to process all possible sounds with imperceptible distortion and good perceived quality for someone listening all day long. In other words, they have to be able to handle every sound and any sound in all possible combinations.

Another approach that is sometimes used in hearing aids, especially for a sloping severe-to-profound hearing loss, is frequency lowering (McDermott, 2011). Lower frequencies (say, up to 1.5 kHz) are processed in the usual way with compressive amplification, while higher frequencies are compressed along the frequency axis (for example, mapping two octaves from 1.5 to 6 kHz down into one octave from 1.5 to 3 kHz, thus extending the range of frequencies that the subject can hear by about an octave). Another approach is to translate a band of higher frequencies to overlap with some lower frequencies. Since both of these approaches are highly nonlinear, care is needed in finding a way to make the sound intelligible, if not natural.

Cochlear implants are another class of hearing aid. Rather than provide an amplified vibration to a somewhat defective cochlea, they provide electrical stimulation to the primary auditory neurons of the spiral ganglion, via an electrode array placed in an otherwise nonfunctional cochlea. It is very challenging to get enough electrodes, distributed over enough of the range of the cochlea, and to control the spread of electrical interaction, with the result that it is impossible to stimulate the nerve in a way that is realistically close to what a functioning cochlea would do. Nevertheless, the implant stimulators are typically based on cochlea-model-like filterbanks, combined with preprocessors that attempt to identify and emphasize speech—and people do learn to interpret the signals as speech. For music listening, implants are much less satisfactory than for

speech, but progress is being made; rhythm perception is good, but melody and timbre are not, suggesting that better spatiotemporal cue patterns are needed (McDermott, 2004; Loizou, 2006).

Interestingly, early cochlear implants with only a single channel, and therefore no frequency–place mapping cue, worked fairly well for some subjects. As Loizou (1998) noted, "It remains a puzzle how some single-channel patients can perform so well given the limited spectral information they receive." This tantalizing sometimes-success of coding via temporal structure in these single-channels devices has not been very successfully replicated in multichannel implants, despite multiple attempts. It is theorized that perhaps the time–space patterns that can be delivered via implants cannot approach the patterns that the cochlear nucleus expects.

Since fine time structure seems to be mostly ineffective in implants, simulations of them for normal-hearing listeners are sometimes done using modulated bandpass noise, as in a noise-excited vocoder. This allows researchers to get some idea what kind of information might be deliverable by the implant processor. Li et al. (2013) have recently shown improved performance in vocoder simulations with normal-hearing subjects and at least partially corresponding improvements in melody recognition by cochlear implant users, for their *harmonic-single-sideband-encoder* strategy that attempts to use some synchrony and other effects to convey melodic and harmonic relationships.

Even if we succeed at delivering a reasonable signal to the auditory nerve, whether via conventional hearing aids or by cochlear implants, there are still other hearing deficits that present a challenge. Generally, impairments beyond the auditory nerve are known as *auditory processing disorder* (APD) (Levitt et al., 2012). APD, like other hearing problems, tends to increase with age, with the result that APD is partially correlated with other measures of hearing loss (Aydelott et al., 2010). The more we learn about auditory processing in the brain, and how it goes wrong, the more we will be able to modify sounds to try to make them easier to process (Kricos, 2006), and the better we will be able to train people to work around the problem (Pichora-Fuller and Levitt, 2012).

My friends in the hearing-aid field—Harry Levitt, Hugh McDermott, James Kates, and Tao Zhang—have been generous with their feedback and criticism of this section. Still, I have presented here my own views, which are somewhat at odds with theirs in some areas.

28.4 Visible Sound

There is a long history of using images of sound, especially speech sounds, as aids to the hard-of-hearing and for other purposes. The *phonautograph* and *manometric flame* (see Figure 28.2) were devices experimented with by Alexander Graham Bell, among others, in attempts to make tracings or visualizations of sound waveforms that a deaf person could try to match with their own voice. He wrote "If we can find the definite shape due to each sound, what an assistance in teaching the deaf and dumb!" (Bruce, 1990; Lepore, 2002). Such attempts were mostly not welcomed by the deaf community, as Bell and others were essentially trying to do away with their sign-language culture and make them fit in with the speaking/hearing culture. But there were also technical reasons why it could not work: fundamentally, it is just too hard to tell much about a sound from its waveform.

Spectrograms, introduced by Steinberg and French (1946) and promoted by Potter et al. (1947), were a good improvement, but still very hard to read and make sense of, and not very practical as a real-time display. The stabilized auditory image, on the other hand, being a "movie" representation suitable for visual input, may be more usable for visualizing sounds.

A good sound visualizer has many potential uses. It will probably never provide a communication channel as good as sign language, but there are many other reasons why deaf or hard-of-hearing people, or cochlear implant users, might benefit from a supplementary channel of sound access—and not just for speech. In the case of speech, the SAI might be an excellent supplement to lipreading, since the pitch and voicing information

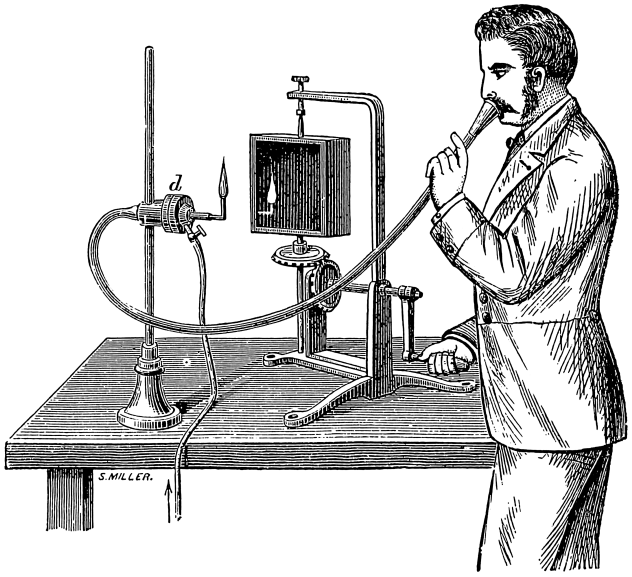


FIG. 432.—König's apparatus for illustrating the quality of vowel tones by a manometric flame.

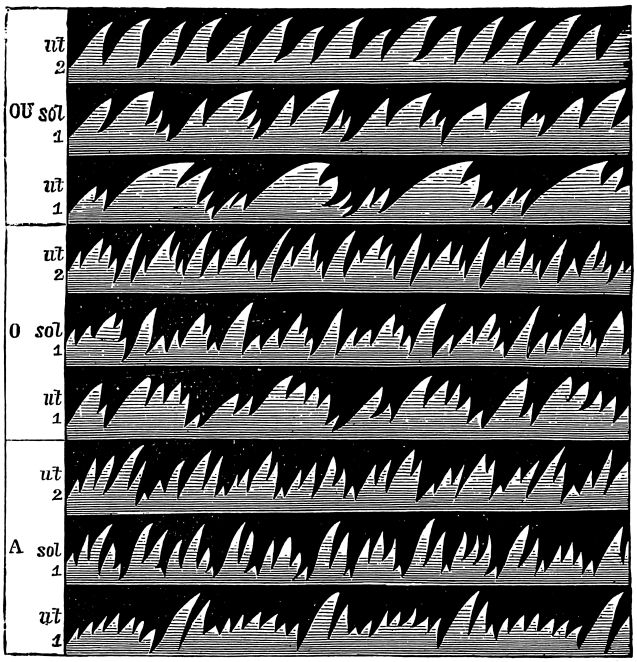


FIG. 433.—Flame pictures of the vowels *OU*, *O*, and *A*.—König.

Figure 28.2: Rudolf König's manometric flame apparatus was among the devices that Alexander Graham Bell used to visualize sound waveforms. The rotating quad mirror converted fast sound-induced modulations of the flame into spatio-temporal patterns—but not stabilized. These images from McKendrick (1889) show how the experimenter might use the apparatus, and the flame patterns that would appear in the mirror in response to several steady vowels, each at several different pitches.

("manner of articulation" information) that shows up in the SAI is a good complement to the mostly "place of articulation" information that comes from watching the lips. A simple eyeglass-mounted visual complement for lipreading, based on manner of articulation, was presented by Upton (1968). It used sound analysis logic driving five lamps to indicate five categories: voiced sound, unvoiced fricative, unvoiced stop, voiced fricative, voiced stop.

For nonspeech, the SAI display might be very good for alerting a deaf person to doorbells, alarms, knocks, barks, squeaks, breakage, music, and a myriad of other sounds that convey useful information about what's going on in the nearby environment, or in the show they are watching. Ho-Ching et al. (2003) report that "our participant was able to detect a number of sounds including: speech, mobile phone calls, chair movement, typing, mouse movement, page turning, papers rustling, footsteps of people entering the office, and a university truck which turned around outside several times," with a spectrogram display. Many such sounds may be easier to distinguish from speech, and from each other, in an SAI display. Matthews et al. (2005) discuss many other types of sounds that deaf users were happy to be able to see in a display.

For normal-hearing users, a good speech display has uses, too. For example, for second-language learning, accent adjustment, singing-voice training, and such, an SAI or cochleagram/pitchogram display may be a good tool for a user to compare their own sounds to the sounds of an instructor—including pitch contours. The ear itself might be a better tool for many kinds of sound comparison, but speech can be confusing: it is very hard for a user to objectively compare their sounds when the target language or accent has different phonemic or prosodic distinctions from those of the user's native language or accent. There is already a significant use of speech displays (usually based on spectrograms) in computer-assisted language learning (CALL) or computer-assisted pronunciation training (CAPT), but the spectrogram is not the best representation for these purposes. An auditory spectrogram augmented with a pitchogram may be a step in the right direction. The use of a pitch contour display has been shown to be useful for English learners of Japanese (Hirata, 2004).

Of course, not all talkers use the same sound for the same speech. One key to a good visual representation is that it should not make talkers of different pitches and different vocal tract lengths look too different, the way spectrograms often do. To the extent that the place dimension of an SAI maps frequency logarithmically, and the lag axis of an SAI or pitchogram maps pitch period logarithmically, the patterns of vowels will mostly just shift in two dimensions with variation of these talker properties. Patterson et al. (2007) has explored such "scale–shift covariant" and other SAI-based normalized representations for speech.

28.5 Diagnosis

Sound is useful for diagnosis of problems, in medical and mechanical fields among others. While there is no particular reason to expect that human hearing is the ideal model for analyzing sounds to make features for a diagnostic classifier, many of the properties of human hearing will probably be helpful in that direction. For example, a stabilized representation of the temporal and spectral structure of a sound, as in an SAI, may capture much of the relevant information and make a good input to a learning system.

Doctors have a long history of diagnosing by listening—they call it *auscultation*, a term introduced by René-Théophile-Hyacinthe Laënnec (1819), the inventor of the stethoscope. Leatham (1951) discusses the use of the *phonocardiogram* for the diagnosis of aortic stenosis. Since that time, many techniques have been explored to help automate the interpretation of such heart sound recordings for various conditions. In their survey of techniques, Rangayyan and Lehner (1986) note that "The heart sound signal has much more information than can be assessed by the human ear or by visual inspection of the signal tracings on paper as currently practiced." Unlike most methods explored so far that are not auditory inspired, Sung et al. (2013) have proposed an auditory-inspired method for analysis of certain types of heart murmurs. Their method uses both a "cochlear spectrogram," based on a linear gammatone filterbank, and a "temporal correlogram" (but by basing the correlogram on the Hilbert envelope of filterbank outputs, it misses the opportunity to use or

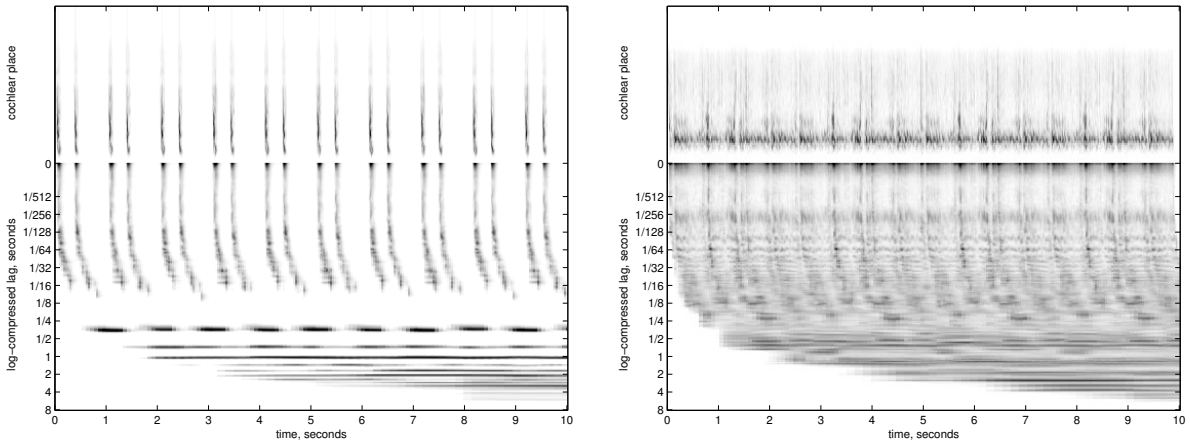


Figure 28.3: Phonocardiograms visualized with the cochleagram/log-lag pitchogram representation of Chapter 21, showing a normal clean low-frequency "lub-dub" heart sound on the left, and a heart with *patent ductus arteriosus* (PDA) on the right. The heart with PDA has a continuous murmur, or "machinery murmur"—essentially a modulated noise from blood squirting under pressure through an opening that should not be there. Other heart problems have more subtle sonic signatures.

represent fine temporal structure—an SAI might work better).

In any of these approaches, it is important to account for the fact that the relevant range of frequencies for heart sounds is somewhat lower than the normal range of human hearing. This low frequency range sometimes makes effective auscultation by human doctors difficult, so machine techniques may be a helpful addition. On a personal note, a friend just had heart surgery for *patent ductus arteriosus* (PDA), a condition normally caught in infancy, but which had gone undiagnosed for 56 years in spite of symptoms including a classic PDA "machinery murmur," visualized in Figure 28.3.

Sound is often used by automobile mechanics, as many of us know from listening to the NPR radio show *Car Talk* with "Click and Clack the Tappet Brothers." Yadav and Kalra (2010) show how to diagnose tappet clicks among other engine noises, using spectrograms and neural networks. Liu et al. (2011) connect their sparse-coding approach to machinery fault diagnosis with the remarkable ability of skilled humans to use their auditory systems to diagnose machines that they are familiar with. There may be a role for better auditory models in such work.

Frequencies above the range of human hearing are also useful in detection and diagnosis. For example, ultrasonic gas leak detectors find signals in the 20–60 kHz range (Naranjo and Baliga, 2009). It would be interesting to see if auditory techniques would find more distinctive features in such signals, to help better classify them.

28.6 Speech and Speaker Recognition

The field of automatic speech recognition (ASR) is perhaps the single biggest application area for the machine hearing approach. As pointed out in Chapter 5, the field has adopted many hearing-inspired ideas, but has in recent decades been dominated by concerns at higher levels than how best to represent the sound input. Speaker recognition and verification has been a smaller but also important part of the speech field. Since speakers differ greatly in pitch, the pitch dimension that is typically ignored in ASR is somewhat more used in speaker recognition.

Much of the early research in machine hearing (including my own) was conducted in the context of ASR.

We surveyed some of the sound representations used in the speech field in Chapter 5. Since there is a huge literature in ASR and in machine hearing applied to ASR, and since some of it has been discussed where relevant throughout this book, we do not attempt to cover this field in more detail here.

With the proliferation of networked mobile and wearable devices, ASR is becoming a widely used and very effective technology. Improvements to ASR from better sound representations, including spatial/binaural processing and more use of pitch, are likely to be increasingly valuable. So we expect this to remain an important area for work on applying a machine hearing approach.

28.7 Music Information Retrieval

Like speech, music is an old and important application area for machine hearing, and is one of the contexts in which important early machine hearing research was conducted, for example at Stanford's CCRMA (Mont-Reynaud, 1992). Music information retrieval (MIR) has become a big field of its own since then.

Much of the leverage in modern recommendation and retrieval systems comes not from the sound, but from metadata and user behavior data—known as *collaborative filtering*. For example, Netflix's movie recommendations are based on the behavior of users, much more than on the content of the movies. Content-based methods remain important, too, and have a lot of room to improve. Content analysis, or meaning from sound, is useful especially for new music, new videos, etc., that do not yet have much user history associated with them; and for other applications, such as the sound retrieval we described in Chapter 26, or tasks where sound is the query. Casey et al. (2008b) describe "content-based music information retrieval" this way:

> In addition to metadata-based systems, information about the content of music can be used to help users find music. Content-based music description identifies what the user is seeking even when he does not know specifically what he is looking for. For example, the Shazam system (shazam.com), described in [(Wang, 2006)], can identify a particular recording from a sample taken on a mobile phone in a dance club or crowded bar and deliver the artist, album, and track title along with nearby locations to purchase the recording or a link for direct online purchasing and downloading. Users with a melody but no other information can turn to the online music service Nayio (nayio.com) which allows one to sing a query and attempts to identify the work.

Sukthankar, Ke, and Hoiem (2006) show how to adapt computer vision techniques that are used in systems such as face detectors to recognize and localize sound objects, and to extract semantically meaningful relationships from sounds, with application to MIR. Weston, Bengio, and Hamel (2011) also describe an approach based both on sound and on multiple semantic relationships:

> Music prediction tasks range from predicting tags given a song or clip of audio, predicting the name of the artist, or predicting related songs given a song, clip, artist name or tag. That is, we are interested in every semantic relationship between the different musical concepts in our database.

Rather than tackle these semantic relationships explicitly or individually, however, they describe a large-scale system design in which the audio features and the semantic features all map to points in a moderate-dimensionality embedding space that is jointly trained to support all of these tasks.

Of course, the performance of such a system will be better if the audio features include, in a usable way, the information needed to help identify musically-relevant aspects of the sounds, including melody, harmony, key, tempo, rhythm, verse/chorus structure, voice, instrumentation, and more complex or subtle properties of artists and performers. Most current work on MIR systems still uses rather impoverished representations such as MFCCs or spectrograms, but with powerful learned embeddings and other modern machine learning

techniques, it should be possible to take advantage of more information, such as from the log-lag SAIs and pitchograms shown in Chapter 21. The melody matching system that we described in Chapter 27 is our first example of using auditory models for MIR.

28.8 Security, Surveillance, and Alarms

Burglar alarms use glass-breakage detectors; gunshot localization systems use time differences between sharp sound events to determine their source; hidden microphones and wire taps pick up talk of conspiracy (and lots of other more innocent things!). But these systems designed for rare outlier problem events are just the tip of the iceberg. In the long run, I expect to see huge value from more routine audio monitoring by machines that hear—to monitor my home, my car, my office or factory, perhaps my pet, and to alert me when something might need attention, or when a rare bird has been heard in my yard.

There will of course be big privacy issues raised by any attempt to monitor sounds, and I make no pretense to knowing how these issues will be resolved. Some who have studied the issue, in the camera domain especially, are optimistic. Brin (1998) writes of a *transparent society* being the result of universal information recording and access. Mann et al. (2003) have coined the term *sousveillance* (looking from below) for the democratized inverse surveillance by the rest of us. Surely we can choose to have the sounds in our own homes captured, analyzed, and maybe recorded; or we can choose not to.

Features that we use for sound retrieval from text queries, as in Chapter 26, should also support going the other direction: from sounds to class labels. An important property of these systems will be their ability to learn while deployed, so that sounds characteristic to the place will be learned, classified, and often ignored, and unknown sounds will stand out as worthy of further investigation. Features representing sounds not recognized might be sent to a central service to see if such sounds are known in other contexts. Specialized services for birdsong, for engine noises, and so forth, might be useful.

28.9 Diarization, Summarization, and Indexing

Analyzing and annotating the soundtrack of a recording is very much like the surveillance problem, except that it doesn't need to happen in real time. If someone has already made a recording, and uploaded to a public site such as YouTube, there is less of a privacy concern in analyzing it. Similar analyses of more private recordings, such as of meetings, would present more of a privacy issue, but it will sometimes be appropriate, and valuable, to analyze and annotate those as well.

For existing sound libraries, there is often no very good description of what the recordings contain. Being able to bulk annotate, summarize, and index sound libraries would be very useful. The speech content will of course be useful, when the recordings contain speech, but other sounds will be just as useful to recognize and index as well. Even simple annotations such as marking the boundaries where the sound changes character, or where loud or interesting events are, will be useful for anyone who wants to listen to parts of a recording to learn what is there.

Clarkson et al. (1998) did early work on analysis of audio from wearable computers, "for obtaining environmental context through audio for applications and user interfaces ... identifying specific auditory events such as speakers, cars, and shutting doors, and auditory scenes such as the office, supermarket, or busy street." Ellis and Lee (2004) recorded continuous personal recordings and explored methods to segment, visualize, browse, and annotate them. They also proposed a scrambling method that would make it difficult or impossible to recover speech from the recording, but still allow classification of different locations, and allow other organizational functions. Reynolds and Torres-Carrasquillo (2005) review methods of segmenting, clustering, and indexing long sound recordings, indicating who is speaking at what time. An important step is to

determine which parts are speech, a problem known as *voice activity detection* (Lee and Ellis, 2006). Truong and Hayes (2009) explore technologies for workplace, educational, and personal lifetime recording, and investigate applications and access methods. Hearing-based representations of sound are likely to be useful in improving all of these systems.

28.10 Have Fun

My favorite recent machine hearing application discovery is in automatic control of coffee roasting (Wilson, 2014):

> The sounds of first crack are qualitatively similar to the sound of popcorn popping while second crack sounds more like the breakfast cereal Rice Krispies in milk. Additional qualitative audible differences between first and second crack are: first crack is louder, first crack is lower in frequency, and individual second cracks occur more frequently within the chorus than first cracks. The purpose of the present work is to quantify these effects as a preliminary step toward the development of an automated acoustical roast monitoring technique.

It could work for corn popping, too.

It is my hope that everyone who reads this book will think of their own applications, and be equipped to work on them. Some will be valuable, commercially or socially, and others might be just great fun. I hope the lower-level bio-mimetic sound analysis layers that I have provided, along with the ideas and examples about how to structure an application at the higher levels, will put readers in a position to actually undertake to build some of these fun and valuable applications.

Color Plates

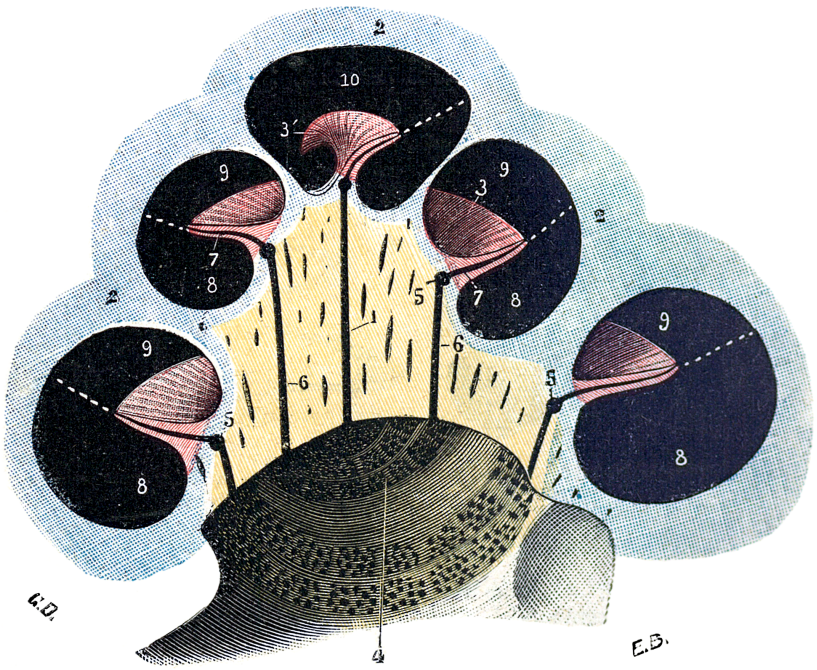


Fig. 890.

Coupe transversale du limaçon osseux : l'un des segments, vu par sa surface de coupe (*demi-schématique*).

Part III introduction. A semischematic transverse cut of the "bony snail," the cochlea, published in color by Leo Testut (1897).

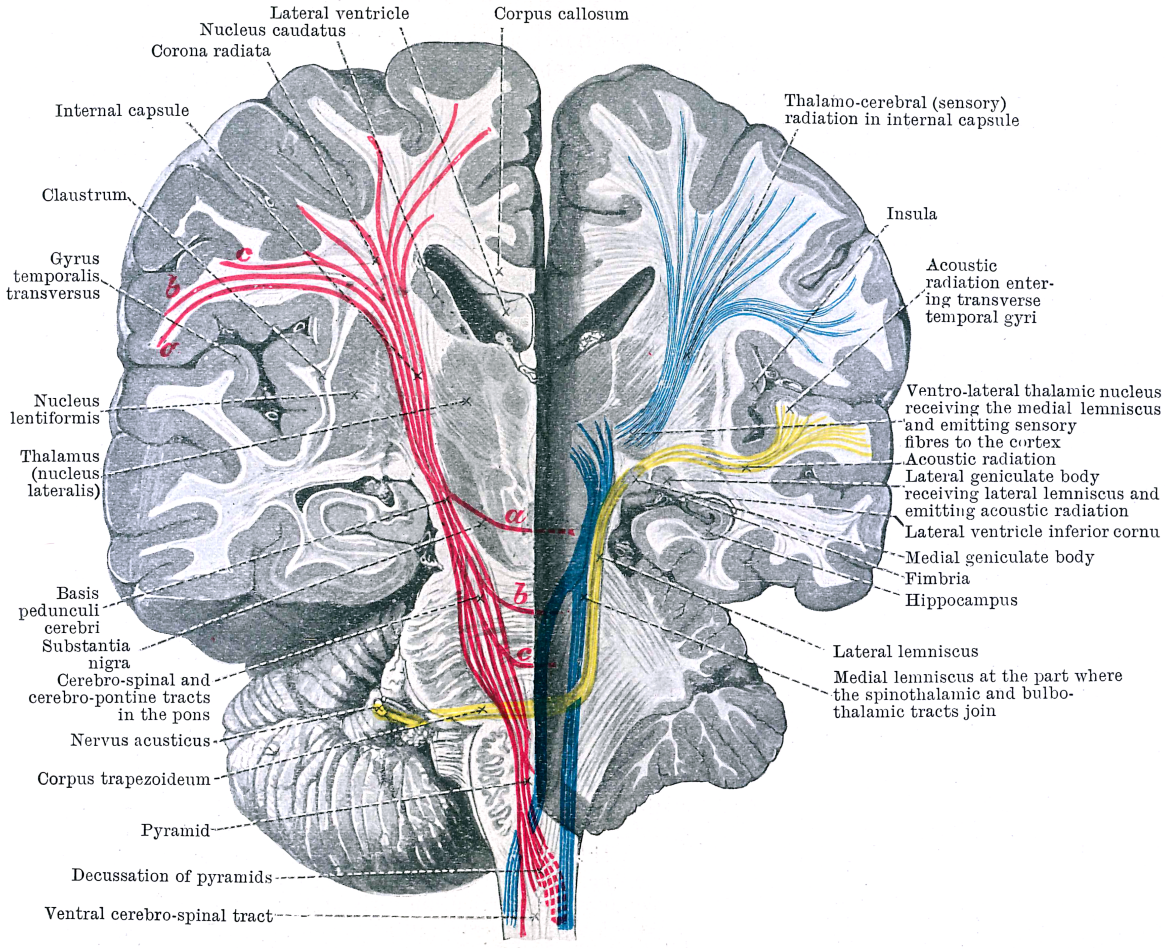


FIG. 506.—A VERTICAL TRANSVERSE SECTION OF THE BRAIN TO SHOW THE WHOLE OF THE CENTRAL ACOUSTIC PATH. The left hemisphere (right side of the figure) is cut on a plane posterior to that of the right. Motor fibres red. Sensory fibres blue. Acoustic fibres yellow.

Part IV introduction. The auditory nervous system was already fairly well mapped out a hundred years ago, as this color illustration from *Cunningham's Text-Book of Anatomy* shows (Cunningham and Robinson, 1918). Auditory fibers are yellow.

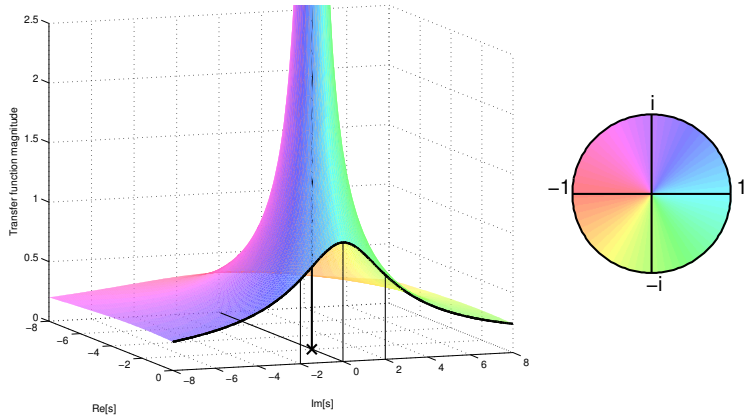


Figure 6.4: The transfer function of the example RC filter of Figure 6.1. The magnitude of $H(s)$ is plotted as a surface height above the complex s plane, while the phase of $H(s)$ determines the hue of the surface color (following the phase–hue legend on the right). For the example filter with $\tau = 0.5$, the transfer function has a singularity at $s = -2 + i0$ (at cross and heavy vertical line). The surface is cut along the imaginary s axis to reveal the frequency response. The frequencies $\omega = 0$ (DC) and $\omega = \pm 2$ (the 3-dB points) are marked by lines.

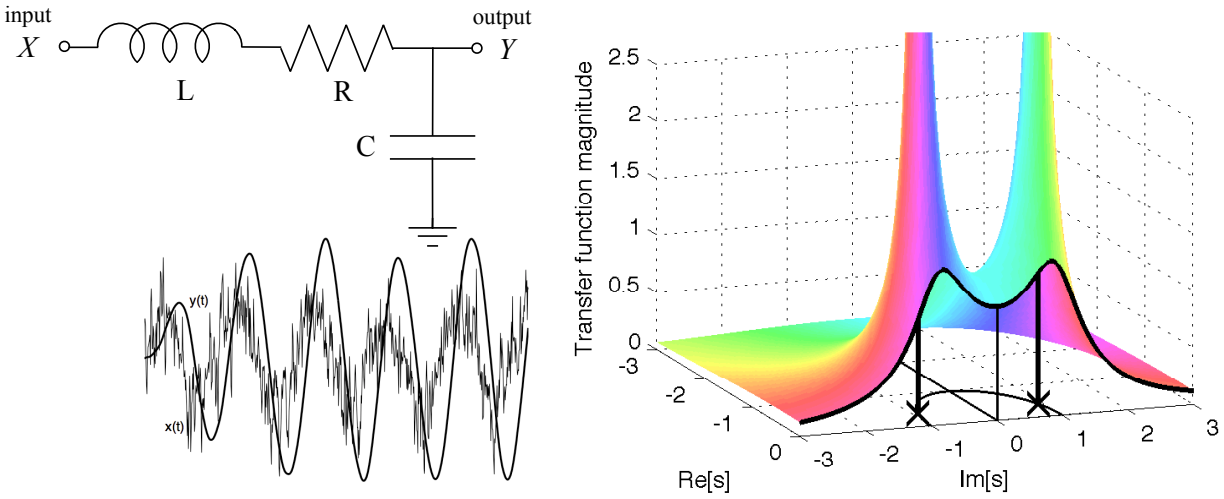


Figure 6.9: Filter A, a second-order resonant lowpass filter, is diagrammed (left) with an example of a noisy input waveform and a corresponding output waveform, which is smooth but has an increased amplitude of the input component that is close to the resonance frequency. The transfer function of filter A (right), plotted as in Figure 6.4, resembles a tent fabric draped over a pair of "tent poles" at the singularities, the two complex pole positions (at crosses and heavy vertical lines).

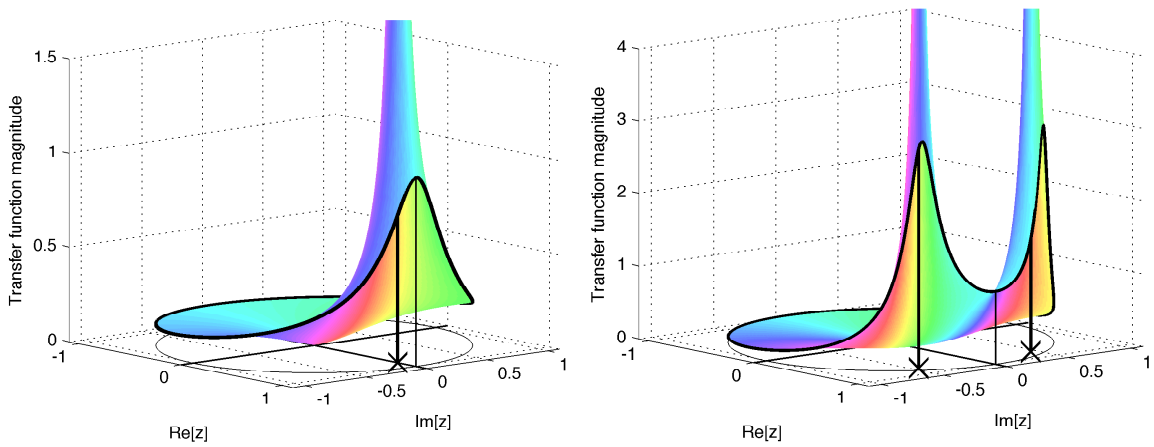


Figure 7.2: Complex transfer functions of one-pole (smoothing) and two-pole (resonator) filters, evaluated inside the unit circle of the z plane. The frequency response is the transfer function evaluated on the unit circle, shown by the dark curves at the circular cut. As in Figure 6.4, phase is mapped to hue; there is one cycle of hue variation around each pole.

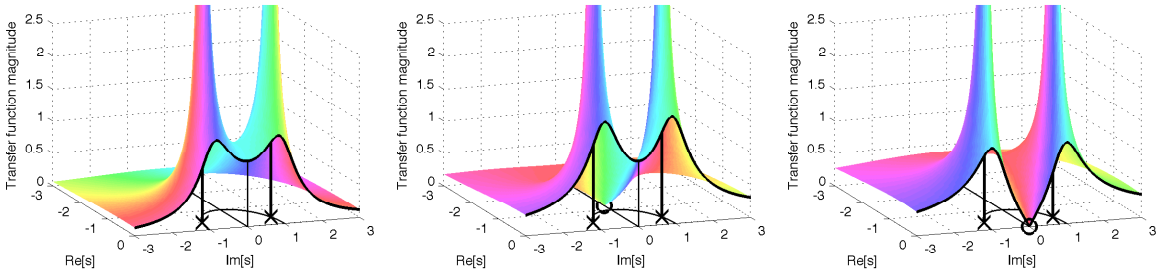


Figure 8.5: The transfer functions of the resonator filters A, B, and C, for natural frequency 1 and damping factor 0.4.

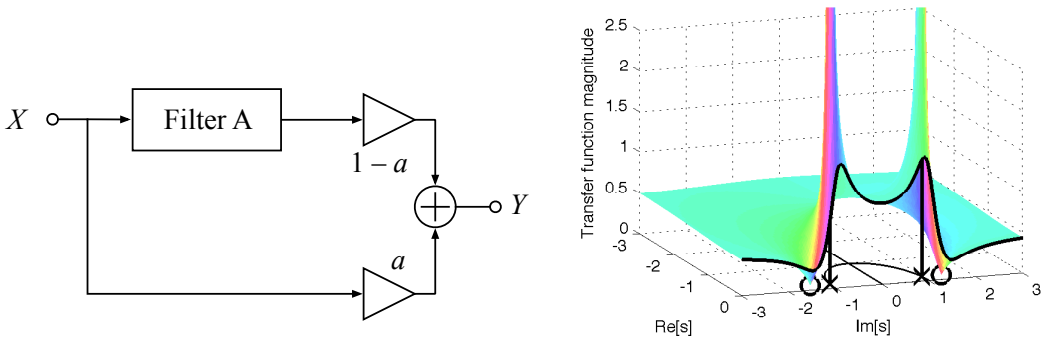


Figure 8.12: Filter D: an asymmetric resonator—schematic and complex transfer function. Adding a straight-through path in parallel to the two-pole resonator of filter A results in a strongly asymmetric peak in the frequency response, involving a complex pair of zeros in addition to the poles inherited from filter A. The ratio of the path gains sets the zero positions. The DC gain is the sum of the path DC gains; as shown, the net DC gain is 1. The illustrated transfer function is for $a = 0.5$ and $\zeta = 0.2$, half the damping of the poles in the illustrations of Figure 8.5, since the zeros near the poles would make the frequency response fairly flat with the higher damping.

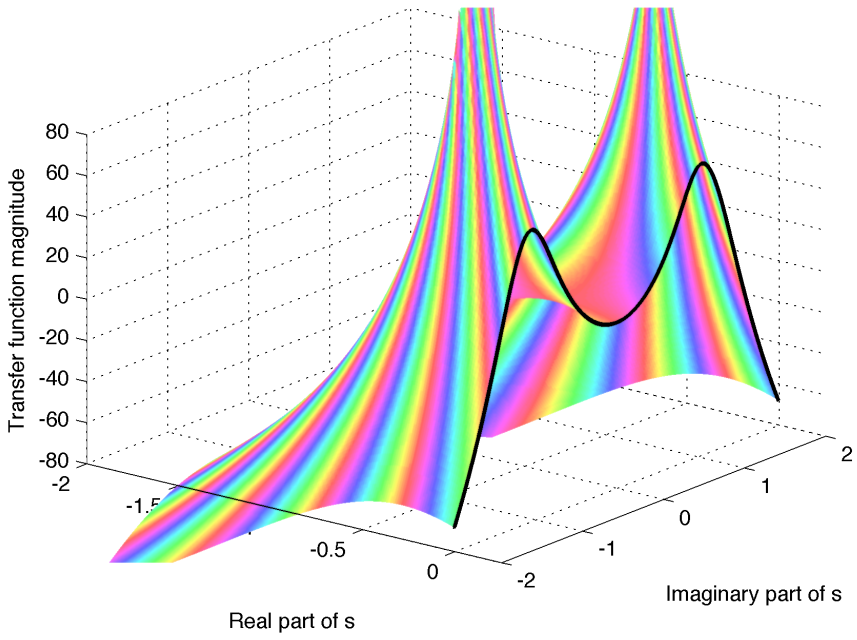


Figure 9.13: The complex transfer function for the Kim et al. filter. The cut line on the imaginary s axis shows the log-magnitude transfer function, with zero frequency in the center. The phase (hue) goes through 10 cycles around each cluster of 10 poles.

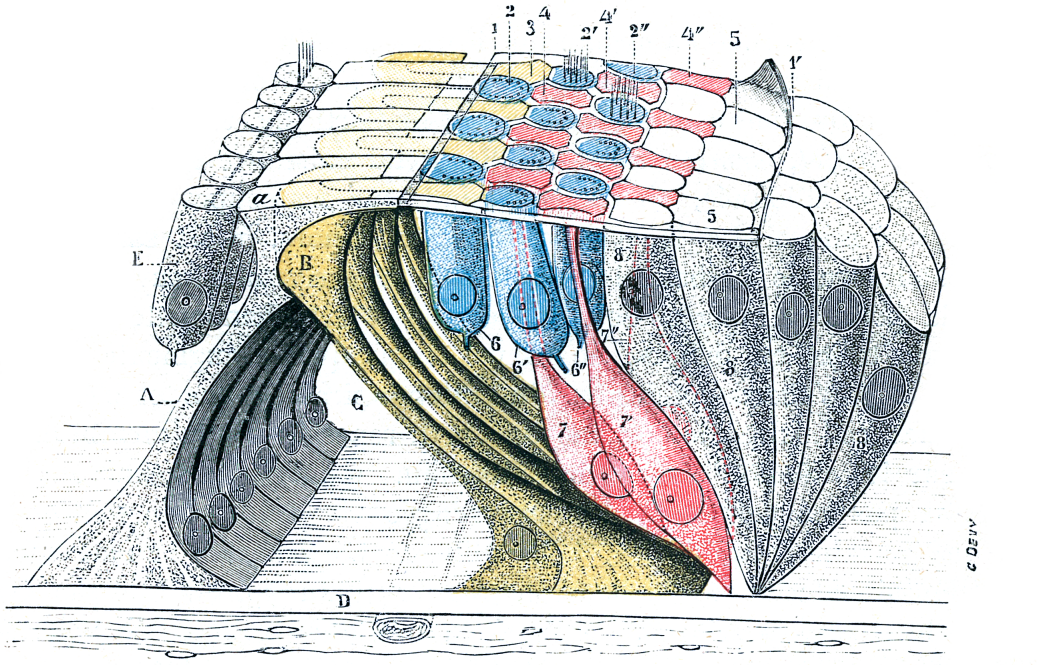


Fig. 918.

La même membrane, avec les cellules qui lui servent de substratum et dont l'empreinte lui donne son aspect réticulé (*schématique*).

Figure 14.2: The three rows of outer hair cells (blue) and one row of inner hair cells (E) sit with their upper ends and hair bundles exposed to endolymph in the scala media through the reticular lamina, but otherwise surrounded by a sealing barrier made up of the pillar cells (A and B), cells of Dieters (red), and other cells of the organ of Corti. This beautiful colored image was published more than 115 years ago (Testut, 1897).

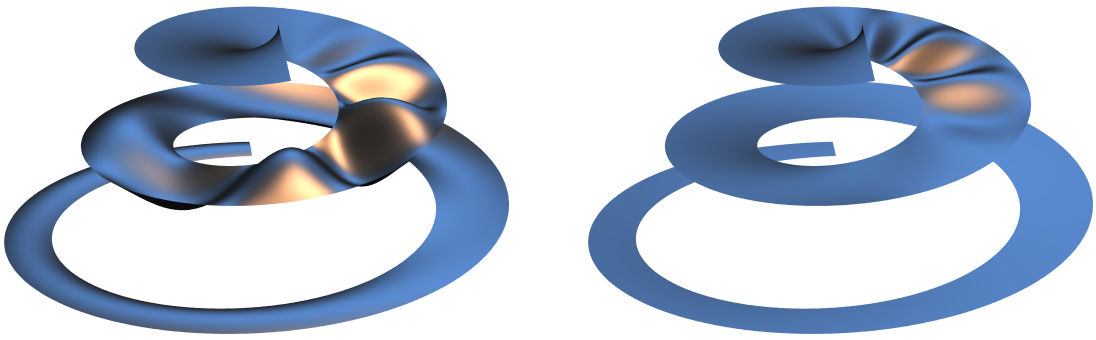


Figure 14.4: The traveling waves shown in Figure 14.3 are here mapped onto a 3D model of the basilar membrane, greatly exaggerated and stylized with colored lights. The active case with 20 dB more gain (right) is rendered for a 30 dB lower input level, so it represents the response on the same scale with a factor of 1000 less input power, corresponding to a cube-root-compressive system (10 dB output level change for 30 dB input level change).

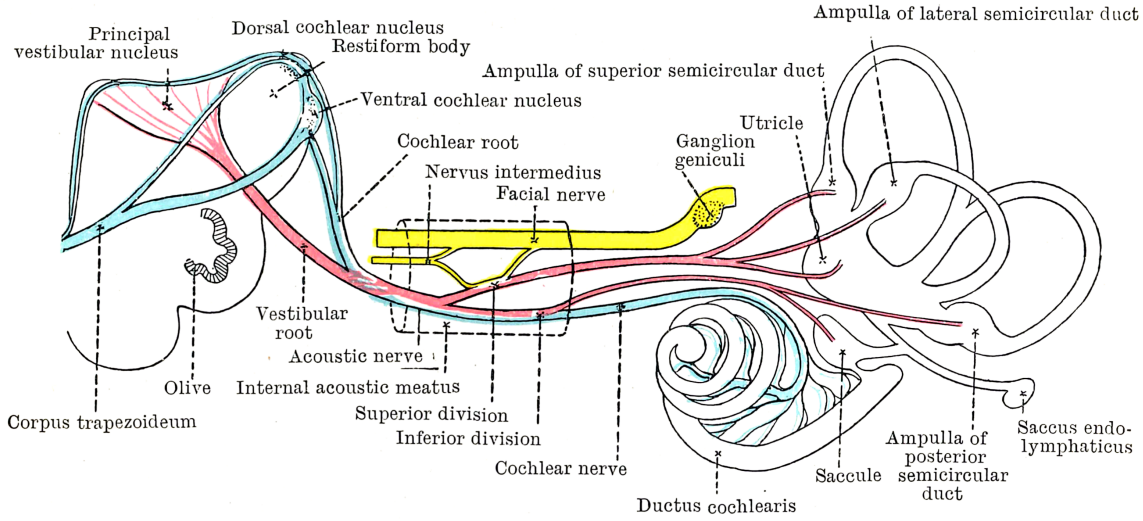


FIG. 587.—SCHEME OF THE ORIGIN AND DISTRIBUTION OF THE ACOUSTIC NERVE.

Figure 20.1: As shown in this color illustration from Cunningham and Robinson (1918), the acoustic nerve, or eighth cranial nerve, includes the cochlear division (blue) that serves hearing, and the vestibular division (red) that serves balance functions. After a stop at the dorsal and ventral divisions of the cochlear nucleus, the auditory pathway branches into the three acoustic stria, one of which, the ventral acoustic stria (the lower one here) goes to the superior olive on both sides, crossing via the trapezoid body. The facial nerve (yellow) takes efferent signals back to the stapedius muscle in the inner ear, via the geniculate ganglion, to serve the protective acoustic reflex.

Bibliography

Abbott, E. A. (**1884**). *Flatland: A Romance of Many Dimensions* (London: Seeley & Co.).

Abramowitz, M. and Stegun, I. A. (**1972**). *Handbook of Mathematical Functions* (Dover).

Aertsen, A. M. H. J. and Johannesma, P. I. M. (**1980**). "Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. Characterization of tonal and natural stimuli," *Biological Cybernetics* **38**, 223–234.

Aggazzotti, A. (**1921**). "Sulla percezione della direzione del suono," *Archivio di Fisiologia* **19**, 33–46.

Ahmed, N., Natarajan, T., and Rao, K. R. (**1974**). "Discrete cosine transfom," *IEEE Transactions on Computers* **100**, 90–93.

Algazi, V. R., Avendano, C., and Duda, R. O. (**2001**a). "Elevation localization and head-related transfer function analysis at low frequencies," *Journal of the Acoustical Society of America* **109**, 1110–1122.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**b). "The CIPIC HRTF database," in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99–102 (IEEE).

Alinaghi, A., Wang, W., and Jackson, P. J. B. (**2013**). "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 684–688.

Allen, J. B. (**1979**). "Cochlear models—1978," in *Models of the Auditory System and Related Signal Processing Techniques: Scandinavian Audiology, Supplementum 9*, edited by B. Hoke and E. de Boer, 1–16 (Almquist & Wiksell).

Allen, J. B. (**1981**). "Cochlear modeling—1980," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, 766–769.

Allen, J. B. (**1983**). "A hair cell model of neural response," in *Mechanics of Hearing*, edited by E. de Boer and M. A. Viergever (Delft University Press).

Allen, J. B. (**2001**). "Nonlinear cochlear signal processing," in *Physiology of the Ear*, edited by A. F. Jahn and J. Santos-Sacchi, 2nd edition, 393–442 (Singular, Thomson Learning).

Allen, J. B. and Fahey, P. F. (**1992**). "Using acoustic distortion products to measure the cochlear amplifier gain on the basilar membrane," *Journal of the Acoustical Society of America* **92**, 178–188.

Allen, J. B. and Neely, S. T. (**1997**). "Modeling the relation between the intensity just-noticeable difference and loudness for pure tones and wideband noise," *Journal of the Acoustical Society of America* **102**, 3628–3646.

Altman, J. A. and Viskov, O. V. (**1977**). "Discrimination of perceived movement velocity for fused auditory image in dichotic stimulation," *Journal of the Acoustical Society of America* **61**, 816–819.

Ambikairajah, E., Black, N. D., and Linggard, R. (**1989**). "Digital filter simulation of the basilar membrane," *Computer Speech and Language* **3**, 105–118.

Andén, J. and Mallat, S. (**2011**). "Multiscale scattering for audio classification," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*.

Andén, J. and Mallat, S. (**2014**). "Deep scattering spectrum," *IEEE Transactions on Signal Processing* **62**, 4114–4128.

Angelo, E. J., Jr. and Papoulis, A. (**1964**). *Pole-Zero Patterns: In the Analysis and Design of Low-order Systems* (McGraw-Hill).

ANSI, ed. (**1960**). *S1.1-1960 Acoustical Terminology* (American National Standards Institute).

Asano, F., Suzuki, Y., Sone, T., Kakehata, S., Satake, M., Ohyama, K., Kobayashi, T., and Takasaka, T. (**1991**). "A digital hearing aid that compensates loudness for sensorineural impaired listeners," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3625–3628.

Ashmore, J. (**2008**). "Cochlear outer hair cell motility," *Physiological Reviews* **88**, 173.

Assmann, P. F. and Katz, W. F. (**2000**). "Time-varying spectral change in the vowels of children and adults," *Journal of the Acoustical Society of America* **108**, 1856–1866.

Assmann, P. F. and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," *Journal of the Acoustical Society of America* **85**, 327–338.

Assmann, P. F. and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *Journal of the Acoustical Society of America* **88**, 680–697.

Assmann, P. F. and Summerfield, Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, 231–308 (Springer).

Atal, B. S. and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America* **50**, 637–655.

Atame, S. and Therese, S. (**2015**). "Singer's voice identification and authentication based on GFCC using k-means clustering and DTW," *IOSR Journal of Computer Engineering* **17**, 86–91.

Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (**2008**). "Cooperative nonlinearities in auditory cortical neurons," *Neuron* **58**, 956.

Aydelott, J., Leech, R., and Crinion, J. (**2010**). "Normal adult aging and the contextual influences affecting speech and meaningful sound perception," *Trends in Amplification* **14**, 218–232.

Baker, R. J. and Rosen, S. (**2006**). "Auditory filter nonlinearity across frequency using simultaneous notched-noise masking," *Journal of the Acoustical Society of America* **119**, 454–462.

Baker, R. J., Rosen, S., and Darling, A. M. (**1998**). "An efficient characterisation of human auditory filtering across level and frequency that is also physiologically reasonable," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, 81–88 (Whurr).

Bale, G. G. P. (**1879**). *The Elements of the Anatomy and Physiology of Man*, student's edition (London: Remington and Co.).

Baluja, S. and Covell, M. (**2008**). "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition* **41**, 3467–3480.

Bar-Yam, Y. (**1997**). *Dynamics of Complex Systems* (Westview Press).

Barchiesi, D., Giannoulis, D., Stowell, D., and Plumbley, M. D. (**2015**). "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine* **32**, 16–34.

Barker, J., Cooke, M., and Ellis, D. P. W. (**2000**a). "Decoding speech in the presence of other sound sources," in *Sixth International Conference on Spoken Language Processing*.

Barker, J., Josifovski, L., Cooke, M., and Green, P. (**2000**b). "Soft decisions in missing data techniques for robust automatic speech recognition," in *Sixth International Conference on Spoken Language Processing*.

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (**2013**). "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language* **27**, 621–633.

Barrington, L., Chan, A., Turnbull, D., and Lanckriet, G. (**2007**). "Audio information retrieval using semantic similarity," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE).

Barth, A. (**1887**). "Report on the progress of otology during the latter half of 1886," *Archives of Otology* **16**, 154–167.

Barton, E. H. (**1908**). *A Text-Book on Sound* (London: Macmillan and Co.).

Baumgarte, F. (**1999**). "A physiological ear model for the emulation of masking," *Journal of Oto-Rhino-Laryngology* **61**, 294–304.

Baumgartner, R., Majdak, P., and Laback, B. (**2014**). "Modeling sound-source localization in sagittal planes for human listeners," *Journal of the Acoustical Society of America* **136**, 791–802.

Bean, M. A. (**2001**). *Probability: The Science of Uncertainty—with Applications to Investments, Insurance, and Engineering* (American Mathematical Society).

Beauvois, M. W. and Meddis, R. (**1991**). "A computer model of auditory stream segregation," *The Quarterly Journal of Experimental Psychology* **43**, 517–541.

Békésy, G. von (**1928**). "Zur Theorie des Hörens; die Schwingungsform der Basilarmembran," *Physik Zeits* **29**, 793–810.

Békésy, G. von (**1956**). "Current status of theories of hearing," *Science* **123**, 779–783.

Békésy, G. von (**1960**). "The pattern of vibrations in the cochlea," in *Experiments in Hearing*, edited by G. von Békésy and E. G. Wever, 403–484 (Krieger).

Békésy, G. von (**1967**). *Sensory Inhibition* (Princeton University Press).

Békésy, G. von (**1974**). "Some biophysical experiments from fifty years ago," *Annual Review of Physiology* **36**, 1–18.

Bell, C. G. and Newell, A. (**1971**). *Computer Structures: Readings and Examples* (McGraw-Hill).

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B. (**2005**). "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing* **13**, 1035–1047.

Bendor, D. and Wang, X. (**2006**). "Cortical representations of pitch in monkeys and humans," *Current Opinion in Neurobiology* **16**, 391–399.

Bennett, S. (**1993**). *A History of Control Engineering, 1930–1955* (Peter Peregrinus Ltd.).

Bergeaud, F. and Mallat, S. (**1995**). "Matching pursuit of images," in *Proceedings, International Conference on Image Processing*, volume 1.

Bertin-Mahieux, T. and Ellis, D. P. W. (**2011**). "Large-scale cover song recognition using hashed chroma landmarks," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*.

Beyer, R. T. (**1999**). *Sounds of Our Times: Two Hundred Years of Acoustics* (Springer).

Bianchi, F., Verhulst, S., and Dau, T. (**2013**). "Experimental evidence for a cochlear source of the precedence effect," *Journal of the Association for Research in Otolaryngology* 1–13.

Billington, D. P. and Billington, D. P., Jr. (**2013**). *Power, Speed, and Form: Engineers and the Making of the Twentieth Century* (Princeton University Press).

Billock, V. A. and Tsou, B. H. (**2011**). "To honor Fechner and obey Stevens: Relationships between psychophysical and neural nonlinearities," *Psychological Bulletin* **137**, 1.

Bilsen, F. A. and Ritsma, R. J. (**1970**). "Some parameters influencing the perceptibility of pitch," *Journal of the Acoustical Society of America* **47**, 469.

Bishop, C. M. (**1995**). *Neural Networks for Pattern Recognition* (Oxford University Press).

Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J., and Schnupp, J. W. H. (**2009**). "Interdependent encoding of pitch, timbre, and spatial location in auditory cortex," *The Journal of Neuroscience* **29**, 2064–2075.

Blackburn, C. C. and Sachs, M. B. (**1990**). "The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons," *Journal of Neurophysiology* **63**, 1191–1212.

Blauert, J. (**1997**). "An introduction to binaural technology," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Inc).

Bode, H. W. (**1945**). *Network Analysis and Feedback Amplifier Design* (D. Van Nostrand Co.).

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (**1963**). "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, 209–243.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**, 113–120.

Boney, L., Tewfik, A. H., and Hamdy, K. N. (**1996**). "Digital watermarks for audio signals," in *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, 473–480.

Borg, E. and Counter, S. A. (**1989**). "The middle-ear muscles," *Scientific American* **261**, 74–80.

Boring, E. G. (**1926**). "Auditory theory with special reference to intensity, volume, and localization," *The American Journal of Psychology* 157–188.

Bottou, L. and Bousquet, O. (**2008**). "The tradeoffs of large scale learning," *Advances in Neural Information Processing Systems* **20**, 161–168.

Bottou, L., Chapelle, O., Decoste, D., and Weston, J. (**2007**). *Large-Scale Kernel Machines* (MIT Press).

Bouguer, P. (**1760**). *Traité d'optique sur la gradation de la lumière* (Paris: H. L. Guerin & L. F. Delatour).

Bouvrie, J., Rosasco, L., and Poggio, T. (**2009**). "On invariance in hierarchical models," *Advances in Neural Information Processing Systems* **22**, 162–170.

Bowlker, T. J. (**1908**). "On the factors serving to determine the direction of sound," *Philosophical Magazine* **15**, 318–332.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press).

Bregman, A. S. and Campbell, J. (**1971**). "Primary auditory stream segregation and perception of order in rapid sequences of tones," *Journal of Experimental Psychology* **89**, 244.

Bregman, A. S. and Pinker, S. (**1978**). "Auditory streaming and the building of timbre.," *Canadian Journal of Psychology* **32**, 19–31.

Breschet, G. (**1836**). *Recherches anatomiques et physiologiques sur l'organe de l'ouïe et sur l'audition dans l'homme et les animaux vertébrés* (Paris: J. B. Baillière).

Bridle, J. S. (**1990**). "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, 227–236 (Springer).

Bridle, J. S. and Brown, M. D. (**1974**). "An experimental automatic word recognition system," Technical Report, Joint Speech Research Unit.

Brin, D. (**1998**). *The Transparent Society* (Perseus Books).

Britannica (**1797**). "Logarithms," in *Encyclopædia Britannica: or, a dictionary of arts, sciences, and miscellaneous literature*, edited by C. Macfarquhar and G. Gleig, volume 10, Part 1, 3rd edition (Edinburgh: A. Bell and C. Macfarquhar).

Britannica (**1911**). "Waves," in *The Encyclopædia Britannica: A dictionary of arts, sciences, literature and general information*, edited by H. Chisholm, volume 28, 427 (New York: Encyclopædia Britannica).

Broadbent, D. E. (**1958**). *Perception and Communication* (Oxford University Press).

Brown, A. M. (**1993**). "Distortion in the cochlea: Acoustic $f_2 - f_1$ at low stimulus levels," *Hearing Research* **70**, 160–166.

Brown, C. P. and Duda, R. O. (**1998**). "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing* **6**, 476–488.

Brown, G. J. and Cooke, M. (**1994**). "Computational auditory scene analysis," *Computer Speech and Language* **8**, 297–336.

Brownell, W. E., Manis, P. B., and Ritz, L. A. (**1979**). "Ipsilateral inhibitory responses in the cat lateral superior olive," *Brain Research* **177**, 189–193.

Bruce, R. V. (**1990**). *Bell: Alexander Graham Bell and the Conquest of Solitude* (Cornell University Press).

Brugge, J. F. (**1992**). "An overview of central auditory processing," in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay, 1–33 (Springer).

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation," *Journal of the Acoustical Society of America* **120**, 4007.

Burkard, R. F., Eggermont, J. J., and Do, M. (**2007**). *Auditory Evoked Potentials: Basic Principles and Clinical Application* (Lippincott Williams and Wilkins).

Cahan, D. (**1993**). *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science* (University of California Press).

Cajal, S. Ramón y (**1909**). *Histologie du système nerveux de l'homme & des vertébrés* (Paris: A. Maloine).

Caldwell, W. F., Glaesser, E., and Stewart, J. L. (**1962**). "Theory and design of an analog ear," in *Biological Prototypes and Synthetic Systems*, edited by E. E. Bernard and M. R. Kare, 97–103 (Plenum Press).

Campbell, G. A. (**1922**). "Physical theory of the electric wave filter," *Bell System Technical Journal* **1**, 1–32.

Capranica, R. R. (**1992**). "The untuning of the tuning curve: Is it time?," in *Seminars in Neuroscience*, volume 4, 401–408 (Elsevier).

Cariani, P. (**1994**). "As if time really mattered: Temporal strategies for neural coding of sensory information," in *Origins: Brain and Self-Organization*, edited by K. H. Pribram, 209–252 (Lawrence Erlbaum Associates).

Cariani, P. (**1999**). "Temporal coding of periodicity pitch in the auditory system: An overview," *Neural Plasticity* **6**, 147–172.

Cariani, P. A. and Delgutte, B. (**1996**a). "Neural correlates of the pitch of complex tones. I. pitch and pitch salience," *Journal of Neurophysiology* **76**, 1698–1716.

Cariani, P. A. and Delgutte, B. (**1996**b). "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch," *Journal of Neurophysiology* **76**, 1717–1734.

Carney, L. H. (**1993**). "A model for the responses of low-frequency auditory-nerve fibers in cat," *Journal of the Acoustical Society of America* **93**, 401–417.

Carney, L. H., McDuffy, M. J., and Shekhter, I. (**1999**). "Frequency glides in the impulse responses of auditory-nerve fibers," *Journal of the Acoustical Society of America* **105**, 2384–2391.

Casey, M., Rhodes, C., and Slaney, M. (**2008**a). "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Transactions on Audio, Speech, and Language Processing* **16**, 1015–1028.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (**2008**b). "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE* **96**, 668–696.

Casseday, J. H. and Covey, E. (**1996**). "A neuroethological theory of the operation of the inferior colliculus," *Brain, Behavior and Evolution* **47**, 311–322.

Casson, H. N. (**1910**). *The History of the Telephone* (Chicago: A. C. McClurg & Co.).

Cedolin, L. and Delgutte, B. (**2005**). "Pitch of complex tones: Rate–place and interspike interval representations in the auditory nerve," *Journal of Neurophysiology* **94**, 347–362.

Chalupper, J. and Fastl, H. (**2002**). "Dynamic loudness model (DLM) for normal and hearing-impaired listeners," *Acta Acustica United with Acustica* **88**, 378–386.

Cheatham, M. A. and Dallos, P. (**2000**). "The dynamic range of inner hair cell and organ of Corti responses," *Journal of the Acoustical Society of America* **107**, 1508–1520.

Chechik, G., Anderson, M. J., Bar-Yosef, O., Young, E. D., Tishby, N., and Nelken, I. (**2006**). "Reduction of information redundancy in the ascending auditory pathway," *Neuron* **51**, 359–368.

Chechik, G., Ie, E., Rehn, M., Bengio, S., and Lyon, R. F. (**2008**). "Large-scale content-based audio retrieval from text queries," in *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, 105–112 (ACM).

Chechik, G. and Nelken, I. (**2012**). "Auditory abstraction from spectro-temporal features to coding auditory entities," *Proceedings of the National Academy of Sciences* **109**, 18968–18973.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America* **25**, 975–979.

Chowdhury, S. A. and Suga, N. (**2000**). "Reorganization of the frequency map of the auditory cortex evoked by cortical electrical stimulation in the big brown bat," *Journal of Neurophysiology* **83**, 1856–1863.

Chung, D.-O., Doh, W., Youn, D.-H., Choi, J.-Y., Woo, H.-C., Kim, D.-W., and Kim, W.-K. (**1996**). "Hearing impairment simulation for the performance evaluation of hearing aid system," in *Engineering in Medicine and Biology Society, 1996: Bridging Disciplines for Biomedicine*, volume 1, 415–416.

Clarkson, B., Sawhney, N., and Pentland, A. (**1998**). "Auditory context awareness via wearable computing," *Energy* **400**, 20.

Clifton, R. K. and Freyman, R. L. (**1997**). "The precedence effect: Beyond echo suppression," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson, 233–256 (Lawrence Erlbaum Associates, Inc).

Cohen, J. R. (**1989**). "Application of an auditory model to speech recognition," *Journal of the Acoustical Society of America* **85**, 2623–2629.

Colburn, H. S. (**1996**). "Computational models of binaural processing," in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, 332–400 (Springer).

Collins, N. (**2005**). "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *AES Convention*, volume 118.

Collobert, R. and Bengio, S. (**2004**). "Links between perceptrons, MLPs and SVMs," in *Proceedings of the Twenty-first International Conference on Machine Learning*, 23 (ACM).

Cook, D. L., Schwindt, P. C., Grande, L. A., and Spain, W. J. (**2003**). "Synaptic depression in the localization of sound," *Nature* **421**, 66–70.

Cook, N. D. (**1986**). *The Brain Code: Mechanisms of Information Transfer and the Role of the Corpus Callosum* (Routledge).

Cook, P. R. (**2001**). *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics* (MIT Press).

Cooke, M. (**1993**). *Modelling Auditory Processing and Organisation* (Cambridge University Press).

Cooke, M. and Ellis, D. P. W. (**2001**). "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication* **35**, 141–177.

Cooke, M., Morris, A., and Green, P. (**1997**). "Missing data techniques for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, 863–866.

Cooper, N. P. and Kemp, D. T. (**2009**). *Concepts and Challenges in the Biophysics of Hearing* (World Scientific Publishing Company).

Cooper, N. P., Pickles, J. O., and Manley, G. A. (**2008**). "Traveling waves, second filters, and physiological vulnerability: A short history of the discovery of active processes in hearing," in *Active Processes and Otoacoustic Emissions in Hearing*, edited by G. A. Manley, R. R. Fay, and A. N. Popper, 39–62 (Springer).

Cooper, N. P. and Rhode, W. S. (**1997**). "Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea," *Journal of Neurophysiology* **78**, 261–270.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (**1996**). "On the Lambert W function," *Advances in Computational Mathematics* **5**, 329–359.

Cortes, C. and Vapnik, V. (**1995**). "Support-vector networks," *Machine Learning* **20**, 273–297.

Cover, T. M. (**1965**). "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers* 326–334.

Cramer, E. M. and Huggins, W. H. (**1958**). "Creation of pitch through binaural interaction," *Journal of the Acoustical Society of America* **30**, 413–417.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (**2006**). "Online passive-aggressive algorithms," *Journal of Machine Learning Research (JMLR)* **7**, 551–585.

Cunningham, D. J. and Robinson, A. (**1918**). *Cunningham's Text-Book of Anatomy* (New York: William Wood and Company).

Dallos, P. (**1973**). *The Auditory Periphery: Biophysics and Physiology* (Academic Press).

Dallos, P. (**1992**). "The active cochlea," *The Journal of Neuroscience* **2**, 4575–4585.

Dallos, P. (**2003**). "Some pending problems in cochlear mechanics," in *Biophysics of the Cochlea: from Molecules to Models*, edited by A. W. Gummer, E. Dalhoff, M. Nowotny, and M. P. Scherer, 97–105 (World Scientific).

Darling, A. M. (**1991**). "Properties and implementation of the GammaTone filter: A tutorial," *Speech, Hearing, and Language: Work in Progress* **5**, 43–61.

Darrow, K. N., Maison, S. F., and Liberman, M. C. (**2006**). "Cochlear efferent feedback balances interaural sensitivity," *Nature Neuroscience* **9**, 1474–1476.

Darwin, C. J. and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," *Journal of the Acoustical Society of America* **107**, 970–977.

Davenport, W. B. (**1953**). "Signal-to-noise ratios in band-pass limiters," *Journal of Applied Physics* **24**, 720–727.

David, E. E., Jr. (**1958**). "Artificial auditory recognition in telephony," *IBM Journal of Research and Development* **2**, 294–309.

Davis, H. (**1957**). "Biophysics and physiology of the inner ear," *Physiological Reviews* **37**, 1–49.

Davis, H. (**1965**). "A model for transducer action in the cochlea," in *Cold Spring Harbor Symposia on Quantitative Biology*, volume 30, 181–190 (Cold Spring Harbor Laboratory Press).

Davis, H. (**1983**). "An active process in cochlear mechanics," *Hearing Research* **9**, 79.

Dawson, J. L. and Lee, T. H. (**2004**). *Feedback Linearization of RF Power Amplifiers* (Springer).

de Boer, E. (**1956**). "On the 'residue' in hearing," Ph.D., Universiteit van Amsterdam.

de Boer, E. (**1976**a). "Cross-correlation function of a bandpass nonlinear network," *Proceedings of the IEEE* **64**, 1443–1444.

de Boer, E. (**1976**b). "On the 'residue' and auditory pitch perception," in *Handbook of Sensory Physiology: Vol. V Part 3: Auditory System*, edited by W. D. Keidel and W. D. Neff, 479–583 (Springer).

de Boer, E. (**1997**). "Cochlear models and minimum phase," *Journal of the Acoustical Society of America* **102**, 3810–3813.

de Boer, E. and de Jongh, H. R. (**1978**). "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique," *Journal of the Acoustical Society of America* **63**, 115–135.

de Boer, E. and Kruidenier, C. (**1990**). "On ringing limits of the auditory periphery," *Biological Cybernetics* **63**, 433–442.

de Boer, E. and Nuttall, A. L. (**1997**). "The mechanical waveform of the basilar membrane. I. Frequency modulations ('glides') in impulse responses and cross-correlation functions," *Journal of the Acoustical Society of America* **101**, 3583.

de Boer, E., Nuttall, A. L., Hu, N., Zou, Y., and Zheng, J. (**2005**). "The Allen–Fahey experiment extended," *Journal of the Acoustical Society of America* **117**, 1260–1266.

de Boer, E. and Viergever, M. A. (**1982**). "Validity of the Liouville–Green (or WKB) method for cochlear mechanics," *Hearing Research* **8**, 131.

De Forest, L. (**1942**). *Television, Today and Tomorrow* (The Dial Press).

Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S., and Yagnik, J. (**2013**). "Fast, accurate detection of 100,000 object classes on a single machine," in *IEEE Conference on Computer Vision and Pattern Recognition*.

Delgutte, B. (**1996**). "Physiological models for basic auditory percepts," in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, 157–220 (Springer).

Delgutte, B. (**1997**). "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver, 507–538 (Blackwell).

Delgutte, B. and Cariani, P. (**1992**). "Coding of the pitch of harmonic and inharmonic complex tones in the interspike intervals of auditory-nerve fibers," in *The Auditory Processing of Speech: From Sounds to Words*, edited by M. E. Schouten, 37–45 (Walter de Gruyter).

Deng, L. and O'Shaughnessy, D. (**2003**). *Speech Processing: A Dynamic and Optimization-Oriented Approach* (CRC Press).

Denny, M. (**2007**). *Blip, Ping, and Buzz: Making Sense of Radar and Sonar* (Johns Hopkins University Press).

Dietz, M., Marquardt, T., Salminen, N. H., and McAlpine, D. (**2013**). "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds," *Proceedings of the National Academy of Sciences* **110**, 15151–15156.

Dietz, M., Marquardt, T., Stange, A., Pecka, M., Grothe, B., and McAlpine, D. (**2014**). "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds II: Single-neuron recordings," *Journal of Neurophysiology* **111**, 1973–1985.

Dinther, R. van and Patterson, R. D. (**2006**). "Perception of acoustic scale and size in musical instrument sounds," *Journal of the Acoustical Society of America* **120**, 2158–2176.

Divenyi, P. (**2005**). *Speech Separation by Humans and Machines* (Kluwer Academic Publishers).

Dong, D. W. and Atick, J. J. (**1995**). "Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Network: Computation in Neural Systems* **6**, 159–178.

Dorf, R. C. (**1974**). *Modern Control Systems* (Addison-Wesley).

Draper, J. C. (**1883**). *A Text-Book on Anatomy, Physiology, and Hygiene*, 6th edition (New York: Harper and Brothers).

Drygajlo, A. and El-Maliki, M. (**1998**). "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 121–124.

Drysdale, C. V. (**1920**). "The eleventh Kelvin lecture: Modern marine problems in war and peace," *Journal of the Institution of Electrical Engineers* **58**, 572–597.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (**2008**). "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 272–279 (ACM).

Duda, R. O., Avendano, C., and Algazi, V. R. (**1999**). "An adaptable ellipsoidal head model for the interaural time difference," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 965–968.

Duda, R. O. and Hart, P. E. (**1973**). *Pattem Classification and Scene Analysis* (John Wiley & Sons).

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattem Classification*, 2nd edition (Wiley-Interscience).

Duda, R. O., Lyon, R. F., and Slaney, M. (**1990**). "Correlograms and the separation of sounds," in *Twenty-Fourth Asilomar Conference on Signals, Systems, and Computers*, 457–461.

Duifhuis, H. (**1976**). "Cochlear non-linearity and second filter: Possible mechanism and implications," *Journal of the Acoustical Society of America* **59**, 408–423.

Duifhuis, H. (**1989**). "Power-law nonlinearities: A review of some less familiar properties," in *Cochlear Mechanisms: Structure, Function, and Models*, 395–403 (Springer).

Duifhuis, H. (**1992**). "Cochlear modelling and physiology," in *The Auditory Processing of Speech: From Sounds to Words*, edited by M. E. Schouten, 15–27 (Walter de Gruyter).

Duifhuis, H. (**2004**). "Comment on 'An approximate transfer function for the dual-resonance nonlinear filter model of auditory frequency selectivity' [*J. Acoust. Soc. Am.* 114, 2112–2117] (L)," *Journal of the Acoustical Society of America* **115**, 1889–1890.

Duifhuis, H. (**2011**). "Hopf-bifurcations and Van der Pol oscillator models of the mammalian cochlea," in *What Fire is in Mine Ears—Progress in Auditory Biomechanics: Proceedings of the 11th International Mechanics of Hearing Workshop*, edited by C. A. Shera and E. S. Olson, 199–205 (AIP).

Duifhuis, H. (**2012**). *Cochlear Mechanics: Introduction to a Time Domain Analysis of the Nonlinear Cochlea* (Springer).

Duifhuis, H. and van de Vorst, J. J. W. (**1980**). "Mechanics and nonlinearity of hair cell stimulation," *Hearing Research* **2**, 493–504.

Duke, T. and Jülicher, F. (**2003**). "Active traveling wave in the cochlea," *Physical Review Letters* **90**, 158101.

Duke, T. A. J. and Jülicher, F. (**2008**). "Critical oscillators as active elements in hearing," in *Active Processes and Otoacoustic Emissions in Hearing*, edited by G. A. Manley, R. R. Fay, and A. N. Popper, 63–92 (Springer).

Dunlap, K. (**1916**). "Tonal volume and pitch," *Journal of Experimental Psychology* **1**, 183.

Dunne, R. A. (**2007**). *A Statistical Approach to Neural Networks for Pattern Recognition* (Wiley-Interscience).

Duverney, G. J. (**1683**). *Traité de l'organe de l'ouïe, contenant la structure, les usages & les maladies de toutes les parties de l'oreille* (Paris: chez Estienne Michallet).

Eaglesfield, C. C. (**1945**). "Carrier frequency amplifiers: The unit step response of amplifiers with single and double circuits," *Wireless Engineer* **22**, 523–532.

Eargle, J. M. (**1994**). *Electroacoustical Reference Data* (Van Nostrand Reinhold).

Edwards, B. (**2007**). "The future of hearing aid technology," *Trends in Amplification* **11**, 31–46.

Eggermont, J. J. (**1993**). "Wiener and Volterra analyses applied to the auditory system," *Hearing Research* **66**, 177–201.

Eggermont, J. J., Aertsen, A., and Johannesma, P. I. M. (**1983**). "Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field," *Hearing Research* **10**, 167–190.

Eguíluz, V. M., Ospeck, M., Choe, Y., Hudspeth, A. J., and Magnasco, M. O. (**2000**). "Essential nonlinearities in hearing," *Physical Review Letters* **84**, 5232–5235.

Ehret, G. and Merzenich, M. M. (**1985**). "Auditory midbrain responses parallel spectral integration phenomena," *Science* **227**, 1245–1247.

Elderton, W. P. and Johnson, N. L. (**1969**). *Systems of Frequency Curves* (Cambridge University Press).

Ellis, D. P. W. (**1997**). "The weft: A representation for periodic sounds," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, 1307–1310.

Ellis, D. P. W. and Cotton, C. (**2007**). "The 2007 LabROSA cover song detection system," *Music Information Retrieval Evaluation eXchange (MIREX)* URL `http://labrosa.ee.columbia.edu/~dpwe/pubs/EllisC07-covers.pdf`.

Ellis, D. P. W. and Lee, K. (**2004**). "Minimal-impact audio-based personal archives," in *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 39–47.

Ellis, D. P. W. and Rosenthal, D. F. (**1998**). "Midlevel representations for computational auditory scene analysis: The weft element," in *Computational Auditory Scene Analysis*, edited by D. F. Rosenthal and H. G. Okuno, 257–272 (Lawrence Erlbaum Associates).

Elmore, W. C. and Heald, M. A. (**1969**). *Physics of Waves* (Courier Dover Publications).

Epp, B., Verhey, J. L., and Mauermann, M. (**2010**). "Modeling cochlear dynamics: Interrelation between cochlea mechanics and psychoacoustics," *Journal of the Acoustical Society of America* **128**, 1870–1883.

Eustaquio-Martín, A. and Lopez-Poveda, E. A. (**2011**). "Isoresponse versus isoinput estimates of cochlear filter tuning," *Journal of the Association for Research in Otolaryngology* **12**, 281–299.

Evans, E. F. (**1980**). "An electronic analogue of single unit recording from the cochlear nerve for teaching and research [proceedings]," *The Journal of Physiology* **298**, 6P.

Evans, E. F. (**1989**). "Cochlear filtering: A view seen through the temporal discharge patterns of single cochlear nerve fibres," in *Cochlear Mechanisms: Structure, Function, and Models*, 241–250 (Springer).

Evans, E. F. and Wilson, J. P. (**1973**). "The frequency selectivity of the cochlea," in *Basic Mechanisms in Hearing*, edited by A. R. Moller, 519–551 (Academic Press).

Evgeniou, T., Pontil, M., and Poggio, T. (**2000**). "Regularization networks and support vector machines," *Advances in Computational Mathematics* **13**, 1–50.

Ewald, J. R. (**1899**). "Zur Physiologie des Labyrinths VI. Mittheilung. Eine neue Hörtheorie," *Pflügers Archiv European Journal of Physiology* **76**, 147–188.

Faller, C. and Merimaa, J. (**2004**). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America* **116**, 3075–3089.

Ferry, E. S. (**1921**). *General Physics and its Application to Industry and Everyday Life* (John Wiley & Sons).

Fettiplace, R. and Kim, K. X. (**2014**). "The physiology of mechanoelectrical transduction channels in hearing," *Physiological Reviews* **94**, 951–986.

Fine, H. B. (**1903**). *The Number-System of Algebra: Treated Theoretically and Historically*, 2nd edition (Boston: Heath).

Flanagan, J. L. (**1960**). "Models for approximating basilar membrane displacement," *Bell System Technical Journal* **39**, 1163–1191.

Flanagan, J. L. (**1962**). "Models for approximating basilar membrane displacement—Part II. Effects of middle-ear transmission and some relations between subjective and physiological behavior," *Bell System Technical Journal* **41**, 959–1009.

Flanagan, J. L. and Guttman, N. (**1960**). "On the pitch of periodic pulses," *Journal of the Acoustical Society of America* **32**, 1308–1319.

Fletcher, H. (**1922**). "The nature of speech and its interpretation," *Bell System Technical Journal* **1**, 129–144.

Fletcher, H. (**1924**). "The physical criterion for determining the pitch of a musical tone," *Physical Review* **23**, 427–437.

Fletcher, H. (**1930**). "A space–time pattern theory of hearing," *Journal of the Acoustical Society of America* **1**, 311–343.

Fletcher, H. (**1940**). "Auditory patterns," *Reviews of Modern Physics* **12**, 47–65.

Fletcher, H. and Munson, W. A. (**1933**). "Loudness, its definition, measurement, and calculation," *Journal of the Acoustical Society of America* **5**, 82–108.

Fletcher, H. and Munson, W. A. (**1937**). "Relation between loudness and masking," *Journal of the Acoustical Society of America* **9**, 1–10.

Formby, C. (**1990**). "Simple triangular approximations of auditory filter shapes," *Journal of Speech, Language and Hearing Research* **33**, 530.

Fornasini, P. (**2008**). *The Uncertainty in Physical Measurements: An Introduction to Data Analysis in the Physics Laboratory* (Springer).

Foster, N. E. V. and Zatorre, R. J. (**2010**). "Cortical structure predicts success in performing musical transformation judgments," *NeuroImage* **53**, 26–36.

Fourier, J. (**1822**). *Theorie Analytique de la Chaleur* (Paris: Firmin Didot).

Fowler, E. P. (**1936**). "A method for the early detection of otosclerosis: a study of sounds well above threshold," *Archives of Otolaryngology—Head and Neck Surgery* **24**, 731–741.

Frasconi, P., Gori, M., and Tesi, A. (**1997**). "Successes and failures of backpropagation: A theoretical investigation," in *Progress in Neural Networks: Architecture*, edited by O. M. Omidvar and C. L. Wilson, volume 5, 205–242 (Ablex Publishing).

Freeman, L. (**1948**). *Physiological Psychology* (D. Van Nostrand Co.).

Friedman, D. H. (**1990**). "Implementation of a nonlinear wave-digital-filter cochlear model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 397–400.

Fröman, N. and Fröman, P. O. (**2002**). *Physical Problems Solved by the Phase-Integral Method* (Cambridge University Press).

Frost, J. (**1838**). *The Class Book of Nature: Comprising Lessons on the Universe, the Three Kingdoms of Nature, and the Form and Structure of the Human Body*, 3rd edition (Hartford: Belknap & Hamersley).

Galambos, R. and Davis, H. (**1943**). "The response of single auditory-nerve fibers to acoustic stimulation," *Journal of Neurophysiology* **6**, 39–57.

Galison, P. (**1997**). *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press).

Gargi, U. and Yagnik, J. (**2008**). "Solving the label resolution problem in supervised video content classification," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 276–282 (ACM).

Garrison, F. H. (**1914**). *An Introduction to the History of Medicine* (Philadelphia: W. B. Saunders and Co.).

Geisler, C. D. (**1998**). *From Sound to Synapse: Physiology of the Mammalian Ear* (Oxford University Press).

Geisler, C. D. and Greenberg, S. (**1986**). "A two-stage nonlinear cochlear model possesses automatic gain control," *Journal of the Acoustical Society of America* **80**, 1359.

Geisler, C. D., Yates, G. K., Patuzzi, R. B., and Johnstone, B. M. (**1990**). "Saturation of outer hair cell receptor currents causes two-tone suppression," *Hearing Research* **44**, 241–256.

Gelfand, S. A. (**1990**). *Hearing: An Introduction to Psychological and Physiological Acoustics*, 2nd edition (M. Dekker).

Gelfand, S. A. (**2004**). *Hearing: An Introduction to Psychological and Physiological Acoustics*, 4th edition (Marcel Dekker).

Gelfand, S. A. and Hochberg, I. (**1976**). "Binaural and monaural speech discrimination under reverberation," *International Journal of Audiology* **15**, 72–84.

Gersho, A. and Gray, R. M. (**1992**). *Vector Quantization and Signal Compression* (Springer).

Giguère, C. and Woodland, P. C. (**1993**). "A wave digital filter model of the entire auditory periphery," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, 708–711.

Gillespie, P. G. and Müller, U. (**2009**). "Mechanotransduction by hair cells: Models, molecules, and mechanisms," *Cell* **139**, 33–44.

Gish, H. (**1990**). "A probabilistic approach to the understanding and training of neural network classifiers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1361–1364.

Glasberg, B. R. (**1982**). *Introduction to the Psychology of Hearing*, 2nd edition (Academic Press).

Glasberg, B. R. and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched noise data," *Hearing Research* **47**, 103–138.

Glasberg, B. R. and Moore, B. C. J. (**2000**). "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise," *Journal of the Acoustical Society of America* **108**, 2318–2328.

Glasberg, B. R., Moore, B. C. J., Patterson, R. D., and Nimmo-Smith, I. (**1984**). "Dynamic range and asymmetry of the auditory filter," *Journal of the Acoustical Society of America* **76**, 419–427.

Glorot, X., Bordes, A., and Bengio, Y. (**2011**). "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, 315–323.

Gold, B. and Morgan, N. (**2000**). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (John Wiley & Sons).

Gold, B., Morgan, N., and Ellis, D. (**2011**). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2nd edition (Wiley-Interscience).

Gold, B. and Rader, C. M. (**1969**). *Digital Processing of Signals* (McGraw-Hill).

Gold, T. (**1948**). "Hearing. II. The physical basis of the action of the cochlea," *Proceedings of the Royal Society of London. Series B, Biological Sciences* **135**, 492–498.

Gold, T. and Pumphrey, R. J. (**1948**). "Hearing. I. The cochlea as a frequency analyzer," *Proceedings of the Royal Society of London. Series B, Biological Sciences* **135**, 462–491.

Golding, N. L. and Oertel, D. (**2012**). "Synaptic integration in dendrites: Exceptional need for speed," *The Journal of Physiology* **590**, 5563–5569.

Golding, N. L., Robertson, D., and Oertel, D. (**1995**). "Recordings from slices indicate that octopus cells of the cochlear nucleus detect coincident firing of auditory nerve fibers with temporal precision," *The Journal of Neuroscience* **15**, 3138–3153.

Goldstein, J. L. (**1967**). "Auditory nonlinearity," *Journal of the Acoustical Society of America* **41**, 676–689.

Goldstein, J. L. (**1990**). "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing Research* **49**, 39–60.

Goldstein, J. L. (**1995**). "Relations among compression, suppression, and combination tones in mechanical responses of the basilar membrane: Data and MBPNL model," *Hearing Research* **89**, 52–68.

Goldstein, J. L., Baer, T., and Kiang, N. Y. S. (**1971**). "A theoretical treatment of latency, group delay, and tuning characteristics for auditory-nerve responses to clicks and tones," in *Physiology of the Auditory System: A Workshop*, edited by M. B. Sachs, 133–141 (National Educational Consultants).

Goldstein, J. L. and Kiang, N. Y. S. (**1968**). "Neural correlates of the aural combination tone $2f_1 - f_2$," *Proceedings of the IEEE* **56**, 981–992.

Grangier, D. and Bengio, S. (**2008**). "A discriminative kernel-based model to rank images from text queries," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **30**, 1371–1384.

Green, D. M. (**1958**). "Detection of complex auditory signals in noise, and the critical band concept," Technical Report No. 82, Electronic Defense Group, Depart of Electrical Engineering, University of Michigan.

Green, G. (**1837**). "The motion of waves in a variable canal of small depth and width," *Transactions of the Cambridge Philosophical Society* **6**, 457–462.

Green, P. and Wood, A. R. (**1986**). "A representational approach to knowledge-based acoustic-phonetic processing in speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1205–1208.

Green, P. D., Cooke, M. P., and Crawford, M. D. (**1995**). "Auditory scene analysis and hidden Markov model recognition of speech in noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 401–404.

Greenberg, S. (**1980**). "Temporal neural coding of pitch and vowel quality," Technical Report, UCLA Working Papers in Phonetics.

Greenberg, S., Geisler, C. D., and Deng, L. (**1986**). "Frequency selectivity of single cochlear-nerve fibers based on the temporal response pattern to two-tone signals," *Journal of the Acoustical Society of America* **79**, 1010–1019.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," *Journal of the Acoustical Society of America* **33**, 1344–1356.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," *Journal of the Acoustical Society of America* **87**, 2592–2605.

Griffiths, T. D., Uppenkamp, S., Johnsrude, I. S., Josephs, O., and Patterson, R. D. (**2001**). "Encoding of the temporal regularity of sound in the human brainstem," *Nature Neuroscience* **4**, 633–637.

Gross, C. G. (**2002**). "Genealogy of the 'grandmother cell'," *The Neuroscientist* **8**, 512–518.

Grothe, B. (**2003**). "New roles for synaptic inhibition in sound localization," *Nature Reviews Neuroscience* **4**, 540–550.

Grothe, B. and Koch, U. (**2011**). "Dynamics of binaural processing in the mammalian sound localization pathway—the role of $GABA_B$ receptors," *Hearing Research* **279**, 43–50.

Grothe, B., Pecka, M., and McAlpine, D. (**2010**). "Mechanisms of sound localization in mammals," *Physiological Reviews* **90**, 983–1012.

Guernsey, M. (**1922**). "A study of liminal sound intensities and the application of Weber's law to tones of different pitch," *The American Journal of Psychology* **33**, 554–569.

Guinan, J. J. (**2010**). "Physiology of the medial and lateral olivocochlear systems," in *Auditory and Vestibular Efferents*, edited by D. K. Ryugo, R. R. Fay, and A. N. Popper, 39–81 (Springer).

Guttman, N. and Flanagan, J. L. (**1964**). "Pitch of high-pass-filtered pulse trains," *Journal of the Acoustical Society of America* **36**, 757–765.

Haas, H. (**1951**). "Über den Einfluss eines Einfachechos auf die Horsamkeit von Sprache," *Acustica* **1**, 49–58.

Hachmeister, J. E. (**2003**). "An abbreviated history of the ear: From Renaissance to present," *The Yale Journal of Biology and Medicine* **76**, 81–86.

Hacker, P. (**1991**). "Seeing, representing and describing: An examination of David Marr's computational theory of vision," in *Investigating Psychology: Sciences of the Mind after Wittgenstein* (London: Routledge).

Hafter, E. R. (**1997**). "Binaural adaptation and the effectiveness of a stimulus beyond its onset," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Inc).

Hald, A. (**2005**). *A History of Probability and Statistics and Their Applications before 1750* (John Wiley & Sons).

Hall, D. A. and Plack, C. J. (**2009**). "Pitch processing sites in the human auditory brain," *Cerebral Cortex* **19**, 576–585.

Hall, J. L. (**1974**). "Two-tone distortion products in a nonlinear model of the basilar membrane," *Journal of the Acoustical Society of America* **56**, 1818–1828.

Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., and Rass, U. (**2005**). "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing* **2005**, 2915–2929.

Hamel, P. and Eck, D. (**2010**). "Learning features from music audio with deep belief networks.," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 339–344.

Hamilton, T. J., Jin, C., van Schaik, A., and Tapson, J. (**2008**). "An active 2-D silicon cochlea," *IEEE Transactions on Biomedical Circuits and Systems* **2**, 30–43.

Hamming, R. W. (**1998**). *Digital Filters*, 3rd edition (Courier Dover Publications).

Harte, J. M., Elliott, S. J., and Rice, H. J. (**2005**). "A comparison of various nonlinear models of cochlear compression," *Journal of the Acoustical Society of America* **117**, 3777–3786.

Hartley, R. V. L. and Fry, T. C. (**1922**). "The binaural location of complex sounds," *Bell System Technical Journal* **1**, 33–42.

Hartmann, W. M. (**1996**). "Pitch, periodicity, and auditory organization," *Journal of the Acoustical Society of America* **100**, 3491–3502.

Hartmann, W. M. (**1997**). "Listening in a room and the precedence effect," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Inc).

Hartmann, W. M. (**1998**). *Signals, Sound, and Sensation* (AIP Press).

Hartung, K. and Trahiotis, C. (**2001**). "Peripheral auditory processing and investigations of the 'precedence effect' which utilize successive transient stimuli," *Journal of the Acoustical Society of America* **110**, 1505–1513.

Hateren, J. H. van and Snippe, H. P. (**2001**). "Information theoretical evaluation of parametric models of gain control in blowfly photoreceptor cells," *Vision Research* **41**, 1851–1865.

Hausdorff, J. M. and Peng, C.-K. (**1996**). "Multiscaled randomness: A possible source of 1/f noise in biology," *Physical Review E* **54**, 2154–2157.

Hawkins, J. E. (**2001**). "Auditory physiological history: A surface view," in *Physiology of the Ear*, edited by A. F. Jahn and J. Santos-Sacchi, 2nd edition (Singular, Thomson Learning).

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *Journal of the Acoustical Society of America* **115**, 833–843.

Haykin, S. (**1994**). *Neural Networks: A Comprehensive Foundation* (Prentice Hall).

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America* **134**, 3029–3038.

Healy, R. D. and Huggins, W. H. (**1974**). "Double poles and the mysterious 't' factor," *IEEE Transactions on Education* **17**, 205.

Heaviside, O. (**1892**). "Contributions to the theory of the propagation of current in wires," in *Electrical papers, Volume 1*, 141–179 (London: Macmillan and Co.).

Hecht, S. (**1924**). "The visual discrimination of intensity and the Weber–Fechner law," *The Journal of General Physiology* **7**, 235–267.

Heeger, D. J. (**1991**). "Nonlinear model of neural responses in cat visual cortex," in *Computational Models of Visual Processing*, edited by M. Landy and J. A. Movshon, 119–133 (MIT Press).

Heeger, D. J. (**1992**). "Normalization of cell responses in cat striate cortex," *Visual Neuroscience* **9**, 181–197.

Heffner, R. S. and Heffner, H. E. (**1992**). "Visual factors in sound localization in mammals," *The Journal of Comparative Neurology* **317**, 219–232.

Heffner, R. S. and Masterton, R. B. (**1990**). "Sound localization in mammals: Brain-stem mechanisms," *Comparative Perception* **1**, 285–314.

Heinz, M. G., Issa, J. B., and Young, E. D. (**2005**). "Auditory-nerve rate responses are inconsistent with common hypotheses for the neural correlates of loudness recruitment," *Journal of the Association for Research in Otolaryngology* **6**, 91–105.

Heller, E. J. (**2013**). *Why You Hear What You Hear: An Experiential Approach to Sound, Music, and Psychoacoustics* (Princeton University Press).

Helmholtz, H. L. F. von (**1863**). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* (Braunschweig: F. Vieweg & Sohn).

Helmholtz, H. L. F. von (**1870**). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*, 3rd edition (Braunschweig: F. Vieweg & Sohn).

Helmholtz, H. L. F. von (**1878**). "The facts of perception," in *Selected Writings of Hermann Helmholtz*, edited by R. Kahl, 366–408 (Middletown, CT: Wesleyan University Press).

Henry, J. (**1851**). "On the limit of perceptibility of a direct and reflected sound," *Proceedings of the American Association for the Advancement of Science* **5**, 42–43.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America* **87**, 1738–1752.

Hermansky, H. and Morgan, N. (**1994**). "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing* **2**, 578–589.

Hermansky, H. and Pavel, M. (**1995**). "Psychophysics of speech engineering systems," in *Proc. 13th International Congress on Phonetic Sciences*, volume 3, 42–49.

Hermansky, H., Tibrewala, S., and Pavel, M. (**1996**). "Towards ASR on partially corrupted speech," in *International Conference on Spoken Language Processing*, volume 1, 462–465 (IEEE).

Herschel, J. F. W. (**1930**). *A Preliminary Discourse on the Study of Natural Philosophy* (London: Longmans).

Hewitt, M. J. and Meddis, R. (**1991**). "An evaluation of eight computer models of mammalian inner hair-cell function," *Journal of the Acoustical Society of America* **90**, 904–917.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (**2006**). "A fast learning algorithm for deep belief nets," *Neural Computation* **18**, 1527–1554.

Hirata, Y. (**2004**). "Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts," *Computer Assisted Language Learning* **17**, 357–376.

Ho-Ching, F. W.-l., Mankoff, J., and Landay, J. A. (**2003**). "Can you see what I hear?: The design and evaluation of a peripheral sound display for the deaf," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 161–168 (ACM).

Holder, W. (**1731**). *A Treatise on the Natural Grounds, and Principles of Harmony* (London: W. Pearson).

Holdsworth, J., Nimmo-Smith, I., Patterson, R. D., and Rice, P. (**1988**). "Implementing a gammatone filter bank," Technical Report, MRC Applied Psychology Unit, SVOS final report, APU report 2341 annex C.

Hornbostel, E. M. von (**1931**). "The time-theory of sound localization: A restatement," in *Report of a Discussion on Audition*, 120–127 (London: The Physical Society).

Hornbostel, E. M. von and Wertheimer, M. (**1920**). "Über die Wahrnehmung der Schallrichtung [On the perception of the direction of sound]," *Sitzungsberichte Akademie der Wissenschaften, Berlin* **15**, 388–396.

Hoshen, Y., Weiss, R. J., and Wilson, K. W. (**2015**). "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4624–4628.

Houtsma, A. J. M., Durlach, N. I., and Braida, L. D. (**1980**). "Intensity perception XI. Experimental results on the relation of intensity resolution to loudness matching," *Journal of the Acoustical Society of America* **68**, 807–813.

Houtsma, A. J. M., Rossing, T. D., and Wagenaars, W. M. (**1987**). *Auditory Demonstrations*, volume (CD and booklet) (Acoustical Society of America).

Howell, W. H. (**1915**). *A Text-Book of Physiology for Medical Students and Physicians*, 6th edition (London: W. B. Saunders and Co.).

Hu, G. and Wang, D. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, 79–82 (IEEE).

Hu, G. and Wang, D. (**2008**). "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America* **124**, 1306–1319.

Hudspeth, A. J. (**1982**). "Extracellular current flow and the site of transduction by vertebrate hair cells," *The Journal of Neuroscience* **2**, 1–10.

Hudspeth, A. J. and Corey, D. P. (**1977**). "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," *Proceedings of the National Academy of Sciences* **74**, 2407–2411.

Huggins, W. H. and Licklider, J. C. R. (**1951**). "Place mechanisms of auditory frequency analysis," *Journal of the Acoustical Society of America* **23**, 290–299.

Hukin, R. W. and Damper, R. I. (**1989**). "Testing an auditory model by resynthesis," in *First European Conference on Speech Communication and Technology*, 1243–1246.

Hummersone, C., Brookes, T., and Mason, R. (**2010**a). "A comparison of computational precedence models for source separation in reverberant environments," in *Audio Engineering Society Convention 128*.

Hummersone, C., Mason, R., and Brookes, T. (**2010**b). "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 1867–1871.

Humphrey, E. J. and Bello, J. P. (**2012**). "Rethinking automatic chord recognition with convolutional neural networks," in *International Conference on Machine Learning and Applications*, volume 2, 357–362 (IEEE).

Humphrey, E. J., Bello, J. P., and LeCun, Y. (**2012**). "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*.

Hurewicz, W. (**1947**). "Filters and servo systems with pulsed data," in *Theory of Servomechanisms*, edited by H. M. James, N. B. Nichols, and R. S. Phillips, 231–261 (McGraw-Hill Book Company).

Huron, D. B. (**2006**). *Sweet Anticipation: Music and the Psychology of Expectation* (MIT Press).

Hurst, C. H. (**1895**). "A new theory of hearing," *Proceedings and Transactions of the Liverpool Biological Society* **9**, 321–353.

Indyk, P. and Motwani, R. (**1998**). "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, 604–613.

Irino, T. and Kawahara, H. (**1993**). "Signal reconstruction from modified auditory wavelet transform," *IEEE Transactions on Signal Processing* **41**, 3549–3554.

Irino, T. and Patterson, R. D. (**1996**). "Temporal asymmetry in the auditory system," *Journal of the Acoustical Society of America* **99**, 2316–2331.

Irino, T. and Patterson, R. D. (**1997**). "A time-domain, level-dependent auditory filter: The gammachirp," *Journal of the Acoustical Society of America* **101**, 412–419.

Irino, T. and Patterson, R. D. (**2001**). "A compressive gammachirp auditory filter for both physiological and psychophysical data," *Journal of the Acoustical Society of America* **109**, 2008–2022.

Irino, T. and Patterson, R. D. (**2006**a). "A dynamic compressive gammachirp auditory filterbank," *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 2222–2232.

Irino, T. and Patterson, R. D. (**2006**b). "A dynamic compressive gammachirp auditory filterbank," *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 2222–2232.

Irino, T., Patterson, R. D., and Kawahara, H. (**2006**). "Speech segregation using an auditory vocoder with event-synchronous enhancements," *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 2212–2221.

Irino, T. and Unoki, M. (**1999**). "An analysis/synthesis auditory filterbank based on an IIR implementation of the gammachirp," *Journal of the Acoustical Society of Japan* **20**, 397–406.

Itakura, F. (**1975**). "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**, 67–72.

Ives, D. T. and Patterson, R. D. (**2008**). "Pitch strength decreases as f0 and harmonic resolution increase in complex tones composed exclusively of high harmonics," *Journal of the Acoustical Society of America* **123**, 2670.

Jacobs, C. E., Finkelstein, A., and Salesin, D. H. (**1995**). "Fast multiresolution image querying," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 277–286 (ACM).

Jaitly, N. and Hinton, G. E. (**2011**). "Learning a better representation of speech sound waves using restricted Boltzmann machines," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5884–5887.

James, W. (**1890**). *The Principles of Psychology* (New York: Dover Publications).

Janssen, T. and Müller, J. (**2007**). "Otoacoustic emissions as a diagnostic tool in a clinical context," in *Active Processes and Otoacoustic Emissions in Hearing*, edited by G. A. Manley, R. R. Fay, and A. N. Popper, 421–460 (Springer).

Jaramillo, F. and Hudspeth, A. J. (**1991**). "Localization of the hair cell's transduction channels at the hair bundle's top by iontophoretic application of a channel blocker," *Neuron* **7**, 409–420.

Jarvis, E. D. (**2004**). "Learned birdsong and the neurobiology of human language," *Annals of the New York Academy of Sciences* **1016**, 749–777.

Jeffress, L. A. (**1948**). "A place theory of sound localization," *Journal of Comparative and Physiological Psychology* **41**, 35–39.

Jeffress, L. A. (**1970**). "Masking," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias, volume 1, 87–114 (Academic Press).

Jensen, K. (**2005**). "A causal rhythm grouping," in *Computer Music Modeling and Retrieval*, 83–95 (Springer).

Jensen, K. (**2007**). "Multiple scale music segmentation using rhythm, timbre, and harmony," *EURASIP Journal on Applied Signal Processing* **2007**, 159–159.

Joachims, T. (**2002**). "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142.

Johannesma, P. I. M. (**1972**). "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Proceedings of the IPO Symposium on Hearing Theory*, edited by B. L. Cardozo, 58–69 (Eindhoven: IPO).

Johannesma, P. I. M. (**1980**). "Narrow band filters and active resonators [Comment on: 1 'Observations on the generator mechanism of stimulus frequency acoustic emissions – two tone suppression' (D.T. Kemp and R.A. Churn). 2 'On the mechanism of the evoked cochlear mechanical response' (H. P. Wit and R. J. Ritsma).]," in *Psychophysical, Physiological and Behavioural Studies in Hearing*, edited by G. van den Brink and F. A. Bilsen, 62–63 (Delft University Press).

Johnson, K. (**2003**). *Acoustic and Auditory Phonetics*, 2nd edition (Wiley-Blackwell).

Johnson, R. M. (**1997**). *Linear Differential and Difference Equations: A Systems Approach for Mathematicians and Engineers* (Albion Publishing).

Johnson, S. L., Eckrich, T., Kuhn, S., Zampini, V., Franz, C., Ranatunga, K. M., Roberts, T. P., Masetto, S., Knipper, M., Kros, C. J., and Marcotti, W. (**2011**). "Position-dependent patterning of spontaneous action potentials in immature cochlear inner hair cells," *Nature Neuroscience* **14**, 711–717.

Johnston, J. D. (**1988**). "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications* **6**, 314–323.

Johnstone, B. M., Patuzzi, R., and Yates, G. K. (**1986**). "Basilar membrane measurements and the travelling wave," *Hearing Research* **22**, 147–153.

Jones, A. T. (**1928**). "The vibration of bells," *Physical Review* **31**, 1092–1102.

Joris, P. X., Carney, L. H., Smith, P. H., and Yin, T. C. T. (**1994**). "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency," *Journal of Neurophysiology* **71**, 1022–1036.

Joris, P. X., Smith, P. H., and Yin, T. C. T. (**1998**). "Coincidence detection in the auditory system: 50 years after Jeffress," *Neuron* **21**, 1235–1238.

Joris, P. X. and Yin, T. C. T. (**1995**). "Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences," *Journal of Neurophysiology* **73**, 1043–1062.

Joris, P. X. and Yin, T. C. T. (**2007**). "A matter of time: Internal delays in binaural processing," *Trends in Neurosciences* **30**, 70–78.

Jungnickel, C. and McCormmach, R. (**1986**). *Intellectual Mastery of Nature. Theoretical Physics from Ohm to Einstein, Volume 1: The Torch of Mathematics, 1800 to 1870* (University Of Chicago Press).

Jutten, C., Hérault, J., and Guérin, A. (**1988**). "IN.C.A.: An independent component analyser based on an adaptive neuromimetic network," in *Artificial Intelligence and Cognitive Sciences*, edited by J. Demongeot, T. Herve, V. Rialle, and C. Roche (Manchester University Press).

Kaiser, A. and Manley, G. A. (**1994**). "Physiology of single putative cochlear efferents in the chicken," *Journal of Neurophysiology* **72**, 2966–2979.

Kaiser, J. F. and David, E. E., Jr. (**1960**). "Reproducing the cocktail party effect," *Journal of the Acoustical Society of America* **32**, 918.

Karam, L. J., McClellan, J. H., Selesnick, I. W., and Burrus, C. S. (**1999**). "Digital filtering," in *Digital Signal Processing Handbook*, edited by V. K. Madisetti and D. B. Williams (CRC Press).

Karino, S., Smith, P. H., Yin, T. C. T., and Joris, P. X. (**2011**). "Axonal branching patterns as sources of delay in the mammalian auditory brainstem: A re-examination," *The Journal of Neuroscience* **31**, 3016–3031.

Karjalainen, M. (**1987**). "Auditory models for speech processing," in *Proceedings of the International Congress of Phonetic Sciences, Tallinn* (Academy of Sciences of the Estonian S.S.R.).

Karklin, Y., Ekanadham, C., and Simoncelli, E. P. (**2012**). "Hierarchical spike coding of sound," in *Advances in Neural Information Processing Systems*, volume 25, 3041–3049.

Kashyap, R. L. (**1970**). "Algorithms for pattern classification," in *Adaptive, Learning and Pattern Recognition Systems*, edited by K. S. Fu and J. M. Mendel, 81–113 (Academic Press).

Kates, J. M. (**1983**). "Auditory spectral analysis model using the chirp z-transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **31**, 148–156.

Kates, J. M. (**1991**). "A time-domain digital cochlear model," *IEEE Transactions on Signal Processing* **39**, 2573–2592.

Kates, J. M. (**1993**a). "Accurate tuning curves in a cochlear model," *IEEE Transactions on Speech and Audio Processing* **1**, 453–462.

Kates, J. M. (**1993**b). "Toward a theory of optimal hearing aid processing," *Journal of Rehabilitation Research and Development* **30**, 39–48.

Kates, J. M. (**2010**). "Understanding compression: Modeling the effects of dynamic-range compression in hearing aids," *International Journal of Audiology* **49**, 395–409.

Katsiamis, A. G., Drakakis, E. M., and Lyon, R. F. (**2006**). "Introducing the differentiated all-pole and one-zero gammatone filter responses and their analog VLSI log-domain implementation," in *IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, volume 1.

Katsiamis, A. G., Drakakis, E. M., and Lyon, R. F. (**2007**). "Practical gammatone-like filters for auditory processing," *EURASIP Journal on Audio, Speech, and Music Processing*. Article ID 63685, 15 pp. doi:10.1155/2007/63685.

Katsiamis, A. G., Drakakis, E. M., and Lyon, R. F. (**2009**). "A biomimetic, 4.5 μW, 120+dB, log-domain cochlea channel with AGC," *IEEE Journal of Solid-State Circuits* **44**, 1006–1022.

Kayser, C., Körding, K. P., and König, P. (**2003**). "Learning the nonlinearity of neurons from natural visual stimuli," *Neural Computation* **15**, 1751–1759.

Kemp, D. T. (**1978**). "Stimulated acoustic emissions from within the human auditory system," *Journal of the Acoustical Society of America* **64**, 1386–1391.

Kemp, D. T. (**1979**). "Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea," *Archives of Oto-Rhino-Laryngology* **224**, 37–45.

Kersten, D. (**2000**). "High-level vision as statistical inference," in *The New Cognitive Neurosciences*, edited by M. S. Gazzaniga, 2nd edition, 353–363 (MIT Press).

Kiang, N. Y. S. (**1965**). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (MIT Press).

Kiang, N. Y. S. and Moxon, E. C. (**1974**). "Tails of tuning curves of auditory-nerve fibers," *Journal of the Acoustical Society of America* **55**, 620–630.

Kick, S. A. and Simmons, J. A. (**1984**). "Automatic gain control in the bat's sonar receiver and the neuroethology of echolocation," *The Journal of Neuroscience* **4**, 2725–2737.

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005**). "The advantage of knowing where to listen," *Journal of the Acoustical Society of America* **118**, 3804–3815.

Killion, M. C. (**1997**). "Hearing aids: Past, present, future: Moving toward normal conversations in noise," *British Journal of Audiology* **31**, 141–148.

Killion, M. C. and Fikret-Pasa, S. (**1993**). "The 3 types of sensorineural hearing loss: Loudness and intelligibility considerations," *Hearing Journal* **46**, 31–36.

Killion, M. C. and Tillman, T. W. (**1982**). "Evaluation of high-fidelity hearing aids," *Journal of Speech, Language and Hearing Research* **25**, 15–25.

Killion, M. C. and Villchur, E. (**1993**). "Kessler was right—partly: But SIN test shows some aids improve hearing in noise," *Hearing Journal* **46**, 31–35.

Kim, D. O. (**1980**). "Cochlear mechanics: Implications of electrophysiological and acoustical observations," *Hearing Research* **2**, 297–317.

Kim, D. O. (**1984**). "Functional roles of the inner- and outer-hair-cell subsystems in the cochlea and brainstem," *Hearing Science: Recent Advances* 241–262.

Kim, D. O., Molnar, C. E., and Pfeiffer, R. R. (**1973**). "A system of nonlinear differential equations modeling basilar-membrane motion," *Journal of the Acoustical Society of America* **54**, 1517–1529.

Kim, D. O., Neely, S. T., Molnar, C. E., and Matthews, J. W. (**1980**). "An active cochlear model with negative damping in the cochlear partition: Comparison with Rhode's ante-and post-mortem results," in *Psychological, Physiological and Behavioral Studies in Hearing*, edited by F. A. Bilsen and G. van den Brink, 7–14 (Delft University Press).

Kim, D. O., Parham, K., Zhao, H.-B., and Ghoshal, S. (**1995**). "The olivocochlear feedback gain control subsystem: Ascending input from the small cell cap of the cochlear nucleus?," in *Active Hearing*, edited by Å. Flock, D. Ottoson, and M. Ulfendahl, 31–51 (Pergamon).

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America* **126**, 1486.

King, A. J. (**1997**). "Sensory processing: Signal selection by cortical feedback," *Current Biology* **7**, R85–R88.

King, D. B. and Wertheimer, M. (**2007**). *Max Wertheimer and Gestalt Theory* (Transaction Publishers).

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R. (**2013**). "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, 1–4 (IEEE).

Klemm, O. (**1920**). "Untersuchungen über die Lokalisation von Schallreizen. 4. Mitteilung: Über den Einfluß des binauralen Zeitunterschiedes auf die Lokalisation," *Archiv für die gesamte Psychologie* **40**, 117–146.

Kletsky, E. J. and Zwislocki, J. J. (**1981**). "A network model of cochlear dynamics," *Journal of the Acoustical Society of America* **69**, S52.

Knudsen, E. I. (**1982**). "Auditory and visual maps of space in the optic tectum of the owl," *The Journal of Neuroscience* **2**, 1177–1194.

Knudsen, E. I., du Lac, S., and Esterly, S. D. (**1987**). "Computational maps in the brain," *Annual Review of Neuroscience* **10**, 41–65.

Knudsen, E. I. and Knudsen, P. F. (**1983**). "Space-mapped auditory projections from the inferior colliculus to the optic tectum in the barn owl (Tyto alba)," *The Journal of Comparative Neurology* **218**, 187–196.

Knudsen, V. O. (**1923**). "The sensibility of the ear to small differences of intensity and frequency," *Physical Review* **21**, 84–102.

Koh, K., Kim, S.-J., and Boyd, S. (**2007**). "An interior-point method for large-scale ℓ_1-regularized logistic regression," *Journal of Machine Learning Research (JMLR)* **8**, 1519–1555.

Kojima, S. (**2003**). *Search for the Origins of Human Speech: Auditory and Vocal Functions of the Chimpanzee* (Trans Pacific Press).

Kollmeier, B., Brand, T., and Meyer, B. (**2008**). "Perception of speech and sound," in *Springer Handbook of Speech Processing*, edited by J. Benesty, M. M. Sondhi, and Y. Huang, 61–82 (Springer Verlag).

Kollmeier, B., Peissig, J., and Hohmann, V. (**1993**). "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *Journal of Rehabilitation Research and Development* **30**, 82–94.

Konishi, M. (**1991**). "Deciphering the brain's codes," *Neural Computation* **3**, 1–18.

Konishi, M. (**1995**). "Neural mechanisms of auditory image formation," in *The Cognitive Neurosciences*, edited by M. S. Gazzaniga, 269–277 (MIT Press).

Konishi, M. (**2003**). "Coding of auditory space," *Annual Review of Neuroscience* **26**, 31–55.

Konishi, M., Sullivan, W. E., and Takahashi, T. (**1985**). "The owl's cochlear nuclei process different sound localization cues," *Journal of the Acoustical Society of America* **78**, 360–364.

Korenberg, M. J. and Hunter, I. W. (**1986**). "The identification of nonlinear biological systems: LNL cascade models," *Biological Cybernetics* **55**, 125–134.

Kouh, M. and Poggio, T. (**2008**). "A canonical neural circuit for cortical nonlinear operations," *Neural Computation* **20**, 1427–1451.

Kricos, P. B. (**2006**). "Audiologic management of older adults with hearing loss and compromised cognitive/psychoacoustic auditory processing capabilities," *Trends in Amplification* **10**, 1–28.

Kristjansson, T. T., Hershey, J. R., Olsen, P. A., Rennie, S. J., and Gopinath, R. A. (**2006**). "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Interspeech: Ninth International Conference on Spoken Language Processing*.

Krogh, A. (**2008**). "What are artificial neural networks?," *Nature Biotechnology* **26**, 195–197.

Kros, C. J. (**2007**). "How to build an inner hair cell: Challenges for regeneration," *Hearing Research* **227**, 3–10.

Krueger, L. E. (**1989**). "Reconciling Fechner and Stevens: Toward a unified psychophysical law," *Behavioral and Brain Sciences* **12**, 251–267.

Kubin, G. and Kleijn, W. B. (**1999**). "On speech coding in a perceptual domain," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 205–208.

Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," *Journal of the Acoustical Society of America* **62**, 157–167.

Kuile, E. ter (**1900**). "Die Uebertragung der Energie von der Grundmembran auf die Haarzellen (The transfer of energy from the basilar membrane to the hair cells)," *Pflügers Archiv European Journal of Physiology* **79**, 146–157.

Kulesza, R. J., Jr. (**2007**). "Cytoarchitecture of the human superior olivary complex: Medial and lateral superior olive," *Hearing Research* **225**, 80–90.

Kulkarni, A. and Colburn, H. S. (**1998**). "Role of spectral detail in sound-source localization," *Nature* **396**, 747–749.

Küpfmüller, K. (**1928**). "Uber die Dynamik der selbsttatigen Verstarkungsregler," *Elektrische Nachrichtentechnik* **5**, 459–467.

Kurzweil, R. (**2012**). *How to Create a Mind: The Secret of Human Thought Revealed* (Viking).

Kuwabara, N. and Suga, N. (**1993**). "Delay lines and amplitude selectivity are created in subthalamic auditory nuclei: The brachium of the inferior colliculus of the mustached bat," *Journal of Neurophysiology* **69**, 1713–1724.

La Bruyère, J. de (**1713**). *The Works of Monsieur De La Bruyère Volume II: Containing The Characters or Manners of the Present Age* (London: Curll, Sanger, and Pemberton).

La Rochefoucauld, O. de and Olson, E. S. (**2007**). "The role of organ of Corti mass in passive cochlear tuning," *Biophysical Journal* **93**, 3434–3450.

Laënnec, R.-T.-H. (**1819**). *De l'auscultation médiate, ou traité du diagnostic des maladies des poumons et du coeur fondé principalement sur ce nouveau moyen d'exploration* (Paris: J.-A. Brosson and J.-S. Chaudé).

Lamb, H. (**1879**). *A Treatise on the Mathematical Theory of the Motion of Fluids* (Cambridge: The University Press).

Lamb, H. (**1895**). *Hydrodynamics* (Cambridge: The University Press).

Langner, G. (**1981**). "Neuronal mechanisms for pitch analysis in the time domain," *Experimental Brain Research* **44**, 450–454.

Langner, G. (**1997**). "Neural processing and representation of periodicity pitch," *Acta Oto-Laryngologica* **117**, 68–76.

Langner, G. (**2005**). "Topographic representation of periodicity information: The 2nd neural axis of the auditory system," in *Plasticity and Signal Representation in the Auditory System*, edited by M. M. M. Josef Syka, 37–51 (Springer).

Langner, G., Albert, M., and Briede, T. (**2002**). "Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (Chinchilla laniger)," *Hearing Research* **168**, 110–130.

Langner, G., Dinse, H. R., and Godde, B. (**2009**). "A map of periodicity orthogonal to frequency representation in the cat auditory cortex," *Frontiers in Integrative Neuroscience* **3**, 1–11.

Langner, G., Sams, M., Heil, P., and Schulze, H. (**1997**). "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography," *Journal of Comparative Physiology A* **181**, 665–676.

Laudanski, J., Coombes, S., Palmer, A. R., and Sumner, C. J. (**2010**). "Mode-locked spike trains in responses of ventral cochlear nucleus chopper and onset neurons to periodic stimuli," *Journal of Neurophysiology* **103**, 1226–1237.

Lazzaro, J., Ryckebusch, S., Mahowald, M. A., and Mead, C. A. (**1989**). "Winner-take-all networks of O(n) complexity," in *Advances in Neural Information Processing Systems*, volume 1, 703–711.

Leatham, A. (**1951**). "The phonocardiogram of aortic stenosis," *British Heart Journal* **13**, 153.

Lee, J.-H., Jung, H.-Y., Lee, T.-W., and Lee, S.-Y. (**2000**). "Speech feature extraction using independent component analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, 1631–1634.

Lee, K. and Ellis, D. P. W. (**2006**). "Voice activity detection in personal audio recordings using autocorrelogram compensation," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1970–1973 (ISCA).

Lepore, J. (**2002**). *A Is for American: Letters and Other Characters in the Newly United States* (Knopf).

Levitt, H. (**2004**). "Compression amplification," in *Compression: From Cochlea to Cochlear Implants*, edited by S. P. Bacon, R. R. Fay, and A. N. Popper, 153–183 (Springer).

Levitt, H. (**2007**). "A historical perspective on digital hearing aids: How digital technology has changed modern hearing aids," *Trends in Amplification* **11**, 7–24.

Levitt, H., Oden, C., Simon, H., Noack, C., and Lotze, A. (**2012**). "Computer-based training methods for age-related APD: Past, present, and future," in *Auditory Processing Disorders: Assessment, Management, and Treatment*, edited by D. S. Geffner and D. Ross-Swain, 2nd edition, 773–801 (Plural Pub.).

Levitt, H. and Rabiner, L. R. (**1967**). "Binaural release from masking for speech and gain in intelligibility," *Journal of the Acoustical Society of America* **42**, 601–608.

Li, X., Nie, K., Imennov, N. S., Rubinstein, J. T., and Atlas, L. E. (**2013**). "Improved perception of music with a harmonic based algorithm for cochlear implants," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **21**, 684–694.

Li, Y., Liu, Z., Shi, P., and Zhang, J. (**2010**). "The hearing gene *Prestin* unites echolocating bats and whales," *Current Biology* **20**, R55–R56.

Li, Y. and Wang, D. (**2009**). "On the optimality of ideal binary time–frequency masks," *Speech Communication* **51**, 230–239.

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," *Experientia* **7**, 128–133, reprinted in *Physiological Acoustics* (1979) (E. D. Schubert, ed.), Dowden, Hutchinson and Ross, Inc.

Licklider, J. C. R. (**1953**). "Hearing," *Annual Review of Psychology* **4**, 89–110.

Licklider, J. C. R. (**1956**). "Auditory frequency analysis," in *Information Theory*, edited by C. Cherry, 253–268, reprinted in *Forty Germinal Papers in Human Hearing* (1969) (J. D. Harris, ed.), *The Journal of Auditory Research* (London: Butterworth).

Licklider, J. C. R. (**1959**). "Three auditory theories," in *Psychology: A Study of a Science. Volume 1. Sensory, Perceptual, and Physiological Formulations Study I. Conceptual and Systematic*, edited by S. Koch, 41–144 (McGraw-Hill).

Lighthill, J. (**1981**). "Energy flow in the cochlea," *Journal of Fluid Mechanics* **106**, 149–213.

Lim, K.-M. and Steele, C. R. (**2002**). "A three-dimensional nonlinear active cochlear model analyzed by the WKB–numeric method," *Hearing Research* **170**, 190–205.

Lindeberg, T. (**1990**). "Scale-space for discrete signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **12**, 234–254.

Lindemann, W. (**1986**). "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *Journal of the Acoustical Society of America* **80**, 1623–1630.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," *Journal of the Acoustical Society of America* **106**, 1633–1654.

Liu, H., Liu, C., and Huang, Y. (**2011**). "Adaptive feature extraction using sparse coding for machinery fault diagnosis," *Mechanical Systems and Signal Processing* **25**, 558–574.

Liu, S.-C., van Schaik, A., Minch, B. A., and Delbruck, T. (**2010**). "Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2027–2030.

Liu, Y.-W. and Neely, S. T. (**2009**). "Outer hair cell electromechanical properties in a nonlinear piezoelectric model," *Journal of the Acoustical Society of America* **126**, 751–761.

Loizou, P. C. (**1998**). "Mimicking the human ear," *IEEE Signal Processing Magazine* **15**, 101–130.

Loizou, P. C. (**2006**). "Speech processing in vocoder-centric cochlear implants," *Advances in Otorhinolaryngology* **64**, 109–143.

Loizou, P. C. and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing* **19**, 47–56.

Lopez-Poveda, E. A. (**2005**). "Spectral processing by the peripheral auditory system: Facts and models," in *Auditory Spectral Processing (Vol. 70 International Review of Neurobiology)*, edited by M. S. Malmierca and D. R. F. Irvine, 7–48 (Academic Press).

Lopez-Poveda, E. A. and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," *Journal of the Acoustical Society of America* **110**, 3107–3118.

Louage, D. H. G., van der Heijden, M., and Joris, P. X. (**2005**). "Enhanced temporal response properties of anteroventral cochlear nucleus neurons to broadband noise," *The Journal of Neuroscience* **25**, 1560.

Luciani, L. (**1917**). *Human Physiology: The Sense Organs*, volume 4 (New York: Macmillan & Co.).

Lukashkin, A. N., Walling, M. N., and Russell, I. J. (**2007**). "Power amplification in the mammalian cochlea," *Current Biology* **17**, 1340–1344.

Lumpkin, E. A. and Hudspeth, A. J. (**1995**). "Detection of ca2+ entry through mechanosensitive channels localizes the site of mechanoelectrical transduction in hair cells," *Proceedings of the National Academy of Sciences* **92**, 10297.

Lutfi, R. A. and Patterson, R. D. (**1984**). "On the growth of masking asymmetry with stimulus intensity," *Journal of the Acoustical Society of America* **76**, 739–745.

Lyon, R. F. (**1981**). "The optical mouse and an architectural methodology for smart digital sensors," in *VLSI Systems and Computations*, edited by H. T. Kung, R. Sproull, and G. Steele, 1–19 (Computer Science Press).

Lyon, R. F. (**1982**). "A computational model of filtering, detection, and compression in the cochlea," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1282–1285.

Lyon, R. F. (**1983**). "A computational model of binaural localization and separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1148–1151.

Lyon, R. F. (**1984**). "Computational models of neural auditory processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3611–3614.

Lyon, R. F. (**1987**). "Speech recognition in scale space," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1265–1268.

Lyon, R. F. (**1990**). "Automatic gain control in cochlear mechanics," in *The Mechanics and Biophysics of Hearing*, edited by P. Dallos, C. D. Geisler, J. W. Matthews, M. Ruggero, and C. R. Steele, 395–420 (Springer).

Lyon, R. F. (**1996**a). "The all-pole gammatone filter and auditory models," in *Forum Acusticum '96* (European Acoustics Association).

Lyon, R. F. (**1996**b). "The all-pole gammatone filter and auditory models," Technical Report, Apple Computer and dicklyon.com.

Lyon, R. F. (**1997**). "All-pole models of auditory filtering," in *Diversity in Auditory Mechanics*, edited by E. R. Lewis, G. R. Long, R. F. Lyon, P. M. Narins, C. R. Steele, and E. Hecht-Poinar, 205–211 (World Scientific Publishing).

Lyon, R. F. (**1998**). "Filter cascades as analogs of the cochlea," in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, edited by T. S. Lande, 3–18 (Kluwer Academic Publishers).

Lyon, R. F. (**2010**). "Machine hearing: An emerging field [Exploratory DSP column]," *IEEE Signal Processing Magazine* **27**, 131–139.

Lyon, R. F. (**2011**a). "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," *Journal of the Acoustical Society of America* **130**, 3893–3904.

Lyon, R. F. (**2011**b). "A pole–zero filter cascade provides good fits to human masking data and to basilar membrane and neural data," in *What Fire is in Mine Ears—Progress in Auditory Biomechanics: Proceedings of the 11th International Mechanics of Hearing Workshop*, edited by C. A. Shera and E. S. Olson, 224–230 (AIP).

Lyon, R. F. (**2014**). "The optical mouse: Early biomimetic embedded vision," in *Advances in Embedded Computer Vision*, edited by B. Kisačanin and M. Gelautz, 3–22 (Springer).

Lyon, R. F. and Dyer, L. (**1986**). "Experiments with a computational model of the cochlea," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1975–1978.

Lyon, R. F., Katsiamis, A. G., and Drakakis, E. M. (**2010**a). "History and future of auditory filter models," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 3809–3812.

Lyon, R. F. and Mead, C. A. (**1988**a). "An analog electronic cochlea," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**, 1119–1134.

Lyon, R. F. and Mead, C. A. (**1988**b). "Cochlear hydrodynamics demystified," Technical Report, California Institute of Technology.

Lyon, R. F., Ponte, J., and Chechik, G. (**2011**). "Sparse coding of auditory features for machine hearing in interference," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5876–5879.

Lyon, R. F., Rehn, M., Bengio, S., Walters, T. C., and Chechik, G. (**2010**b). "Sound retrieval and ranking using sparse auditory representations," *Neural Computation* **22**, 2390–2416.

Magnasco, M. O. (**2003**). "A wave traveling over a Hopf instability shapes the cochlear tuning curve," *Physical Review Letters* **90**, 058101.

Makhoul, J. and Cosell, L. (**1976**). "LPCW: An LPC vocoder with linear predictive spectral warping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 466–469.

Mallat, S. and Zhang, Z. (**1993**). "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing* **41**, 3397–3415.

Mallock, A. (**1908**). "Note on the sensibility of the ear to the direction of explosive sounds," *Proceedings of the Royal Society of London: Series A* **80**, 110–112.

Mann, S., Nolan, J., and Wellman, B. (**2003**). "Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments," *Surveillance and Society* **1**, 331–355.

Markel, J. E. and Gray, A. H. (**1982**). *Linear Prediction of Speech* (Springer).

Marolt, M. (**2008**). "A mid-level representation for melody-based retrieval in audio collections," *IEEE Transactions on Multimedia* **10**, 1617–1625.

Marr, D. (**1982**). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Company).

Marrill, T., Hartley, A. K., Evans, T. G., Bloom, B. H., Park, D. M. R., Hart, T. P., and Darley, D. L. (**1963**). "CYCLOPS-1: A second-generation recognition system," in *Proceedings of the Fall Joint Computer Conference*, 27–33.

Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press).

Massie, D. (**2012**). "Coefficient interpolation for the Max Mathews phasor filter," in *Audio Engineering Society Convention 133*.

Mathevon, N., Koralek, A., Weldele, M., Glickman, S. E., and Theunissen, F. E. (**2010**). "What the hyena's laugh tells: Sex, age, dominance and individual signature in the giggling call of Crocuta crocuta," *BMC Ecology* **10**, 9.

Mathews, M. V. (**1959**). "Extremal coding for speech transmission," *Journal of the Acoustical Society of America* **31**, 113–113.

Mathews, M. V. (**1961**). "An acoustic compiler for music and psychological stimuli," *Bell System Technical Journal* **40**, 677–694.

Mathews, M. V. (**1963**). "Energy detection for human auditory detection," in *Time Series Analysis*, edited by M. Rosetblatt, 349–361 (John Wiley & Sons).

Mathews, M. V. and Pfafflin, S. M. (**1965**). "Effect of filter type on energy-detection models for auditory signal detection," *Journal of the Acoustical Society of America* **38**, 1055–1056.

Mathews, M. V. and Pierce, J. R. (**1980**). "Harmony and nonharmonic partials," *Journal of the Acoustical Society of America* **68**, 1252–1257.

Mathews, M. V. and Smith, J. O. (**2003**). "Methods for synthesizing very high Q parametrically well behaved two pole filters," in *Proceedings of the Stockholm Musical Acoustics Conference (SMAC 2003)* (Royal Swedish Academy of Music).

Matthews, T., Fong, J., and Mankoff, J. (**2005**). "Visualizing non-speech sounds for the deaf," in *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, 52–59 (ACM).

Mauchly, J. W. (**1942**). "The use of high-speed vacuum tube devices for calculating," in *(1982) The Origins of Digital Computers: Selected Papers*, edited by B. Randell, 355–358 (Springer).

Mayer, A. M. (**1876**). "Researches in acoustics, No. VIII," *Philosophical Magazine* **2**, 500–507.

Mayer, A. M. (**1878**). *Sound: A Series of Simple, Entertaining, and Inexpensive Experiments in the Phenomena of Sound: For the Use of Students of Every Age* (New York: D. Appleton and Company).

McAdams, S. (**1982**). "Spectral fusion and the creation of auditory images," in *Music, Mind, and Brain: The Neuropsychology of Music*, edited by M. Clynes, 279–298 (Plenum Press).

McDermott, H. J. (**2004**). "Music perception with cochlear implants: A review," *Trends in Amplification* **8**, 49–82.

McDermott, H. J. (**2011**). "A technical comparison of digital frequency-lowering algorithms available in two current hearing aids," *PLoS One* **6**, e22358.

McDermott, J. H. and Simoncelli, E. P. (**2011**). "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron* **71**, 926–940.

McDonnell, M. D. and Abbott, D. (**2009**). "What is stochastic resonance? definitions, misconceptions, debates, and its relevance to biology," *PLoS Computational Biology* **5**, e1000348.

McFadden, D. and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences," *Journal of the Acoustical Society of America* **59**, 634–639.

McKendrick, J. G. (**1889**). *A Text Book of Physiology*, volume II – Special Physiology (New York: Macmillan and Co.).

McKendrick, J. G. (**1899**). *On the Physiological Perception of Musical Tone: being the seventh Robert Boyle lecture* (London: Henry Frowde).

McKendrick, J. G. and Gray, A. A. (**1900**). "The ear," in *Text-Book of Physiology*, edited by E. A. Sharpey-Schäfer, volume 2, 1149–1205 (Edinburgh and London: Pentland).

Mead, C. A. (**1990**). "Neuromorphic electronic systems," *Proceedings of the IEEE* **78**, 1629–1636.

Mead, C. A. (**2002**). *Collective Electrodynamics: Quantum Foundations of Electromagnetism* (The MIT Press).

Meddis, R. (**1986**). "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America* **79**, 702–711.

Meddis, R. (**1988**). "Simulation of auditory-neural transduction: Further studies," *Journal of the Acoustical Society of America* **83**, 1056–1063.

Meddis, R. and Hewitt, M. J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *Journal of the Acoustical Society of America* **89**, 2866–2882.

Meddis, R. and Hewitt, M. J. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America* **91**, 233–245.

Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A. (**2001**). "A computational algorithm for computing nonlinear auditory frequency selectivity," *Journal of the Acoustical Society of America* **109**, 2852–2861.

Meijering, E. (**2002**). "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proceedings of the IEEE* **90**, 319–342.

Mellinger, D. K. (**1991**). "Event formation and separation in musical sound," Ph.D. thesis, Stanford University.

Mermelstein, P. (**1976**). "Distance measures for speech recognition—psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, edited by C. H. Chen, 374–388 (Academic Press).

Merzenich, M. M. and Kaas, J. H. (**1980**). "Principles of organization of sensory-perceptual systems in mammals," in *Progress in Psychobiology and Physiological Psychology*, edited by J. M. Sprague, volume 9, 1–42 (Academic Press).

Mesgarani, N. and Chang, E. F. (**2012**). "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature* **485**, 233–236.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (**2014**). "Phonetic feature encoding in human superior temporal gyrus," *Science* **343**, 1006–1010.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (**2008**). "Phoneme representation and classification in primary auditory cortex," *Journal of the Acoustical Society of America* **123**, 899–909.

Meyer, M. (**1907**). "An introduction to the mechanics of the inner ear," *Science Series, University Missouri Studies* **2**, 1–140.

Meyer, M. F. (**1899**). "Zur Theorie des Hörens," *Arch. ges. Physiol Pflügers's* **78**, 346–362.

Middleton, D. (**1948**). "Some general results in the theory of noise through non-linear devices," *Quarterly of Applied Mathematics* **5**, 445–498.

Miller, S. and Childers, D. (**2012**). *Probability and Random Processes: With Applications to Signal Processing and Communications* (Academic Press).

Milne, A. A. (**1926**). *Winnie-the-Pooh* (London: Methuen & Co.).

Minsky, M. L. and Papert, S. (**1969**). *Perceptrons: An Introduction to Computational Geometry* (MIT Press).

Mohamed, A.-M. O. (**2006**). *Principles and Applications of Time Domain Electrometry in Geoenvironmental Engineering* (Taylor & Francis).

Møller, A. R. (**1962**). "Acoustic reflex in man," *Journal of the Acoustical Society of America* **34**, 1524–1534.

Møller, A. R. (**2003**). *Auditory Physiology* (Academic Press).

Mont-Reynaud, B. (**1992**). "Machine hearing research at CCRMA: An overview," in *Center for Computer Research in Music and Acoustics: Research Overview*, 24–32 (Department of Music, Stanford University).

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th edition (Emerald Group Publishing).

Moore, B. C. J. and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," *Hearing Research* **28**, 209–225.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society* **45**, 224–240.

Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (**1990**). "Auditory filter shapes at low center frequencies," *Journal of the Acoustical Society of America* **88**, 132–140.

Moreau, L., Sontag, E., and Arcak, M. (**2003**). "Feedback tuning of bifurcations," *Systems and Control Letters* **50**, 229–239.

Morgan, N., Bourlard, H., and Hermansky, H. (**2004**). "Automatic speech recognition: An auditory perspective," in *Speech Processing in the Auditory System*, edited by S. Greenberg, 309–338 (Springer).

Mountain, D. C. and Hubbard, A. E. (**1996**). "Computational analysis of hair cell and auditory nerve processes," in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, 121–156 (Springer).

Müller, J. (**1838**). *Handbuch der Physiologie des Menschen*, volume 2, 3nd edition (Coblenz: J. Holscher).

Müller, M. and Ewert, S. (**2010**). "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 649–662.

Muncey, R. W. and Nickson, A. F. B. (**1964**). "The listener and room acoustics," *Journal of Sound and Vibration* **1**, 141–147.

Murché, V. T. (**1884**). *Animal Physiology: A Specific Subject of Instruction in Public Elementary Schools* (Glasgow: Blackie).

Naranjo, E. and Baliga, S. (**2009**). "Expanding the use of ultrasonic gas leak detectors: A review of gas release characteristics for adequate detection," *International Gases and Instrumentation* **3**, 24–29.

Narasimhan, S. V. and Veena, S. (**2005**). *Signal Processing: Principles and Implementation* (Alpha Science International Ltd.).

Neely, S. T. and Kim, D. O. (**1983**). "An active cochlear model showing sharp tuning and high sensitivity," *Hearing Research* **9**, 123–130.

Neisser, U. (**1967**). *Cognitive Psychology* (Appleton-Century-Crofts).

Nolle, A. W. (**1948**). "Adjustment speed of automatic-volume-control systems," *Proceedings of the Institute of Radio Engineers* **36**, 911–916.

November, J. A. (**2012**). *Biomedical Computing: Digitizing Life in the United States* (JHU Press).

O'Callaghan, C. (**2007**). *Sounds: A Philosophical Theory* (Oxford University Press).

Oertel, D., Bal, R., Gardner, S. M., Smith, P. H., and Joris, P. X. (**2000**). "Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus," *Proceedings of the National Academy of Sciences* **97**, 11773–11779.

Ohm, G. S. (**1843**). "Ueber die Definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher tonbildenden Vorrichtungen," *Poggendorff's Annalen der Physik und Chemie* **59**, 513–565.

Olshausen, B. A. and Field, D. J. (**2004**). "Sparse coding of sensory inputs," *Current Opinion in Neurobiology* **14**, 481–487.

Oppenheim, A. V. and Schafer, R. W. (**2009**). *Discrete-Time Signal Processing*, 3rd edition (Prentice Hall).

Oppenheim, A. V., Schafer, R. W., and Stockham, T. G. (**1968**). "Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE* **56**, 1264–1291.

Oppenheim, A. V. and Willsky, A. S. (**1997**). *Signals and Systems*, 2nd edition (Prentice Hall).

Orr, G. and Müller, K.-R. (**1998**). *Neural Networks: Tricks of the Trade* (Springer).

O'Shaughnessy, D. (**1987**). *Speech Communication: Human and Machine* (Addison-Wesley).

Ospeck, M., Eguíluz, V. M., and Magnasco, M. O. (**2001**). "Evidence of a Hopf bifurcation in frog hair cells," *Biophysical Journal* **80**, 2597–2607.

Otnes, R. K. and Enochson, L. D. (**1968**). "An algorithm for digital one-third octave analysis," *Sound and Vibration* **2**, 17–20.

Painter, T. and Spanias, A. (**2000**). "Perceptual coding of digital audio," *Proceedings of the IEEE* **88**, 451–515.

Palaz, D., Magimai-Doss, M., and Collobert, R. (**2015**). "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4295–4299.

Paliwal, M. and Kumar, U. A. (**2009**). "Neural networks and statistical techniques: A review of applications," *Expert Systems with Applications* **36**, 2–17.

Palmer, A. R. and Russell, I. J. (**1986**). "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells," *Hearing Research* **24**, 1–15.

Palomäki, K. J., Brown, G. J., and Wang, D. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication* **43**, 361–378.

Papoulis, A. (**1962**). *The Fourier Integral and its Applications* (McGraw Hill).

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (**2012**). "Reconstructing speech from human auditory cortex," *PLoS Biology* **10**, e1001251.

Patterson, R. D. (**1974**). "Auditory filter shape," *Journal of the Acoustical Society of America* **55**, 802–809.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," *Journal of the Acoustical Society of America* **59**, 640–654.

Patterson, R. D. (**2000**). "Auditory images: How complex sounds are represented in the auditory system," *Journal of the Acoustical Society of America* **21**, 183–190.

Patterson, R. D., Dinther, R. van, and Irino, T. (**2007**). "The robustness of bio-acoustic communication and the role of normalization," in *Proceedings of the 19th International Congress on Acoustics*, ppa–07–011 (Curran).

Patterson, R. D., Handel, S., Yost, W. A., and Datta, A. J. (**1996**). "The relative strength of the tone and noise components in iterated rippled noise," *Journal of the Acoustical Society of America* **100**, 3286–3294.

Patterson, R. D. and Holdsworth, J. (**1996**). "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing and Language Processing* **3**, 547–563.

Patterson, R. D. and Irino, T. (**1998**). "Modeling temporal asymmetry in the auditory system," *Journal of the Acoustical Society of America* **104**, 2967–2979.

Patterson, R. D., Ives, D. T., Walters, T. C., and Lyon, R. F. (**2013**). "Modelling the distortion produced by cochlear compression," in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, 81–88 (Springer).

Patterson, R. D. and Moore, B. C. J. (**1986**). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore, 123–177 (Academic Press).

Patterson, R. D. and Nimmo-Smith, I. (**1980**). "Off-frequency listening and auditory-filter asymmetry," *Journal of the Acoustical Society of America* **67**, 229–245.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An efficient auditory filterbank based on the gammatone function," Technical Report, MRC Applied Psychology Unit, SVOS final report, APU report 2341 annex B.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *Journal of the Acoustical Society of America* **72**, 1788–1803.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner, 429–446 (Pergamon Press).

Patterson, R. D., Unoki, M., and Irino, T. (**2003**). "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *Journal of the Acoustical Society of America* **114**, 1529–1542.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., and Griffiths, T. D. (**2002**). "The processing of temporal pitch and melody information in auditory cortex," *Neuron* **36**, 767–776.

Patuzzi, R. (**1996**). "Cochlear micromechanics and macromechanics," in *The Cochlea*, edited by P. Dallos, 186–257 (Springer).

Pearson, K. (**1916**). "Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **216**, 429–457.

Pecharsky, V. K. and Zavalij, P. Y. (**2009**). *Fundamentals of Powder Diffraction and Structural Characterization of Materials*, 2nd edition (Springer).

Percival, G. and Tzanetakis, G. (**2013**). "An effective, simple tempo estimation method based on self-similarity and regularity," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 241–245.

Pérez, J. P. A., Pueyo, S. C., and López, B. C. (**2011**). *Automatic Gain Control: Techniques and Architectures for RF Receivers* (Springer).

Pérot, A. (**1921**). "Sur la sensation d'orientation dans l'audition naturelle," *Le Journal de Physique et le Radium* **2**, 97–106.

Perrett, W. (**1919**). *Some Questions of Phonetic Theory* (Cambridge: W. Heffer & Sons Ltd.).

Perrole, M. (**1797**). "A philosophical memoir, containing—1. experiments relative to the propagation of sound in different solid and fluid mediums.—and 2. an experimental enquiry into the cause of the resonance of musical instruments. with annotations," in *A Journal of Natural Philosophy, Chemistry and the Arts, Volume 1*, edited by W. Nicholson, 411–419 (G. G. and J. Robinson).

Petersen, T. L. and Boll, S. F. (**1983**). "Critical band analysis–synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **31**, 656–663.

Peterson, L. C. and Bogert, B. P. (**1950**). "A dynamical theory of the cochlea," *Journal of the Acoustical Society of America* **22**, 369–381.

Pfafflin, S. M. and Mathews, M. V. (**1962**). "Energy-detection model for monaural auditory detection," *Journal of the Acoustical Society of America* **34**, 1842–1853.

Pfeiffer, R. R. (**1970**). "A model for two-tone inhibition of single cochlear-nerve fibers," *Journal of the Acoustical Society of America* **48**, 1373–1378.

Phillips, D. P. (**2001**). "Introduction to the central audiotory nervous system," in *Physiology of the Ear*, edited by A. F. Jahn and J. Santos-Sacchi, 613–638 (Cengage Learning).

Pichora-Fuller, M. K. and Levitt, H. (**2012**). "Speech comprehension training and auditory and cognitive processing in older adults," *American Journal of Audiology* **21**, 351–357.

Pick, G. F. (**1980**). "Theoretical dependence of cochlear fibre discharge rate versus intensity function on frequency: Evidence for basilar membrane nonlinearity?," *Hearing Research* **2**, 559–564.

Pienkowski, M., Shaw, G., and Eggermont, J. J. (**2009**). "Wiener–Volterra characterization of neurons in primary auditory cortex using Poisson-distributed impulse train inputs," *Journal of Neurophysiology* **101**, 3031–3041.

Pierce, A. H. (**1901**). *Studies in Auditory and Visual Space Perception* (New York: Longmans, Green, and Co.).

Pierce, J. R. (**1991**). "Periodicity and pitch perception," *Journal of the Acoustical Society of America* **90**, 1889–1893.

Pierce, J. R. (**1999**). "The nature of musical sound," in *The Psychology of Music*, edited by D. Deutsch, 2nd edition, 1–24 (Academic Press).

Pierce, J. R. and David, E. E., Jr. (**1958**). *Man's World of Sound* (Doubleday & Co.).

Plinge, A., Hennecke, M. H., and Fink, G. A. (**2010**). "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control*.

Plomp, R. (**1964**). "The ear as a frequency analyzer," *Journal of the Acoustical Society of America* **36**, 1628–1636.

Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff).

Plomp, R. (**1976**). *Aspects of Tone Sensation* (Academic Press).

Plomp, R. (**1988**). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," *Journal of the Acoustical Society of America* **83**, 2322–2327.

Plomp, R. (**2002**). *The Intelligent Ear: On the Nature of Sound Perception* (Lawrence Erlbaum Associates).

Plomp, R., Pols, L. C. W., and van de Geer, J. P. (**1967**). "Dimensional analysis of vowel spectra," *Journal of the Acoustical Society of America* **41**, 707.

Poggio, T., Torre, V., and Koch, C. (**1985**). "Computational vision and regularization theory," *Nature* **317**, 314–319.

Pollak, J. (**1886**). "Über die Function des Musculus tensor tympani," *Wiener medizinische Jahrbuch* **1**, 555–582.

Poole, H. W., Lambert, L., Woodford, C., and Moschovitis, C. J. P. (**2005**). *The Internet: A Historical Encyclopedia* (ABC-CLIO).

Popper, A. N. and Fay, R. R. (**1992**). *The Mammalian Auditory Pathway: Neurophysiology* (Springer).

Potter, R. K., Kopp, G. A., and Green, H. C. (**1947**). *Visible Speech* (D. van Nostrand).

Pressnitzer, D. and Patterson, R. D. (**2001**). "Distortion products and the perceived pitch of harmonic complex tones," in *Physiological and Psychophysical Bases of Auditory Function*, edited by D. Breebaart, A. J. M. Houtsma, A. Kohlrausch, V. Prijs, and R. Schoonhoven, 97–104 (Shaker BV).

Probst, R., Grevers, G., and Iro, H. (**2006**). *Basic Otorhinolaryngology: A Step-by-Step Learning Guide*, 2nd edition (Thieme).

Quackenbos, J. D., Mayer, A. M., and Nipher, F. E. (**1891**). *Appletons' School Physics: Embracing the Results of the Most Recent Researches in the Several Departments of Natural Philosophy* (New York: American Book Company).

Rabiner, L. R. and Schafer, R. W. (**2007**). *Introduction to Digital Speech Processing* (Now Publishers Inc.).

Rabiner, L. R. and Schafer, R. W. (**2010**). *Theory and Applications of Digital Speech Processing* (Prentice Hall).

Ragazzini, J. R. and Zadeh, L. A. (**1952**). "The analysis of sampled-data systems," *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry* **71**, 225–234.

Raj, B. and Stern, R. M. (**2005**). "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine* **22**, 101–116.

Rajan, K. and Bialek, W. (**2013**). "Maximally informative 'stimulus energies' in the analysis of neural responses to natural signals," *PLoS One* **8**, e71959.

Rameau, J. P. (**1722**). *Traité de l'harmonie* (Paris: Jean-Baptiste-Christophe Ballard).

Rameau, J. P. and Gossett, P. (**1971**). *Treatise on Harmony* (Dover Publications).

Rangayyan, R. M. and Lehner, R. J. (**1986**). "Phonocardiogram signal analysis: A review," *Critical Reviews in Biomedical Engineering* **15**, 211–236.

Ranke, O. F. (**1931**). *Die Gleichrichter-Resonanztheorie: eine Erweiterung der Helmholtzschen Resonanztheorie des Gehörs durch physikalische Untersuchung der Flüssigkeitsschwingungen in der Cochlea (The Rectifier Resonance Theory: An Extension of Helmholtz's Resonance Theory of Hearing by Physical Examination of the Fluid Vibrations in the Cochlea)* (Munich: J. F. Lehmanns Verlag).

Ranke, O. F. (**1950**). "Theory of operation of the cochlea: A contribution to the hydrodynamics of the cochlea," *Journal of the Acoustical Society of America* **22**, 772–777.

Ranson, S. W. (**1920**). *The Anatomy of the Nervous System: From the Standpoint of Development and Function* (W. B. Saunders Co.).

Rasetshwane, D. M., Gorga, M. P., and Neely, S. T. (**2014**). "Signal-processing strategy for restoration of cross-channel suppression in hearing-impaired listeners," *IEEE Transactions on Biomedical Engineering* **61**, 64–75.

Rauschecker, J. P. and Scott, S. K. (**2009**). "Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing," *Nature Neuroscience* **12**, 718–724.

Ravuri, S. and Ellis, D. P. W. (**2010**). "Cover song detection: From high scores to general classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 65–68.

Recio-Spinoso, A., Fan, Y.-H., and Ruggero, M. A. (**2011**). "Basilar-membrane responses to broadband noise modeled using linear filters with rational transfer functions," *IEEE Transactions on Biomedical Engineering* **58**, 1456–1465.

Recio-Spinoso, A., Narayan, S. S., and Ruggero, M. A. (**2009**). "Basilar membrane responses to noise at a basal site of the chinchilla cochlea: Quasi-linear filtering," *Journal of the Association for Research in Otolaryngology* **10**, 471–484.

Redheffer, R. M. (**1991**). *Differential Equations: Theory and Applications* (Jones and Bartlett).

Reichenbach, T. and Hudspeth, A. J. (**2014**). "The physics of hearing: Fluid mechanics and the active process of the inner ear," *Reports on Progress in Physics* **77**, 076601.

Reiss, L. A. J. and Young, E. D. (**2005**). "Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus," *The Journal of Neuroscience* **25**, 3680–3691.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," *Science* **212**, 947–949.

Ren, T., He, W., and Porsov, E. (**2011**). "Localization of the cochlear amplifier in living sensitive ears," *PLoS One* **6**.

Retzius, G. (**1884**). *Das Gehörorgan der Reptilien, der Vögel und die Saugethiere* (Stockholm: Samson and Wallin).

Reynolds, D. A. and Torres-Carrasquillo, P. (**2005**). "Approaches and applications of audio diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, 953–956.

Rhode, W. S. (**1971**). "Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique," *Journal of the Acoustical Society of America* **49**, 1218–1231.

Rhode, W. S. (**1978**). "Some observations on cochlear mechanics," *Journal of the Acoustical Society of America* **64**, 158–176.

Rhode, W. S. and Greenberg, S. (**1992**). "Physiology of the cochlear nuclei," in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay, 94–152 (Springer).

Rhode, W. S. and Greenberg, S. (**1994**). "Lateral suppression and inhibition in the cochlear nucleus of the cat," *Journal of Neurophysiology* **71**, 493–514.

Riesenhuber, M. and Poggio, T. (**2000**). "Models of object recognition," *Nature Neuroscience* **3**, 1199–1204.

Rifkin, R. and Klautau, A. (**2004**). "In defense of one-vs-all classification," *Journal of Machine Learning Research (JMLR)* **5**, 101–141.

Rifkin, R. and Lippert, R. A. (**2007**). "Notes on Regularized Least-Squares," Technical Report, Massachusetts Institute of Technology.

Rijntjes, M., Weiller, C., Bormann, T., and Musso, M. (**2012**). "The dual loop model: Its relation to language and other modalities," *Frontiers in Evolutionary Neuroscience* **4**.

Ritsma, R. J. (**1970**). "Periodicity detection," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff).

Roads, C. (**1996**). *The Computer Music Tutorial* (MIT Press).

Roaf, H. E. (**1922**). "The analysis of sound waves by the cochlea," *Philosophical Magazine Series 6* **43**, 349–354.

Robert, A. and Eriksson, J. L. (**1999**). "A composite model of the auditory periphery for simulating responses to complex sounds," *Journal of the Acoustical Society of America* **106**, 1852–1864.

Roberts, W. M. and Rutherford, M. A. (**2008**). "Linear and nonlinear processing in hair cells," *Journal of Experimental Biology* **211**, 1775–1780.

Robinson, D. W. and Dadson, R. S. (**1956**). "A re-determination of the equal-loudness relations for pure tones," *British Journal of Applied Physics* **7**, 166–181.

Robles, L. and Ruggero, M. A. (**2001**a). "Mechanics of the mammalian cochlea," *Physiological Reviews* **81**, 1305–1352.

Robles, L. and Ruggero, M. A. (**2001**b). "Mechanics of the mammalian cochlea," *Physiological Reviews* **81**, 1305–1352.

Rodríguez, J., Neely, S. T., Patra, H., Kopun, J., Jesteadt, W., Tan, H., and Gorga, M. P. (**2010**). "The role of suppression in psychophysical tone-on-tone masking," *Journal of the Acoustical Society of America* **127**, 361–369.

Roe, A. W., Pallas, S. L., Hahm, J.-O., and Sur, M. (**1990**). "A map of visual space induced in primary auditory cortex," *Science* **250**, 818–820.

Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K.-D., and Kollmeier, B. (**2008**). "Combined source tracking and noise reduction for application in hearing aids," in *ITG Conference on Voice Communication (SprachKommunikation)*, 1–4 (VDE).

Roma, G., Nogueira, W., and Herrera, P. (**2013**). "Recurrence quantification analysis features for environmental sound recognition," in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, 1–4.

Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," *Journal of the Acoustical Society of America* **114**, 2236–2252.

Rose, A. (**1948**). "The sensitivity performance of the human eye on an absolute scale," *Journal of the Optical Society of America* **38**, 196–208.

Rose, A. (**1973**). *Vision: Human and Electronic* (Plenum Press).

Rose, J. E., Brugge, J. F., Anderson, D. J., and Hind, J. E. (**1967**). "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey," *Journal of Neurophysiology* **30**, 769–793.

Rose, J. E., Hind, J. E., Anderson, D. J., and Brugge, J. F. (**1971**). "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey," *Journal of Neurophysiology* **34**, 685–699.

Rosen, S. and Baker, R. J. (**1994**). "Characterising auditory filter nonlinearity," *Hearing Research* **73**, 231–243.

Rosen, S., Baker, R. J., and Darling, A. (**1998**). "Auditory filter nonlinearity at 2 kHz in normal hearing listeners," *Journal of the Acoustical Society of America* **103**, 2539–2550.

Rosen, S. and Howell, P. (**2011**). *Signals and Systems for Speech and Hearing*, 2nd edition (Emerald Group).

Rosenberg, A. E. (**1965**). "Effect of masking on the pitch of periodic pulses," *Journal of the Acoustical Society of America* **38**, 747–758.

Rosenblatt, F. (**1957**). "The perceptron: A perceiving and recognition automaton," Technical Report, Cornell Aeronautical Laboratory.

Rosenthal, D. F. and Okuno, H. G. (**1998**). *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates).

Rossing, T. D. (**2007**). "A brief history of acoustics," in *Springer Handbook of Acoustics*, edited by T. D. Rossing, 9–23 (Springer).

Roush, J. (**2001**). *Screening for Hearing Loss and Otitis Media in Children* (Thomson Learning).

Rubin, H. and Atkinson, J. F. (**2001**). *Environmental Fluid Mechanics* (CRC Press).

Ruggero, M. A. (**1992**). "Responses to sound of the basilar membrane of the mammalian cochlea," *Current Opinion in Neurobiology* **2**, 449–456.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (**1997**). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," *Journal of the Acoustical Society of America* **101**, 2151–2163.

Ruggero, M. A., Robles, L., and Rich, N. C. (**1992**). "Two-tone suppression in the basilar membrane of the cochlea: Mechanical basis of auditory-nerve rate suppression," *Journal of Neurophysiology* **68**, 1087–1099.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (**1986**). "Learning internal representations by error backpropagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, edited by D. E. Rumelhart and J. L. McClelland, 318–362 (MIT Press).

Rumelhart, D. E. and McClelland, J. L. (**1987**). *Parallel Distributed Processing*, volume 1 Foundations (MIT Press).

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (**2005**). "Spatiotemporal elements of macaque V1 receptive fields," *Neuron* **46**, 945–956.

Rutherford, W. (**1887**). "A lecture on the sense of hearing: Delivered before the British Association at Birmingham on September 6th, 1886," *Lancet* **I**, 2–6.

Sainath, T. N., Kingsbury, B., Mohamed, A.-R., and Ramabhadran, B. (**2013**). "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 297–302.

Sakaguchi, H., Tokita, J., Müller, U., and Kachar, B. (**2009**). "Tip links in hair cells: Molecular composition and role in hearing loss," *Current Opinion in Otolaryngology and Head and Neck Surgery* **17**, 388–393.

Samuel, A. G. (**1981**). "Phonemic restoration: Insights from a new methodology," *Journal of Experimental Psychology: General* **110**, 474.

Sarpeshkar, R. (**2000**). "Traveling waves versus bandpass filters: The silicon and biological cochlea," in *Proceedings of the International Symposium on Recent Developments in Auditory Mechanics*, edited by H. W. et al., 216–222.

Schafer, T. H., Gales, R. S., Shewmaker, C. A., and Thompson, P. O. (**1950**). "The frequency selectivity of the ear as determined by masking experiments," *Journal of the Acoustical Society of America* **22**, 490.

Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, Rijksuniversiteit Groningen.

Scherer, P. (**1959**). "Uber die Ortungsmöglichkeit verschiedener stereophonischer Aufnahmeverfahren (On the ability of various stereophonic sound-collection processes to define position)," *Nachrichtentechnische Fachberichte* **15**, 36–42.

Schiffman, H. R. (**1990**). *Sensation and Perception: An Integrated Approach*, 3rd edition (John Wiley & Sons).

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Dosch, H. G., Bleeck, S., Stippich, C., and Rupp, A. (**2005**). "Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference," *Nature Neuroscience* **8**, 1241–1247.

Schneider, T. and Brennan, R. (**1997**). "A multichannel compression strategy for a digital hearing aid," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 411–414.

Schofield, B. R. (**2010**). "Structural organization of the descending auditory pathway," in *The Oxford Handbook of Auditory Science: The Auditory Brain*, edited by A. Rees and A. R. Palmer, 43–64 (Oxford University Press).

Schofield, D. (**1985**). "Visualisations of speech based on a model of the peripheral auditory system," Technical Report, UK National Physical Laboratory.

Schouten, J. F. (**1970**). "The residue revisited," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff).

Schreiner, C. E. (**1991**). "Functional topographies in the primary auditory cortex of the cat," *Acta Oto-Laryngologica* **111**, 7–16.

Schreiner, C. E. and Winer, J. A. (**2007**). "Auditory cortex mapmaking: Principles, projections, and plasticity," *Neuron* **56**, 356–365.

Schroeder, M. R. (**1973**). "An integrable model for the basilar membrane," *Journal of the Acoustical Society of America* **53**, 429–434.

Schroeder, M. R. (**1975**). "Amplitude behavior of the cubic difference tone," *Journal of the Acoustical Society of America* **58**, 728–732.

Schroeder, M. R. (**2004**). *Computer Speech: Recognition, Compression, Synthesis*, 2nd edition (Springer).

Schroeder, M. R. and Atal, B. S. (**1962**). "Generalized short-time power spectra and autocorrelation functions," *Journal of the Acoustical Society of America* **34**, 1679–1683.

Schroeder, M. R. and Hall, J. L. (**1974**). "A model for mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America* **55**, 1055–1060.

Schulze, H., Hess, A., Ohl, F. W., and Scheich, H. (**2002**). "Superposition of horseshoe-like periodicity and linear tonotopic maps in auditory cortex of the Mongolian gerbil," *European Journal of Neuroscience* **15**, 1077–1084.

Schwartz, O. and Simoncelli, E. P. (**2001**). "Natural signal statistics and sensory gain control," *Nature Neuroscience* **4**, 819–825.

Schwede, G. W. (**1983**). "An algorithm and architecture for constant-Q spectrum analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1384–1387.

Scripture, E. W. (**1902**). *The Elements of Experimental Phonetics* (New York: C. Scribner's Sons).

Sebastiani, F. (**2002**). "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)* **34**, 1–47.

Seebeck, A. (**1841**). "Beobachtungen über einige Bedingungen der Entstehung von Tönen," *Annalen der Physik und Chemie* **53**, 417–436.

Seltzer, M. L., Raj, B., and Stern, R. M. (**2000**). "Classifier-based mask estimation for missing feature methods of robust speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 538–541.

Serrà Julià, J. (**2011**). "Identification of versions of the same musical composition by processing audio descriptions," Ph.D. thesis, Universitat Pompeu Fabra.

Serre, T., Wolf, L., and Poggio, T. (**2005**). "Object recognition with features inspired by visual cortex," in *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 994–1000.

Sethares, W. A. (**2005**). *Tuning, Timbre, Spectrum, Scale* (Springer).

Sewell, W. F. (**1984**). "The effects of furosemide on the endocochlear potential and auditory-nerve fiber tuning curves in cats," *Hearing Research* **14**, 305–314.

Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1992**). "Across frequency integration in a model of lateralization," *Journal of the Acoustical Society of America* **91**, 2276–2279.

Shambaugh, G. E. (**1910**). "The physiology of the cochlea: Sammelreferat," *The Annals of Otology, Rhinology and Laryngology* **19**, 618–630.

Shamma, S. A. (**1985**). "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *Journal of the Acoustical Society of America* **78**, 1622–1632.

Shamma, S. A. (**2003**). "Encoding sound timbre in the auditory system," *IETE Journal of Research* **49**, 145–156.

Shanmugam, K. S. (**1975**). "Comments on 'Discrete cosine transform'," *IEEE Transactions on Computers* **100**, 759–759.

Shannon, C. E. and Weaver, W. (**1948**). "The mathematical theory of communication," *Bell System Technical Journal* **27**, 379–423.

Sharma, J., Angelucci, A., and Sur, M. (**2000**). "Induction of visual orientation modules in auditory cortex," *Nature* **404**, 841–847.

Shen, J. and Dietterich, T. G. (**2009**). "A family of large margin linear classifiers and its application in dynamic environments," *Statistical Analysis and Data Mining* **2**, 328–345.

Shepard, R. N. (**1964**). "Circularity in judgments of relative pitch," *Journal of the Acoustical Society of America* **36**, 2346–2353.

Shepherd, W. T. (**1911**). "The discrimination of articulate sounds by raccoons," *The American Journal of Psychology* **22**, 116–119.

Shera, C. A. (**2001**). "Intensity-invariance of fine time structure in basilar-membrane click responses: Implications for cochlear mechanics," *Journal of the Acoustical Society of America* **110**, 332–348.

Shera, C. A. (**2007**). "Laser amplification with a twist: Traveling-wave propagation and gain functions from throughout the cochlea," *Journal of the Acoustical Society of America* **122**, 2738–2758.

Shera, C. A. and Zweig, G. (**1991**). "A symmetry suppresses the cochlear catastrophe," *Journal of the Acoustical Society of America* **89**, 1276.

Siebert, W. M. (**1974**). "Ranke revisited—a simple short-wave cochlear model," *Journal of the Acoustical Society of America* **56**, 594–600.

Siebert, W. M. (**1986**). *Circuits, Signals, and Systems* (MIT Press).

Simmons, J. A. and Simmons, A. M. (**2011**). "Bats and frogs and animals in between: Evidence for a common central timing mechanism to extract periodicity pitch," *Journal of Comparative Physiology A* **197**, 585–594.

Simon, H. A. (**1981**). *The Sciences of the Artificial* (MIT Press).

Slaney, M. (**1993**). "An efficient implementation of the Patterson–Holdsworth auditory filter bank," Technical Report 35, Apple Computer.

Slaney, M. (**1998**). "A critique of pure audition," in *Computational Auditory Scene Analysis*, edited by D. F. Rosenthal and H. G. Okuno, 27–41 (Lawrence Erlbaum Associates).

Slaney, M. (**2002**). "Semantic-audio retrieval," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, 4108–4111 (IEEE).

Slaney, M. (**2005**). "The history and future of CASA," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, 199–211 (Kluwer Academic Publishers).

Slaney, M. and Lyon, R. F. (**1990**). "A perceptual pitch detector," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, S6b.3 (V.1) 357–360.

Slaney, M. and Lyon, R. F. (**1993**). "On the importance of time—a temporal representation of sound," in *Visual Representations of Speech Signals*, edited by M. Cooke, S. Beet, and M. Crawford, 95–116 (John Wiley & Sons).

Slaney, M., Naar, D., and Lyon, R. F. (**1994**). "Auditory model inversion for sound separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, 77–80.

Smith, E. C. and Lewicki, M. S. (**2006**). "Efficient auditory coding," *Nature* **439**, 978–982.

Smith, J. O. (**2007**). *Introduction to Digital Filters: With Audio Applications* (W3K Publishing).

Smith, L. S. and Collins, S. (**2007**). "Determining itds using two microphones on a flat panel during onset intervals with a biologically inspired spike-based technique," *IEEE Transactions on Audio, Speech, and Language Processing* **15**, 2278–2286.

Smith, R. L. and Zwislocki, J. J. (**1975**). "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biological Cybernetics* **17**, 169–182.

Smoorenburg, G. F. (**1970**). "Pitch of two-tone complexes," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff).

Smoorenburg, G. F. (**1972**). "Combination tones and their origin," *Journal of the Acoustical Society of America* **52**, 615–632.

Sokolovski, A. (**1974**). "Minimum audible field curve for the cat (monoaural) compared to the minimum audible field curve for man (binaural)," *International Journal of Audiology* **13**, 432–436.

Sorge, G. A. (**1745**). *Vorgemach der musicalischen Composition, oder: Ausführliche, ordentliche und zur heutigen Praxis hinlänglichen Anweisung zum General-Baß* (Lobenstein: im Verlag des Autoris).

Sra, S., Nowozin, S., and Wright, S. J. (**2011**). *Optimization for Machine Learning* (MIT Press).

Srinivasan, S., Roman, N., and Wang, D. (**2006**). "Binary and ratio time–frequency masks for robust speech recognition," *Speech Communication* **48**, 1486–1501.

Steele, C. R. and Taber, L. A. (**1979**). "Comparison of WKB calculations and experimental results for three-dimensional cochlear models," *Journal of the Acoustical Society of America* **65**, 1007–1018.

Steinberg, J. C. and French, N. R. (**1946**). "The portrayal of visible speech," *Journal of the Acoustical Society of America* **18**, 4–18.

Steinmetz, C. P. (**1893**). "Complex quantities and their use in electrical engineering," *AIEE Proceedings of the International Electrical Congress, Chicago* 33–74.

Steinmetz, C. P. (**1910**). "Transmission line equations: Calculation of the electrical constants of transmission circuits," *Electrical World* **LV**, 1653–1654.

Stern, R. M. and Trahiotis, C. (**1995**). "Models of binaural interaction," *Handbook of Perception and Cognition* **6**, 347–386.

Stevens, S. S. (**1936**). "A scale for the measurement of a psychological magnitude: Loudness," *Psychological Review* **43**, 405–416.

Stevens, S. S. (**1961**). "To honor Fechner and repeal his law: A power function, not a log function, describes the operating characteristic of a sensory system," *Science* **133**, 80–86.

Stevens, S. S., Egan, J. P., and Miller, G. A. (**1947**). "Methods of measuring speech spectra," *Journal of the Acoustical Society of America* **19**, 771–780.

Stevens, S. S. and Newman, E. B. (**1936**). "The localization of actual sources of sound," *The American Journal of Psychology* **48**, 297–306.

Stevens, S. S. and Volkmann, J. (**1940**). "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology* **53**, 329–353.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America* **8**, 185–190.

Stewart, G. N. (**1901**). "Theories of hearing," in *American Year-Book of Medicine and Surgery*, volume 6, 561–562 (New York: W. B. Saunders).

Stewart, J. L. (**1963**). "Limits to animal discrimination and recognition in a noise-free external environment," *IEEE Transactions on Military Electronics* **7**, 116–131.

Stewart, J. L. (**1966**). "Neural-like analyzing system," US patent 3469034.

Stewart, J. L. (**1967**). "Status of the Santa Rita Technology analog model," Technical Report, Santa Rita Technology Inc.

Stewart, J. L. (**1979**). *The Bionic Ear* (Covox Co.).

Stillingfleet, B. (**1771**). *Principles of Power and Harmony* (London: J. and H. Hughes).

Stillwell, J. (**2010**). *Mathematics and Its History*, 3rd edition (Springer).

Strube, H. W. (**1985**). "A computationally efficient basilar-membrane model," *Acustica* **58**, 207–214.

Strutt, J. W. (Baron Rayleigh) (**1876**). "On our perception of the direction of a source of sound," *Proceedings of the Musical Association* **2**, 75–84.

Strutt, J. W. (Baron Rayleigh) (**1877**). "Acoustical observations: Perception of the direction of a source of sound," *Philosophical Magazine Series 5* **3**, 456–458.

Strutt, J. W. (Baron Rayleigh) (**1878**). *The Theory of Sound, Vol. II* (London: Macmillan and Co.).

Strutt, J. W. (Baron Rayleigh) (**1907**). "On our perception of sound direction," *Philosophical Magazine Series 6* **13**, 214–232.

Suga, N. (**2008**). "Role of corticofugal feedback in hearing," *Journal of Comparative Physiology A* **194**, 169–183.

Suga, N., Gao, E., Zhang, Y., Ma, X., and Olsen, J. F. (**2000**). "The corticofugal system for hearing: Recent progress," *Proceedings of the National Academy of Sciences* **97**, 11807–11814.

Suga, N., Ma, X., Gao, E., Sakai, M., and Chowdhury, S. A. (**2003**). "Descending system and plasticity for auditory signal processing: Neuroethological data for speech scientists," *Speech Communication* **41**, 189–200.

Sukthankar, R., Ke, Y., and Hoiem, D. (**2006**). "Semantic learning for audio applications: A computer vision approach," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 112.

Sullivan, W. E. and Konishi, M. (**1986**). "Neural map of interaural phase difference in the owl's brainstem," *Proceedings of the National Academy of Sciences* **83**, 8400–8404.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., and Meddis, R. (**2002**). "A revised model of the inner-hair cell and auditory-nerve complex," *Journal of the Acoustical Society of America* **111**, 2178–2188.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., and Meddis, R. (**2003**a). "Adaptation in a revised inner-hair cell model," *Journal of the Acoustical Society of America* **113**, 893–901.

Sumner, C. J., O'Mard, L. P., Lopez-Poveda, E. A., and Meddis, R. (**2003**b). "A nonlinear filter-bank model of the guinea-pig cochlear nerve: Rate responses," *Journal of the Acoustical Society of America* **113**, 3264–3274.

Sung, P.-H., Wang, J.-N., Chen, B.-W., Jang, L.-S., and Wang, J.-F. (**2013**). "Auditory-inspired heart sound temporal analysis for patent ductus arteriosus," in *International Conference on Orange Technologies (ICOT)*, 231–234 (IEEE).

Swerup, C. (**1978**). "On the choice of noise for the analysis of the peripheral auditory system," *Biological Cybernetics* **29**, 97–104.

Takeno, S., Harrison, R. V., Mount, R. J., Wake, M., and Harada, Y. (**1994**). "Induction of selective inner hair cell damage by carboplatin," *Scanning Microscopy* **8**, 97–106.

Tan, Q. and Carney, L. H. (**2003**). "A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide," *Journal of the Acoustical Society of America* **114**, 2007–2020.

Tanner, W. P., Swets, J. A., and Green, D. M. (**1956**). "Some general properties of the hearing mechanism," Technical Report No. 30, Electronic Defense Group, Department of Electrical Engineering, University of Michigan.

Tartini, G. (**1754**). *Trattato di musica secondo la vera scienza dell' armonia* (Padua).

Tchorz, J. and Kollmeier, B. (**1999**). "A psychoacoustical model of the auditory periphery as front-end for ASR," in *ASA-EAA-DEGA Joint Meeting on Acoustics*.

Temchin, A. N., Recio-Spinoso, A., and Ruggero, M. A. (**2011**). "Timing of cochlear responses inferred from frequency-threshold tuning curves of auditory-nerve fibers," *Hearing Research* **272**, 178–186.

Terasawa, H., Slaney, M., and Berger, J. (**2006**). "A statistical model of timbre perception," in *ITRW on Statistical and Perceptual Audio Processing*, 18–23 (International Speech Communication Association).

Terhardt, E. (**1970**). "Frequency analysis and periodicity detection in the sensation of roughness and periodicity pitch," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff).

Terhardt, E. (**1974**). "Pitch, consonance, and harmony," *Journal of the Acoustical Society of America* **55**, 1061–1069.

Terman, F. E. (**1932**). *Radio Engineering* (McGraw-Hill).

Testut, L. (**1897**). *Traité d'anatomie humaine : anatomie descriptive, histologie, développement*, volume 2, 3rd edition (Paris: Octave Doin).

Thiers, F. A., Nadol, J. B., Jr., and Liberman, M. C. (**2008**). "Reciprocal synapses between outer hair cells and their afferent terminals: Evidence for a local neural network in the mammalian cochlea," *Journal of the Association for Research in Otolaryngology* **9**, 477–489.

Thompson, S. P. (**1877**). "On binaural audition," *Philosophical Magazine Series 5* **4**, 274–276.

Todd, C. C., Davidson, G. A., Davis, M. F., Fielder, L. D., Link, B. D., and Vernon, S. (**1994**). "AC-3: Flexible perceptual coding for audio transmission and storage," in *Audio Engineering Society Convention 96*.

Todd, N. P. M. (**1994**). "The auditory 'primal sketch': A multiscale model of rhythmic grouping," *Journal of New Music Research* **23**, 25–70.

Tollin, D. J. (**1998**). "Computational model of the lateralisation of clicks and their echoes," in *Proceedings of the NATO Advanced Study Institute on Computational Hearing*, edited by S. Greenberg and M. Slaney, 77–82 (NATO).

Tollin, D. J. and Yin, T. C. T. (**2005**). "Interaural phase and level difference sensitivity in low-frequency neurons in the lateral superior olive," *The Journal of Neuroscience* **25**, 10648–10657.

Tolonen, T. and Karjalainen, M. (**2000**). "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing* **8**, 708–716.

Trahiotis, C. and Robinson, D. E. (**1979**). "Auditory psychophysics," *Annual Review of Psychology* **30**, 31–61.

Trautwein, P., Hofstetter, P., Wang, J., Salvi, R., and Nostrant, A. (**1996**). "Selective inner hair cell loss does not alter distortion product otoacoustic emissions," *Hearing Research* **96**, 71–82.

Treisman, A. (**1964**). "Monitoring and storage of irrelevant messages in selective attention," *Journal of Verbal Learning and Verbal Behavior* **3**, 449–459.

Troland, L. T. (**1929**). "The psychophysiology of auditory qualities and attributes," *Journal of General Psychology* **2**, 28–58.

Troland, L. T. (**1930**). *The Principles of Psychophysiology: A Survey of Modern Scientific Psychology*, volume 2 (D. Van Nostrand Co.).

Trotter, C. (**1878**). "Note on 'Fechner's law'," *The Journal of Physiology* **1**, 60–65.

Truong, K. N. and Hayes, G. R. (**2009**). "Ubiquitous computing for capture and access," *Foundations and Trends in Human–Computer Interaction* **2**, 95–171.

Tsai, W.-H., Yu, H.-M., and Wang, H.-M. (**2008**). "Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval," *Journal of Information Science and Engineering* **24**, 1669–1687.

Tucker, D. G. (**1946**). "Transient response of tuned-circuit cascades," *Wireless Engineer* **23**, 250–258.

Tukey, J. W. (**1957**). "On the comparative anatomy of transformations," *The Annals of Mathematical Statistics* 602–632.

Turian, J., Bergstra, J., and Bengio, Y. (**2009**). "Quadratic features and deep architectures for chunking," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 245–248.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (**2008**). "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing* **16**, 467.

Turnbull, L. (**1887**). *A Clinical Manual of the Diseases of the Ear* (Philadelphia: Lippincott).

Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (**2009**). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," *Journal of the Acoustical Society of America* **125**, 2374–2386.

Tyndall, J. (**1867**). *Sound* (New York: Appleton).

Unoki, M., Irino, T., Glasberg, B. R., Moore, B. C. J., and Patterson, R. D. (**2006**). "Comparison of the roex and gammachirp filters as representations of the auditory filter," *Journal of the Acoustical Society of America* **120**, 1474–1492.

Unoki, M., Irino, T., and Patterson, R. D. (**2001**). "Improvement of an IIR asymmetric compensation gammachirp filter," *Acoustical Science and Technology* **22**, 426–430.

Upton, H. (**1968**). "Proceedings of the conference on speech-analyzing aids for the deaf: Wearable eyeglass speechreading aid," *American Annals of the Deaf* **113**, 116–330.

Van Compernolle, D. (**1991**). "Development of a computational auditory model," Technical Report IPO no. 784, Institute for Perception Research, Eindhoven.

Van De Water, T. R. and Staecker, H. (**2006**). *Otolaryngology: Basic Science and Clinical Review* (Thieme).

van den Berg, E. and Friedlander, M. P. (**2008**). "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing* **31**, 890–912.

van den Raadt, M. P. M. G. and Duifhuis, H. (**1990**). "A generalized Van der Pol-oscillator cochlea model," in *The Mechanics and Biophysics of Hearing*, edited by P. Dallos, C. D. Geisler, J. W. Matthews, M. A. Ruggero, and C. R. Steele, 227–234 (Springer).

van der Heijden, M. (**2005**). "Cochlear gain control," *Journal of the Acoustical Society of America* **117**, 1223–1233.

van der Heijden, M. (**2014**). "Frequency selectivity without resonance in a fluid waveguide," *Proceedings of the National Academy of Sciences* **111**, 14548–14552.

van der Heijden, M. and Joris, P. X. (**2003**). "Cochlear phase and amplitude retrieved from the auditory nerve at arbitrary frequencies," *The Journal of Neuroscience* **23**, 9194–9198.

van der Heijden, M. and Versteegh, C. P. C. (**2015**). "Energy flux in the cochlea: Evidence against power amplification of the traveling wave," *Journal of the Association for Research in Otolaryngology* **16**, 581–597.

van der Pol, B. (**1926**). "On 'relaxation-oscillations'," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 978–992.

Van Immerseel, L. and Peeters, S. (**2003**). "Digital implementation of linear gammatone filters: Comparison of design methods," *Acoustics Research Letters Online* **4**, 59–64.

van Netten, S. M. and Duifhuis, H. (**1983**). "Modelling an active, nonlinear cochlea," in *Mechanics of Hearing*, 143–151 (Springer).

van Schaik, A., Fragnière, E., and Vittoz, E. (**1996**). "Improved silicon cochlea using compatible lateral bipolar transistors," *Advances in Neural Information Processing Systems* 671–677.

Vaseghi, S. V. (**2007**). *Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications* (John Wiley & Sons).

Velenovsky, D. S., Cetas, J. S., Price, R. O., Sinex, D. G., and McMullen, N. T. (**2003**). "Functional subregions in primary auditory cortex defined by thalamocortical terminal arbors: An electrophysiological and anterograde labeling study," *The Journal of Neuroscience* **23**, 308–316.

Verhulst, P. F. (**1845**). "Recherches mathématiques sur la loi d'accroissement de la population," *Nouveaux mémoires de l'académie royale des sciences et belles-lettres de Bruxelles* **18**, 1–38.

Verhulst, S., Bianchi, F., and Dau, T. (**2013**). "Cochlear contributions to the precedence effect," in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, 283–291 (Springer).

Versteegh, C. P. C. and van der Heijden, M. (**2013**). "The spatial buildup of compression and suppression in the mammalian cochlea," *Journal of the Association for Research in Otolaryngology* **14**, 523–545.

Vetter, P., Smith, F. W., and Muckli, L. (**2014**). "Decoding sound and imagery content in early visual cortex," *Current Biology* **24**, 1256–1262.

Villchur, E. (**1973**). "Signal processing to improve speech intelligibility in perceptive deafness," *Journal of the Acoustical Society of America* **53**, 1646–1657.

Villchur, E. (**1974**). "Simulation of the effect of recruitment on loudness relationships in speech," *Journal of the Acoustical Society of America* **56**, 1601–1611.

Wainwright, M. J., Schwartz, O., and Simoncelli, E. P. (**2002**). "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," in *Probabilistic Models of the Brain: Perception and Neural Function*, edited by R. Rao and B. A. Olshausen, 203–222 (MIT Press).

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," *The American Journal of Psychology* **62**, 315–336.

Walters, T. C. (**2011**). "Auditory-based processing of communication sounds," Ph.D. thesis, University of Cambridge.

Walters, T. C., Ross, D. A., and Lyon, R. F. (**2013**). "The intervalgram: An audio feature for large-scale cover-song recognition," in *From Sounds to Music and Emotions*, 197–213 (Springer).

Wang, A. (**2003**). "An industrial strength audio search algorithm," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*.

Wang, A. (**2006**). "The Shazam music recognition service," *Communications of the ACM* **49**, 44–48.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, 181–197 (Springer).

Wang, D. and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley Interscience).

Wang, K. and Shamma, S. A. (**1994**). "Self-normalization and noise-robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing* **2**, 421–435.

Wang, Y., Narayanan, A., and Wang, D. (**2014**). "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing* **22**, 1849–1858.

Ward, W. D. (**1970**). "Musical perception," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias, volume 1, 407–447 (Academic Press).

Warr, W. B. (**1992**). "Organization of olivocochlear efferent systems in mammals," in *The Mammalian Auditory Pathway: Neuroanatomy*, edited by D. B. Webster, A. N. Popper, and R. R. Fay, 410–448 (Springer).

Warren, E. H. and Liberman, M. C. (**1989**). "Effects of contralateral sound on auditory-nerve responses. I. contributions of cochlear efferents," *Hearing Research* **37**, 89–104.

Watt, H. J. (**1920**). "A theory of binaural hearing," *The British Journal of Psychology* **11**, 162–171.

Watts, L. (**2000**). "The mode-coupling Liouville–Green approximation for a two-dimensional cochlear model," *Journal of the Acoustical Society of America* **108**, 2266–2271.

Watts, L. (**2010**). "Real-time, high-resolution simulation of the auditory pathway, with application to cell-phone noise reduction," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 3821–3824.

Watts, L. (**2012**). "Reverse-engineering the human auditory pathway," in *Advances in Computational Intelligence*, edited by J. Liu, C. Alippi, B. Bouchon-Meunier, G. W. Greenwood, and H. A. Abbass, 47–59 (Springer).

Watts, L., Kerns, D. A., Lyon, R. F., and Mead, C. A. (**1992**). "Improved implementation of the silicon cochlea," *IEEE Journal of Solid State Circuits* **27**, 692–700.

Watts, L., Lyon, R. F., and Mead, C. A. (**1991**). "A bidirectional analog VLSI cochlear model," in *Advanced Research in VLSI*, edited by C. Sequin, 153–163 (MIT Press).

Webster, J. C., Miller, P. H., Thompson, P. O., and Davenport, E. W. (**1952**). "The masking and pitch shift of pure tones near abrupt changes in a thermal noise spectrum," *Journal of the Acoustical Society of America* **24**, 147–152.

Wegel, R. L. and Lane, C. E. (**1924**). "The auditory masking of one sound by another and its probable relation to the dynamics of the inner ear," *Physical Review* **23**, 266–285.

Weintraub, M. (**1984**). "The GRASP sound separation system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 18A.6.1–18A.6.4.

Weintraub, M. (**1987**). "Sound separation and auditory perceptual organization," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten, 125–134 (Springer).

Weiss, K. M. and Buchanan, A. (**2004**). *Genetics and the Logic of Evolution* (Wiley-IEEE).

Wen, B. and Boahen, K. (**2012**). "A biomorphic active cochlear model in silico," in *Integrated Microsystems: Electronics, Photonics, and Biotechnology*, edited by K. Iniewski, 207–235 (CRC Press).

Wenzel, E. M. (**1992**). "Three-dimensional virtual acoustic displays," in *Multimedia Interface Design*, edited by M. M. Blattner and R. B. Dannenberg, 257–288 (ACM Press).

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**). "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America* **94**, 111–111.

Werner, S., Harczos, T., and Brandenburg, K. (**2010**). "Overview of numerical models of cell types in the cochlear nucleus," in *Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing*, edited by J. M. Buchholtz, T. Dau, J. Christensen-Dalsgaard, and T. Poulsen, 61–70 (Danavox Jubilee Foundation).

Weston, J., Bengio, S., and Hamel, P. (**2011**). "Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval," *Journal of New Music Research* **40**, 337–348.

Wever, E. G. (**1949**). *A Theory of Hearing* (John Wiley & Sons).

Wever, E. G. (**1962**). "Development of traveling-wave theories," *Journal of the Acoustical Society of America* **34**, 1319–1324.

Wever, E. G. and Bray, C. W. (**1930**a). "Action currents in the auditory nerve in response to acoustical stimulation," *Proceedings of the National Academy of Sciences* **16**, 344.

Wever, E. G. and Bray, C. W. (**1930**b). "Present possibilities for auditory theory," *Psychological Review* **37**, 365–380.

Wheeler, H. A. (**1928**). "Automatic volume control for radio receiving sets," *Proceedings of the Institute of Radio Engineers* **16**, 30–34.

White, M. W. (**1986**). "Compression systems for hearing aids and cochlear prostheses," *Journal of Rehabilitation Research and Development* **23**, 25–39.

Whitehead, M. L., Lonsbury-Martin, B. L., Martin, G. K., and McCoy, M. J. (**1996**). "Otoacoustic emissions: Animal models and clinical observations," in *Clinical Aspects of Hearing*, edited by T. R. Van De Water, A. N. Popper, and R. R. Fay, 199–257 (Springer).

Whitman, B. and Rifkin, R. (**2002**). "Musical query-by-description as a multiclass learning problem," in *IEEE Workshop on Multimedia Signal Processing*, 153–156.

Widrow, B. and Hoff, M. E. (**1960**). "Adapting switching circuits," in *Proceedings of the IRE Western Electronics Show and Convention, WESCON*, 96–104.

Wilson, H. A. and Myers, C. S. (**1908**). "The influence of binaural phase differences on the localisation of sounds," *The British Journal of Psychology* **2**, 363–385.

Wilson, J. P. (**1973**). "A sub-miniature capacitive probe for vibration measurements of the basilar membrane," *Journal of Sound and Vibration* **30**, 483–493.

Wilson, J. P. (**1992**). "Cochlear mechanics," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner, 71–84 (Pergamon Press).

Wilson, K. W. and Darrell, T. (**2006**). "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 2156–2164.

Wilson, P. S. (**2014**). "Coffee roasting acoustics," *Journal of the Acoustical Society of America* **135**, EL265–EL269.

Winer, J. A. (**2006**). "Decoding the auditory corticofugal systems," *Hearing Research* **212**, 1–8.

Winter, I. M., Palmer, A. R., Wiegrebe, L., and Patterson, R. D. (**2003**). "Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication* **41**, 135–149.

Wintz, P. A. (**1972**). "Transform picture coding," *Proceedings of the IEEE* **60**, 809–820.

Witkin, A. P. (**1983**). "Scale-space filtering," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence—Volume 2*, 1019–1022 (Morgan Kaufmann).

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. S., and Kollmeier, B. (**1997**). "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *Acta Acustica United with Acustica* **83**, 684–699.

Wong, P. C. M., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., and Zatorre, R. J. (**2008**). "Volume of left Heschl's gyrus and linguistic pitch learning," *Cerebral Cortex* **18**, 828–836.

Woodruff, J. and Wang, D. (**2013**). "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech, and Language Processing* **21**, 606–815.

Wrightson, T. (**1918**). *Inquiry into the Analytical Mechanism of the Internal Ear* (London: Macmillan).

Wu, Z. and Cao, Z. (**2005**). "Improved mfcc-based feature for robust speaker identification," *Tsinghua Science and Technology* **10**, 158–161.

Yadav, S. K. and Kalra, P. K. (**2010**). "Automatic fault diagnosis of internal combustion engine based on spectrogram and artificial neural network," in *Proceedings of 10th WSEAS International Conference on Robotics, Control, and Manufacturing Technology*, 101–107.

Yaeger, L. S., Webb, B. J., and Lyon, R. F. (**1998**). "Combining neural networks and context-driven search for online, printed handwriting recognition in the Newton," *AI Magazine* **19**, 73–89.

Yagnik, J., Strelow, D., Ross, D. A., and Lin, R.-S. (**2011**). "The power of comparative reasoning," in *IEEE International Conference on Computer Vision (ICCV)*, 2431–2438.

Yang, W. Y. (**2009**). *Signals and Systems with MATLAB*, chapter Continuous-Time Systems and Discrete-Time Systems, 292–293 (Springer).

Yang, X., Wang, K., and Shamma, S. A. (**1992**). "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory* **38**, 824–839.

Ye, Y., Machado, D. G., and Kim, D. O. (**2000**). "Projection of the marginal shell of the anteroventral cochlear nucleus to olivocochlear neurons in the cat," *The Journal of Comparative Neurology* **420**, 127–138.

Yerkes, R. M. (**1920**). *The New World of Science: Its Development During the War* (New York: The Century Co.).

Yin, T. C. T. (**1994**). "Physiological correlates of the precedence effect and summing localization in the inferior colliculus of the cat," *The Journal of Neuroscience* **14**, 5170–5186.

Yin, T. C. T. and Chan, J. C. K. (**1990**). "Interaural time sensitivity in medial superior olive of cat," *Journal of Neurophysiology* **64**, 465–488.

Yin, T. C. T., Chan, J. C. K., and Carney, L. H. (**1987**). "Effects of interaural time delays of noise stimuli on low-frequency cells in the cat's inferior colliculus. III. Evidence for cross-correlation," *Journal of Neurophysiology* **58**, 562–583.

Yoon, Y.-J., Puria, S., and Steele, C. R. (**2007**). "Intracochlear pressure and derived quantities from a three-dimensional model," *Journal of the Acoustical Society of America* **122**, 952–966.

Yoon, Y.-J., Steele, C. R., and Puria, S. (**2011**). "Feed-forward and feed-backward amplification model from cochlear cytoarchitecture: An interspecies comparison," *Biophysical Journal* **100**, 1–10.

Yost, W. A. (**1991**). "Auditory image perception and analysis: The basis for hearing," *Hearing Research* **56**, 8–18.

Yost, W. A. (**2007**). *Fundamentals of Hearing: An Introduction*, 5th edition (Academic Press).

Yost, W. A. (**2009**). "Pitch perception," *Attention, Perception, and Psychophysics* **71**, 1701–1715.

Young, E. D. and Calhoun, B. M. (**2005**). "Nonlinear modeling of auditory-nerve rate responses to wideband stimuli," *Journal of Neurophysiology* **94**, 4441–4454.

Young, E. D. and Davis, K. A. (**2002**). "Circuitry and function of the dorsal cochlear nucleus," in *Integrative Functions in the Mammalian Auditory Pathway*, edited by A. N. P. Donata Oertel Richard R. Fay, 160–206 (Springer).

Young, E. D. and Oertel, D. (**2003**). "Cochlear nucleus," in *The Synaptic Organization of the Brain*, edited by G. M. Shepherd, 5th edition, 125–163 (Oxford University Press).

Young, E. D. and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *Journal of the Acoustical Society of America* **66**, 1381–1403.

Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., Chung, P.-H., Ho, C.-H., Chang, C.-F., Wei, Y.-H., et al. (**2010**). "Feature engineering and classifier ensemble for KDD Cup 2010," in *Proceedings of the KDD Cup 2010 Workshop*, 1–16.

Yu, Y.-Q., Xiong, Y., Chan, Y.-S., and He, J. (**2004**). "Corticofugal gating of auditory information in the thalamus: An in vivo intracellular recording study," *The Journal of Neuroscience* **24**, 3060–3069.

Yund, E. W. and Buckles, K. M. (**1995**). "Multichannel compression hearing aids: Effect of number of channels on speech discrimination in noise," *Journal of the Acoustical Society of America* **97**, 1206.

Zacksenhouse, M., Johnson, D. H., Williams, J., and Tsuchitani, C. (**1998**). "Single-neuron modeling of LSO unit responses," *Journal of Neurophysiology* **79**, 3098.

Zbilut, J. P. and Webber, C. L. (**2006**). "Recurrence quantification analysis," in *Wiley Encyclopedia of Biomedical Engineering* (Wiley Online Library).

Zhang, M., Kalinec, G. M., Urrutia, R., Billadeau, D. D., and Kalinec, F. (**2003**). "ROCK-dependent and ROCK-independent control of cochlear outer hair cell electromotility," *Journal of Biological Chemistry* **278**, 35644–35650.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (**2001**). "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *Journal of the Acoustical Society of America* **109**, 648–670.

Zhao, B. and Müller, U. (**2015**). "The elusive mechanotransduction machinery of hair cells," *Current Opinion in Neurobiology* **34**, 172–179.

Zhao, X. and Wang, D. (**2013**). "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 7204–7208.

Zobel, O. J. (**1924**). "Transmission characteristics of electric wave filters," *Bell System Technical Journal* **3**, 567–620.

Zurek, P. M. (**1980**). "The precedence effect and its possible role in the avoidance of interaural ambiguities," *Journal of the Acoustical Society of America* **67**, 952–964.

Zurek, P. M. (**1981**). "Spontaneous narrowband acoustic signals emitted by human ears," *Journal of the Acoustical Society of America* **69**, 514–523.

Zurek, P. M. (**1987**). "The precedence effect," in *Directional Hearing*, edited by W. A. Yost and G. Gourevitch, 85–105 (Springer).

Zurek, P. M. and Saberi, K. (**2003**). "Lateralization of two-transient stimuli," *Perception and Psychophysics* **65**, 95–106.

Zweig, G. (**1991**). "Finding the impedance of the organ of Corti," *Journal of the Acoustical Society of America* **89**, 1229–1254.

Zweig, G., Lipes, R., and Pierce, J. R. (**1976**). "The cochlear compromise," *Journal of the Acoustical Society of America* **59**, 975–982.

Zwicker, E. (**1979**). "A model describing nonlinearities in hearing by active processes with saturation at 40 dB," *Biological Cybernetics* **35**, 243–250.

Zwicker, E. (**1986**). "A hardware cochlear nonlinear preprocessing model with active feedback," *Journal of the Acoustical Society of America* **80**, 146–153.

Zwicker, E., Fastl, H., and Dallmayr, C. (**1984**). "BASIC program for calculating the loudness of sounds from their 1/3 oct band spectra according to ISO 532 B," *Acustica* **55**, 63–67.

Zwicker, E. and Peisl, W. (**1990**). "Cochlear preprocessing in analog models, in digital models and in human inner ear," *Hearing Research* **44**, 209–216.

Zwicker, E. and Scharf, B. (**1965**). "A model of loudness summation," *Psychological Review* **72**, 3–26.

Zwislocki, J. J. (**1950**). "Theory of the acoustical action of the cochlea," *Journal of the Acoustical Society of America* **22**, 778.

Zwislocki, J. J., Szymko, Y. M., and Hertig, L. Y. (**1997**). "The cochlea is an automatic gain control system after all," in *Diversity in Auditory Mechanics*, edited by E. R. Lewis, G. R. Long, R. F. Lyon, P. M. Narins, C. R. Steele, and E. Hecht-Poinar, 354–360 (World Scientific Publishing).

Author Index

Abbott, Derek 307
Abbott, Edwin A. 370
Abdallah, Samer 425
Abramowitz, Milton 246
Aertsen, Ad M. H. J. 159, 180
Aggazzotti, Alberto 368
Ahmed, Nasir 77
Albani, Stephan 460
Albert, Monika 28, 392
Algazi, V. Ralph 373, 374
Alinaghi, Atiyeh 382
Allen, Jont B. 15, 48, 51, 191, 244, 269, 273, 297, 310, 315
Allerhand, Michael 60, 167, 232, 235, 342, 443
Altman, John A. 342
Ambikairajah, Eliathamby 279
Andén, Joakim 422
Anderson, David J. 24, 269, 273, 333
Anderson, Michael J. 395
Angelo, E. James, Jr. 135
Angelucci, Alessandra 394
Arbogast, Tanya L. 385
Arcak, Murat 297
Arruda, Marianne 373
Asano, Futoshi 459
Ashmore, Jonathan 267
Assmann, Peter F. 66, 344, 384, 387, 404
Atal, Bishnu S. 75, 83
Atame, Swati 86
Atencio, Craig A. 424
Atick, Joseph J. 394
Atkinson, Joseph F. 223
Atlas, Les E. 461
Avendano, Carlos 373, 374
Aydelott, Jennifer 461

Baer, Thomas 48, 108
Baker, Richard J. 231, 235–237, 241, 243
Bal, Ramazan 338
Bale, George Gilley P. 14
Baliga, Shankar 464
Baluja, Shumeet 442, 443, 453
Bar-Yam, Yaneer 12
Bar-Yosef, Omer 395
Barchiesi, Daniele 422
Barker, Jon 381, 387, 388, 391
Barrington, Luke 428
Barth, Adolf 379
Barton, Edwin Henry 231
Baumgarte, Frank 277
Baumgartner, Robert 373
Bean, Michael A. 161
Beauvois, Michael W. 384
Békésy, Georg von 21, 23, 26, 227, 258, 324, 425
Bell, C. Gordon 208
Bello, Juan P. 420–422, 425
Bendor, Daniel 391
Bengio, Samy 265, 275, 281, 402, 427, 428, 430, 433, 434, 437, 443, 465
Bengio, Yoshua 418, 420, 422
Bennett, Stuart 120
Bergeaud, François 431
Berger, Jonathan 428
Bergstra, James 420, 422
Bertin-Mahieux, Thierry 452
Beyer, Robert T. 15, 22
Bialek, William 423
Bianchi, Federica 380
Billadeau, Daniel D. 273
Billington, David P. 92
Billington, David P., Jr. 92
Billock, Vincent A. 41
Bilsen, Frans A. 54

Bishop, Christopher M. 419
Bizley, Jennifer K. 391
Black, Norman D. 279
Blackburn, Carol C. 338
Blauert, Jens 381, 382
Bleeck, Stefan 391
Bloom, Burton H. 67
Boahen, Kwabena 277
Bode, Hendrik W. 36
Bogert, Bruce P. 80, 229, 256
Boll, Steven F. 41, 388
Boney, Laurence 456
Bordes, Antoine 418
Borg, Erik 379
Boring, Edwin G. 21
Bormann, Tobias 13
Bottou, Léon 416, 427
Bouguer, Pierre 39
Bourlard, Hervé 85
Bousquet, Olivier 416
Bouvrie, Jake 422
Bowlker, Thomas J. 374
Boyd, Stephen 436
Braida, Louis D. 40
Brand, Thomas 370
Brandenburg, Karlheinz 339
Bray, Charles W. 21, 24, 43, 333
Bregman, Albert S. 10, 46, 67, 383, 384, 386
Brennan, Robert 458, 459
Breschet, Gilbert 4
Bridle, John S. 81, 425
Briede, Thorsten 28, 392
Brin, David 466
Britannica 31, 218
Broadbent, Donald E. 385
Brookes, Tim 381
Brown, Ann M. 181
Brown, C. Phillip 371, 373
Brown, Guy J. 67, 68, 382, 385, 388, 389
Brown, Michael D. 81
Brownell, William E. 66
Bruce, Ian C. 243
Bruce, Robert V. 461
Brugge, John F. 24, 269, 273, 333, 379
Brungart, Douglas S. 388
Buchanan, Anne 66
Buckles, Krista M. 459
Burkard, Robert F. 379
Burrus, C. Sidney 129

Cahan, David 15
Caldwell, William F. 28, 256
Calhoun, Barbara M. 180
Campbell, George A. 210
Campbell, Jeffrey 383, 386
Cao, Zhigang 86
Capranica, Robert R. 184
Cariani, Peter A. 23, 28, 59, 60, 336, 339
Carney, Laurel H. 169, 243, 246, 247, 281, 289, 338, 375
Carrell, Thomas D. 65
Casey, Michael A. 453, 465
Casseday, John H. 392
Casson, Herbert Newton 456
Cedolin, Leonardo 59
Cetas, Justin S. 19
Chalupper, Josef 51, 459
Chan, Antoni 428
Chan, Joseph C. K. 375
Chan, Ying-Shing 386
Chandra, Tushar 436
Chang, Chun-Fu 420
Chang, Edward F. 386, 394, 395
Chang, Peter S. 388
Chapelle, Olivier 427
Cheatham, Mary Ann 309
Chechik, Gal 265, 275, 281, 383, 395, 427, 428, 430, 433, 440, 443
Chen, Bo-Wei 463
Cherry, E. Colin 383
Cheung, Connie 395
Childers, Donald 163
Choe, Yong 181, 182
Choi, Jae-Yeon 459
Chou, Jung-Wei 420
Chowdhury, Syed A. 389, 394
Christensen, Heidi 381
Chung, Dong-Ook 459
Chung, Po-Han 420
Clarkson, Brian 466
Clifton, Rachel K. 381
Cohen, Jordan R. 85
Colburn, H. Steven 373, 381
Collins, Nick 425

Collins, Steve 381
Collobert, Ronan 402, 421
Cook, Daniel L. 378
Cook, Norman D. 28
Cook, Perry R. 15
Cooke, Martin P. 7, 68, 127, 173, 344, 384, 385, 387, 388, 391
Coombes, Stephen 339
Cooper, Nigel P. 25, 184, 255, 299
Corey, David P. 305, 307
Corless, Robert M. 197
Cortes, Corinna 416, 418, 434, 435
Cosell, Lynn 81
Cotton, Courtenay 442, 446, 449–451
Counter, S. Allen 379
Covell, Michele 442, 443, 453
Cover, Thomas M. 407
Covey, Ellen 392
Cramer, Elliot M. 54
Crammer, Koby 434
Crawford, Malcolm D. 388
Crinion, Jennifer 461
Crone, Nathan E. 394
Culling, John F. 67
Cunningham, Daniel John 332, 334, 471, 476

Dadson, Robert S. 48
Dallmayr, Christopf 50
Dallos, Peter 7, 15, 255, 256, 276, 277, 309
Damper, Robert I. 456
Darley, D. Lucille 67
Darling, Angela M. 173, 174, 237, 241, 243
Darrell, Trevor 381
Darrow, Keith N. 270, 379
Darwin, Chris J. 385
Datta, A. Jaysurya 59
Dau, Torsten 380
Daudet, Laurent 425
Davenport, E. W. 232, 242
Davenport, Wilbur B. 180
David, Edward E., Jr. 19, 35, 71, 84, 85, 382
David, Stephen V. 394
Davidson, Grant A. 456
Davies, Mike 425
Davis, Hallowell 24, 256, 293, 309, 333
Davis, Kevin A. 339
Davis, Mark F. 456
Dawson, Joel Lawrence 180
de Boer, Egbert 56, 59, 108, 167, 169, 180, 181, 222, 235, 246, 273
De Forest, Lee 191
de Jongh, H. R. 235, 246
de La Bruyère, Jean ii
de La Rochefoucauld, Ombeline 256
Dean, Thomas 425
Decoste, Dennis 427
Dekel, Ofer 434
Delbruck, Tobias 334
Delcroix, Marc 381
Delgutte, Bertrand 52, 58, 59, 336
Deng, Li 24, 71
Denny, Mark 392
Dietterich, Thomas G. 436
Dietz, Mathias 380
Dinse, Hubert R. 392
Divenyi, Pierre 385
Do, Manuel 379
Doh, Won 459
Dong, Dawei W. 394
Dorf, Richard C. 204
Dosch, H. Günter 391
Drakakis, Emmanuel M. 167, 233, 237, 244, 275
Draper, John Christopher 14
Drygajlo, Andrzej 388
Drysdale, Charles Vickery 368
du Lac, Sascha 28
Duchi, John 436
Duda, Richard O. 68, 342, 371, 373, 374, 383, 384, 417
Duifhuis, Hendrikus 180, 183, 231, 240, 295, 296
Duke, Thomas A. J. 181, 182
Dunlap, Knight 22
Dunne, Robert A. 416
Durlach, Nathaniel I. 40
Duverney, Guichard Joseph 257
Duxbury, Chris 425
Dyer, Lounette 245

Eaglesfield, Charles C. 114, 177
Eargle, John M. 48
Eck, Douglas 420, 421
Eckrich, Tobias 336
Edwards, Brent 460
Egan, James P. 81

Eggermont, Jos J. 180, 379
Eggers, Joachim 459
Eguíluz, Víctor M. 181
Ehret, Gunter 28
Ekanadham, Chaitanya 421, 425
El-Maliki, Mounir 388
Elderton, William Palin 137
Elliott, Stephen J. 181
Ellis, Daniel P. W. 68, 133, 384, 385, 387, 391, 442, 446, 449–452, 466, 467
Elmore, William C. 223
Enochson, Loren D. 41
Epp, Bastian 45
Eriksson, Jan L. 167
Esterly, Steven D. 28
Eustaquio-Martín, Almudena 187
Evans, Edward F. 25, 273, 333
Evans, Thomas G. 67
Evgeniou, Theodoros 416
Ewald, J. Rich 21
Ewert, Sebastian 452

Fahey, Paul F. 273
Faller, Christof 381
Fan, Yun-Hui 108
Fastl, Hugo 50, 51
Fay, Richard R. 428
Ferry, Ervin Sidney 368, 369
Fettiplace, Robert 307
Field, David J. 428, 430
Fielder, Louis D. 456
Fikret-Pasa, Selda 457
Fine, Henry Burchard 38
Fink, Gernot A. 382
Finkelstein, Adam 453
Fischer, Eghart 459
Flanagan, James L. 56, 229, 244
Fletcher, Harvey 21, 23, 51, 240, 258, 333
Flinker, Adeen 394
Fong, Janette 463
Formby, Craig 236
Fornasini, Paolo 147
Foster, Nicholas E. V. 391
Fourier, Joseph 21
Fowler, Edmund Prince 457
Fragnière, Eric 238
Franz, Christoph 336
Frasconi, Paolo 416
Freeman, Larry 24
French, Norman R. 461
Freyman, Richard L. 381
Friedlander, Michael P. 436
Friedman, David H. 277
Fritz, Jonathan B. 394
Fröman, Nanny 219
Fröman, Per Olof 219
Frost, John 5
Fry, Thornton C. 368

Galambos, Robert 333
Gales, Robert S. 232
Galison, Peter 208
Gallun, Frederick J. 385
Gao, Enquan 386, 389, 392
Gardner, Stephanie M. 338
Gargi, Ullas 17
Garrison, Fielding Hudson 47
Geisler, C. Daniel 24, 186, 259, 273, 297
Gelfand, Stanley A. 23, 365
Gersho, Allen 431
Ghoshal, Sarbani 379
Giannoulis, Dimitrios 422
Giguère, Christian 277
Gillespie, Peter G. 307, 309
Gish, Herbert 412
Glaesser, Ewald 28, 256
Glasberg, Brian R. 48, 81, 229, 236, 237, 241, 243, 266
Glickman, Stephen E. 420
Glorot, Xavier 418
Godde, Ben 392
Goebel, Rainer 391
Goetze, Stefan 460
Gold, Bernard 71, 133, 155
Gold, Thomas 14, 24, 249, 256
Golding, Nace L. 338
Goldstein, Julius L. 57, 69, 108, 179–181, 240, 245
Gonnet, Gaston H. 197
Gopinath, Ramesh A. 386
Gorga, Michael P. 51, 459
Gori, Marco 416
Gossett, Philip 62
Goto, Masataka 465
Grande, Lucinda A. 378

Grangier, David 427, 433, 434, 437
Gray, Albert A. 257
Gray, Augustine H 83
Gray, Robert M. 431
Green, David M. 230, 232
Green, George 218
Green, Harriet C. 461
Green, Phil D. 9, 381, 388, 391
Greenberg, Steven 24, 273, 324, 338
Greenwood, Donald D. 81, 263
Grevers, Gerhard 15
Griffiths, Timothy D. 456
Gross, Charles G. 392
Grothe, Benedikt 376, 378, 380
Guérin, Anne 12
Guernsey, Martha 40
Guinan, John J., Jr. 270, 326, 379
Guttman, Newman 56
Guzman, Sandra J. 381

Haas, Helmut 380
Hachmeister, Jorge E. 258
Hacker, Peter 10
Hafter, Ervin R. 381
Hahm, Jong-On 394
Hald, Anders 120
Hall, Deborah A. 391
Hall, Joseph L. 295, 310
Hamacher, Volkmar 459
Hamdy, Khaled N. 456
Hamel, Philippe 420, 421, 465
Hamilton, Tara J. 277
Hamming, Richard W. 113
Hanauer, Suzanne L. 83
Handel, Stephen 59
Harada, Yasuo 267
Harczos, Tamás 339
Hare, David E. G. 197
Harrison, Robert V. 267
Hart, Peter E. 383, 417
Hart, Timothy P. 67
Harte, James M. 181
Hartley, Alice K. 67
Hartley, Ralph V. L. 368
Hartmann, William M. 59, 60, 115, 381
Hartung, Klaus 27, 68, 380
Hausdorff, Jeffrey M. 319
Hawkins, Joseph E 258
Hawley, Monica L. 67
Hayes, Gillian R. 467
Haykin, Simon 411
He, Jufang 386
He, Wenxuan 263, 270
Heald, Mark A. 223
Healy, Eric W. 459, 460
Healy, Michael J. R. 80
Healy, Richard D. 161
Heaviside, Oliver 95
Hecht, Selig 39
Heeger, David J. 423
Heffner, Henry E. 46
Heffner, Rickye S. 46, 378
Heil, Peter 19, 28
Heinz, Michael G. 52, 243
Heller, Eric J. 15
Helmholtz, Hermann Ludwig F. von 14–16, 21, 135, 384, 419, 441
Hennecke, Marius H. 382
Henry, Joseph 379
Hérault, Jeanny 12
Hermansky, Hynek 84, 85, 388
Herrera, Perfecto 422
Herschel, John F. W. 61
Hershey, John R. 386
Hertig, Linda Y. 273
Hess, Andreas 28
Hewitt, Michael J. 27, 60, 68, 312, 344, 384, 387
Hind, Joseph E. 24, 269, 273, 333
Hinton, Geoffrey E. 409, 418, 422
Hirata, Yukari 463
Ho, Chia-Hua 420
Ho-Ching, F. Wai-ling 463
Hochberg, Irving 365
Hoff, Marcian E. 403, 412
Hofstetter, Philip 267
Hohmann, Volker 459, 460
Hoiem, Derek 465
Holder, William 61, 62
Holdsworth, John 60, 167, 173, 232, 235, 342, 430, 443
Hornbostel, Erich M. von 368
Hoshen, Yedid 421
Houtsma, Adrian J. M. 40, 53
Howell, Peter 113

Howell, William Henry 39, 47
Hsieh, Hsun-Ping 420
Hu, Guoning 388
Hu, Ning 273
Hu, Yi 388
Huang, Yixiang 464
Hubbard, Allyn E. 338
Hudspeth, A. James 181, 182, 273, 305, 307
Huggins, William H. 21, 54, 161
Hukin, Robert W. 385, 456
Hummersone, Christopher 381
Humphrey, Eric J. 420–422
Hunter, Ian W. 180
Hurewicz, Witold 120
Huron, David B. 11
Hurst, Charles Herbert 21, 257

Ie, Eugene 428, 430, 433
Imennov, Nikita S. 461
Indyk, Piotr 420
Irino, Toshio 66, 169–171, 173, 235, 237, 240, 241, 243, 245, 246, 349, 456, 459, 463
Iro, Heinrich 15
Issa, John B. 52
Itakura, Fumitada 83
Ives, D. Timothy 302, 443, 452

Jackson, Philip J. B. 382
Jacobs, Charles E. 453
Jaitly, Navdeep 422
James, William 40
Jang, Ling-Sheng 463
Janssen, Thomas 181
Jaramillo, Fernán 307
Jarvis, Erich D. 392
Jeffress, Lloyd A. 13, 25, 27, 66, 231, 365, 370
Jeffrey, David J. 197
Jensen, Kristoffer 359
Jesteadt, Walt 51
Jin, Craig 277
Joachims, Thorsten 434
Johannesma, Peter I. M. 159, 180, 246, 295
Johnson, Don H. 377
Johnson, Keith 83, 395
Johnson, Norman Lloyd 137
Johnson, Roy Michael 127
Johnson, Stuart L. 336
Johnsrude, Ingrid S. 456
Johnston, James D. 456
Johnstone, Brian M. 25, 297
Jones, Arthur Taber 52
Joris, Philip X. 28, 268, 338, 370, 374, 375, 377, 378
Josephs, Oliver 456
Josifovski, Ljubomir 388, 391
Jülicher, Frank 181, 182
Jung, Ho-Young 422
Jungnickel, Christa 15
Jutten, Christian 12

Kaas, Jon H. 28
Kachar, Bechara 307
Kaiser, Alexander 273
Kaiser, James F. 382
Kakehata, Seiji 459
Kalinec, Federico 273
Kalinec, Gilda M. 273
Kalra, Prem Kumar 464
Kammeyer, Karl-Dirk 460
Karam, Lina J. 129
Karino, Shotaro 370
Karjalainen, Matti 80, 455
Karklin, Yan 421, 425
Kashyap, Rangasami L. 405
Kates, James M. 41, 275, 277, 279, 459
Katsiamis, Andreas G. 167, 233, 237, 244, 275
Katz, William F. 404
Kawahara, Hideki 66, 456
Kayser, Christoph 423
Ke, Yan 465
Kellermann, Walter 381
Kemp, David T. 25, 255, 258
Kerns, Douglas A. 238
Kersten, Daniel 386
Keshet, Joseph 434
Kiang, Nelson Y. S. 45, 57, 108, 135, 181, 333, 335
Kick, Shelley A. 273
Kidd, Gerald, Jr. 385
Killion, Mead C. 457, 458
Kim, Dong-Wook 459
Kim, Duck On 25, 89, 173, 175, 183, 191, 241, 245, 256, 270, 271, 273, 294, 315, 379
Kim, Gibak 388
Kim, Kyunghee X. 307
Kim, Seung-Jean 436
Kim, Won-Ki 459

King, Andrew J. 391
King, D. Brett 368
Kingsbury, Brian 421
Kinoshita, Keisuke 381
Kistler, Doris J. 373
Klautau, Aldebaro 412
Kleijn, W. Bastiaan 456
Klemm, Otto 368
Kletsky, Earl J. 114
Knight, Robert T. 394
Knipper, Marlies 336
Knudsen, Eric I. 19, 28, 45, 46
Knudsen, Phyllis F. 45
Knudsen, Vern O. 40
Knuth, Donald E. 197
Kobayashi, Toshimitsu 459
Koch, Christof 408
Koch, Ursula 378
Koh, Kwangmoo 436
Kojima, Shozo 63
Kollmeier, Birger 86, 370, 459, 460
König, Peter 423
Konishi, Masakazu 14, 19, 45, 374, 392
Kopp, George A. 461
Kopun, Judy 51
Koralek, Aaron 420
Körding, Konrad P 423
Korenberg, Michael J. 180
Kornagel, Ulrich 459
Kouh, Minjoon 424
Kricos, Patricia B 461
Kristjansson, Trausti T. 386
Krogh, Anders 411
Kros, Corné J. 336
Krueger, Lester E. 49
Kruidenier, C. 167
Kubin, Gernot 456
Kuhn, George F. 372, 373
Kuhn, Stephanie 336
Kulesza, Randy J., Jr. 378
Kulkarni, Abhijit 373
Kumar, Usha A. 416
Küpfmüller, Karl 191, 193
Kurzweil, Ray 401
Kuwabara, Nobuyuki 389

Laback, Bernhard 373
Laënnec, René-Théophile-Hyacinthe 463
Lamb, Horace 259, 261
Lambert, Laura 331
Lanckriet, Gert 428
Landay, James A. 463
Lane, Clarence E. 51, 212, 229, 256, 258
Langner, Gerald 19, 28, 56, 338, 392
Laudanski, Jonathan 339
Lazzaro, John 425
Leatham, Aubrey 463
LeCun, Yann 420, 421
Lee, Jong-Hwan 422
Lee, Keansub 466, 467
Lee, Soo-Young 422
Lee, Te-Won 422
Lee, Thomas H. 180
Leech, Robert 461
Lehner, Richard J. 463
Leman, Marc 465
Lepore, Jill 461
Levitt, Harry 67, 457, 458, 460, 461
Lewicki, Michael S. 421, 422, 425
Li, Xing 461
Li, Ying 293
Li, Yipeng 388
Liberman, M. Charles 270, 319, 326, 379
Licklider, Joseph C. R. 13, 18, 19, 21, 25, 27, 28, 53, 54, 56, 114, 281, 341–343
Lighthill, James 221, 259
Lim, Kian-Meng 262
Lin, Ruei-Sung 425
Lindeberg, Tony 164
Lindemann, W. 67
Linggard, Robert 279
Link, Brian D. 456
Lipes, Richard 3, 207, 219, 221, 229, 256, 263
Lippert, Ross A. 408, 412
Litovsky, Ruth Y. 67, 381
Liu, Chengliang 464
Liu, Haining 464
Liu, Shih-Chii 334
Liu, Yi-Wen 270
Liu, Zhen 293
Lo, Hung-Yi 420
Loizou, Philip C. 388, 461
Lonsbury-Martin, Brenda L. 45
López, Belén Calvo 201

Lopez-Poveda, Enrique A. 180, 187, 230, 240, 312
Lotze, Al 461
Lou, Jing-Kai 420
Louage, Dries H. G. 378
Lu, Yang 388
Luciani, Luigi 258
Lukashkin, Andrei N. 262, 267
Lumpkin, Ellen A. 307
Lutfi, Robert A. 241
Lyon, Richard F. 10, 27, 56, 60, 67, 68, 114, 166, 167, 187, 191, 222, 229, 233, 234, 237, 238, 240, 241, 243–245, 256, 260, 263, 265, 268, 273, 275, 277, 279, 281, 302, 315, 342, 344, 384, 387, 391, 414, 418, 425, 427, 428, 430, 433, 440, 441, 443, 455, 456

Ma, Ning 381
Ma, Xiaofeng 386, 389, 392
Maas, Roland 381
Machado, Duarte G. 271
Magimai-Doss, Mathew 421
Magnasco, Marcelo O. 181, 182
Mahowald, Misha A. 425
Maison, Stéphane F. 270, 379
Majdak, Piotr 373
Makhoul, John 81
Mallat, Stéphane G. 422, 431
Mallock, Arnulph 368, 369
Manis, Paul B. 66
Mankoff, Jennifer 463
Manley, Geoffrey A. 25, 184, 273
Mann, Steve 466
Marcotti, Walter 336
Markel, John E 83
Marolt, Matija 442
Marquardt, Torsten 380
Marr, David 7, 11
Marrill, Tom 67
Martin, Glen K. 45
Masetto, Sergio 336
Mason, Christine R. 385
Mason, Russell 381
Massaro, Dominic W. 65
Massie, Dana 157, 283
Masterton, R. Bruce 378
Mathevon, Nicolas 420
Mathews, Max V. 63, 232, 283, 399
Matthews, John W. 256
Matthews, Tara 463
Mauchly, John W. 117
Mauermann, Manfred 45
Mayer, Alfred Marshall 8, 51, 139
McAdams, Stephen 342
McAlpine, David 378, 380
McClellan, James H. 129
McClelland, James L. 409
McCormmach, Russell 15
McCoy, Marcy J. 45
McDermott, Hugh J. 461
McDermott, Josh H. 423
McDonnell, Mark D. 307
McDuffy, Megean J. 169, 246, 281, 289
McFadden, Dennis 370
McKendrick, John Gray 257, 462
McKenzie, Todd G. 420
McKeown, Dennis 60, 167, 232, 235, 342, 443
McMullen, Nathaniel T. 19
Mead, Carver A. 12, 109, 238, 241, 256, 260, 263, 268, 277, 279, 425
Meddis, Ray 27, 60, 68, 180, 240, 312, 344, 384, 387
Meijering, Erik 125
Mellinger, David K. 7, 384
Merimaa, Juha 381
Mermelstein, Paul 76, 81
Merzenich, Michael M. 28
Mesgarani, Nima 386, 394, 395
Meyer, Bernd 370
Meyer, Max F. 21, 258
Middleton, David 180
Miller, George A. 81
Miller, Paul H. 232, 242
Miller, Scott 163
Milne, Alan Alexander 16
Milroy, Robert 232, 237, 241
Minch, Bradley A. 334
Minsky, Marvin L. 402
Mohamed, Abdel-Mohsen Onsy 213
Mohamed, Abdel-Rahman 421
Møller, Aage R. 222, 379
Molnar, Charles E. 89, 173, 175, 183, 241, 256, 294
Monaghan, Jessica J. M. 404
Mont-Reynaud, Bernard 11, 385, 465
Moore, Brian C. J. 23, 48, 81, 229, 237, 241, 243, 266

Moreau, Luc 297
Morgan, Nelson 71, 84, 85, 133
Morris, Andrew C. 388
Moschovitis, Christos J. P. 331
Motwani, Rajeev 420
Mount, Richard J. 267
Mountain, David C. 338
Movshon, J. Anthony 422
Moxon, Edwin C. 135
Muckli, Lars 395
Müller, Johannes 21
Müller, Jörg 181
Müller, Klaus-Robert 407, 417
Müller, Meinard 452
Müller, Ulrich 307
Muncey, Robert W. 67
Munson, Wilden A. 51
Murché, Vincent T. 14
Musso, Mariacristina 13
Myers, Charles Samuel 367

Naar, Daniel 456
Nadol, Joseph B., Jr. 319
Nakatani, Tomohiro 381
Naranjo, Edward 464
Narasimhan, S. V. 127
Narayan, Shyamla S. 248, 273, 277
Narayanan, Arun 388
Natarajan, T. 77
Neely, Stephen T. 25, 48, 51, 256, 270, 459
Neisser, Ulric 385
Nelken, Israel 383, 395
Newell, Allen 208
Newman, Edwin B. 67, 81, 365, 370, 379, 380
Nickson, Arthur Francis Bennie 67
Nie, Kaibao 461
Nimmo-Smith, Ian 167, 173, 232, 234, 235, 237, 241, 242
Nipher, Francis Eugene 139
Noack, Carla 461
Nogueira, Waldo 422
Nolan, Jason 466
Nolle, Alfred Wilson 201
Nostrant, A. 267
November, James A. 331
Nowozin, Sebastian 416
Nuttall, Alfred L. 169, 273

O'Callaghan, Casey 23
Oden, Chris 461
Oertel, Donata 338, 339
Ohl, Frank W. 28
Ohm, Georg S. 14, 21
Ohyama, Kenji 459
Okuno, Hiroshi G. 385
Olsen, John F. 386, 392
Olsen, Peder A. 386
Olshausen, Bruno A. 428, 430
Olson, Elizabeth S. 256
O'Mard, Lowel P. 180, 240, 312
Oppenheim, Alan V. 78, 113, 128, 133
Orr, Geneviève 407, 417
O'Shaughnessy, Douglas 71, 81
Osindero, Simon 418
Ospeck, Mark 181, 182
Otnes, Robert K. 41

Painter, Ted 456
Palaz, Dimitri 421
Paliwal, Mukta 416
Pallas, Sarah L. 394
Palmer, Alan R. 333, 338, 339
Palomäki, Kalle J. 382
Papert, Seymour 402
Papoulis, Athanasios 135, 161, 177, 312
Parham, Kourosh 379
Park, David M. R. 67
Parrish, Todd B. 391
Pasanen, Edward G. 370
Pasley, Brian N. 394
Patra, Harisadhan 51
Patterson, Roy D. 59–61, 66, 167, 169–171, 173, 229, 232, 234, 235, 237, 240–243, 245, 246, 297, 299, 302, 338, 342, 349, 404, 428, 430, 443, 452, 456, 459, 463
Patuzzi, Robert B. 25, 273, 297
Pavel, Misha 84, 388
Pearson, Karl 177
Pecharsky, Vitalij K. 233
Pecka, Michael 378, 380
Peeters, Stefaan 173
Peisl, Wolfgang 277, 279, 280
Peissig, Jurgen 459, 460
Peng, Chung-Kang 319
Penhune, Virginia B. 391

Pentland, Alex 466
Percival, Graham 80
Pérez, Juan Pablo Alegre 201
Pérot, Alfred 368
Perrett, Wilfrid 22
Perrole, M. 10
Peters, Robert W. 241
Petersen, Tracy L. 41
Peterson, Liss C. 229, 256
Pfafflin, Sheila M 232
Pfeiffer, Russell R. 89, 173, 175, 180, 183, 240, 241, 294
Phillips, Dennis P. 15
Pichora-Fuller, M. Kathleen 461
Pick, Graham F. 180
Pickles, James O. 25, 184
Pienkowski, Martin 180
Pierce, Arthur Henry 366
Pierce, John R. 3, 19, 35, 56, 63, 91, 207, 219, 221, 229, 256, 263
Pinker, Steven 383, 384
Pisoni, David B. 65
Plack, Christopher J. 391
Plinge, Axel 382
Plomp, Reinier 14, 60, 61, 68, 76, 77, 458
Plumbley, Mark D. 422
Poggio, Tomaso 408, 416, 422, 424
Pollak, Josef 379
Pols, Louis C. W. 76, 77
Ponte, Jay 440
Pontil, Massimiliano 416
Poole, Hilary W. 331
Popper, Arthur N. 428
Porsov, Edward 263, 270
Potter, Ralph K. 461
Pressnitzer, Daniel 297, 299
Price, Robin O. 19
Probst, Rudolf 15
Puder, Henning 459
Pueyo, Santiago Celma 201
Pumphrey, R. J. 14
Puria, Sunil 261, 262

Quackenbos, John Duncan 139

Rabiner, Lawrence R. 67, 71, 72, 84
Rader, Charles M. 155
Ragazzini, John R. 120
Raj, Bhiksha 388
Rajan, Kanaka 423
Ramabhadran, Bhuvana 421
Rameau, Jean Philippe 62
Ramón y Cajal, Santiago 376
Ranatunga, Kishani M. 336
Rangayyan, Rangaraj M. 463
Ranke, Otto Friedrich 229, 258, 259
Ranson, Stephen Walter 390
Rao, Kamisetty R. 77
Rasetshwane, Daniel M. 459
Rass, Uwe 459
Rauschecker, Josef P. 46
Ravuri, Suman 442, 450
Rayleigh, *see* Strutt, John William
Recio-Spinoso, Alberto 108, 246, 248, 273, 277
Redheffer, Raymond M. 96
Rehn, Martin 265, 275, 281, 427, 428, 430, 433, 443
Reichenbach, Tobias 273
Reiss, Lina A. J. 339
Remez, Robert E. 65
Ren, Tianying 263, 270
Rennie, Steven J. 386
Retzius, Gustaf 251
Reynolds, Douglas A. 466
Rhode, William S. 45, 69, 184, 231, 268, 269, 299, 315, 324, 338
Rhodes, Christophe 453, 465
Rice, Henry J. 181
Rice, Peter 167, 173, 232, 235
Rich, Nola C. 245, 259, 277, 297
Riesenhuber, Maximilian 422
Rifkin, Ryan 408, 412, 428, 433
Rijntjes, Michel 13
Ritsma, Roelof J. 54, 57
Ritz, Louis A. 66
Roads, Curtis 41
Roaf, Herbert Eldon 258
Robert, Arnaud 167
Roberts, Neil 391
Roberts, Terri P. 336
Roberts, William M. 181
Robertson, Donald 338
Robinson, Arthur 332, 334, 471, 476
Robinson, Donald E. 181
Robinson, Douglas W. 48
Robinson, Ken 60, 167, 232, 235, 342, 443

Robles, Luis 57, 245, 246, 255, 259, 277, 297
Rodríguez, Joyce 51
Roe, Anna W. 394
Rohdenburg, Thomas 460
Roma, Gerard 422
Roman, Nicoleta 68, 388, 389
Rosasco, Lorenzo 422
Rose, Albert 191, 273
Rose, Jerzy E. 24, 269, 273, 333
Rosen, Stuart 113, 231, 235–237, 241, 243
Rosenberg, Aaron E. 56
Rosenblatt, Frank 402
Rosenthal, David F. 68, 385
Rosenzweig, Mark R. 67, 365, 379, 380
Ross, David A. 441
Rossing, Thomas D. 23, 53
Roush, Jackson 69
Roy, Anil K. 391
Rubin, Hillel 223
Rubin, Philip E. 65
Rubinstein, Jay T. 461
Ruggero, Mario A. 57, 108, 186, 245, 246, 248, 255, 259, 273, 277, 297
Rumelhart, David E. 409
Rupp, André 391
Russell, Ian J. 262, 267, 333
Rust, Nicole C. 422
Rutherford, Mark A. 181
Rutherford, William 21, 22, 333
Ruzon, Mark A 425
Ryckebusch, Sylvie 425

Saberi, Kourosh 381
Sachs, Murray B. 59, 333, 338
Sadehh, Abdulmalek 391
Sainath, Tara N. 421
Sakaguchi, Hirofumi 307
Sakai, Masashi 389
Salesin, David H. 453
Salminen, Nelli H. 380
Salvi, Richard 267
Sams, Mikko 19, 28
Samuel, Arthur G. 65
Sandler, Mark B. 425
Sarpeshkar, Rahul 114, 231, 235, 238, 279
Satake, Mitsuaki 459
Sawhney, Nitin 466
Schafer, Ronald W. 71, 72, 78, 84, 128, 133
Schafer, Tillman H. 232
Scharf, Bertram 50
Scheffers, Michael T. M. 66
Scheich, Henning 28
Scherer, P. 370
Scherg, Michael 391
Schiffman, Harvey Richard 23
Schneider, Peter 391
Schneider, Todd 458, 459
Schnupp, Jan W. H. 391
Schofield, Brett R. 11
Schofield, David 232, 235
Schouten, Jan Frederik 43, 53, 56
Schreiner, Christoph E. 19, 28, 341, 424
Schroeder, Manfred R. 69, 71, 75, 219, 310
Schulze, Holger 19, 28
Schwartz, Odelia 422, 424
Schwede, Gary W. 41
Schwindt, Peter C. 378
Scott, Sophie K. 46
Scripture, Edward Wheeler 15
Sebastiani, Fabrizio 420
Seebeck, August 14
Segal, Mark 425
Sehr, Armin 381
Selesnick, Ivan W. 129
Seltzer, Michael L. 388
Serrà Julià, Joan 442, 443, 449
Serre, Thomas 422
Sethares, William A. 63
Sewell, William F. 267
Shackleton, Trevor M. 27, 68
Shalev-Shwartz, Shai 434, 436
Shambaugh, George E. 257, 258
Shamma, Shihab A. 324, 394, 423, 428, 456
Shanmugam, K. Sam 77
Shannon, Claude E. 125
Sharma, Jitendra 394
Sharpee, Tatyana O. 424
Shaw, Greg 180
Shekhter, Ilya 169, 246, 281, 289
Shen, Jianqiang 436
Shepard, Roger Newland 61
Shepherd, William T. 10
Shera, Christopher A. 246, 254, 256, 268
Shewmaker, C. A. 232

Shi, Peng 293
Shlens, Jonathon 425
Siebert, William M. 36, 113, 147, 259
Silverman, Bernard W. 391
Simmons, Andrea Megela 391
Simmons, James A. 273, 391
Simon, Helen 461
Simon, Herbert A. 11, 12
Simoncelli, Eero P. 421–425
Simpson, Brian D. 388
Sinex, Donal G. 19
Singer, Yoram 434, 436
Slaney, Malcolm 11, 60, 68, 166, 167, 173, 234, 237, 342, 384, 387, 391, 428, 453, 456, 465
Sluming, Vanessa 391
Smith, Evan C. 421, 422, 425
Smith, Fraser W. 395
Smith, Julius O 283
Smith, Leslie S. 381
Smith, Philip H. 28, 338, 370, 374, 375
Smith, Robert L. 191
Smoorenburg, Guido F. 57, 296
Snippe, Herman P. 327
Sokolovski, Alexander 24
Sone, Toshio 459
Sontag, Eduardo 297
Sorge, Georg Andreas 15
Spain, William J. 378
Spanias, Andreas 456
Specht, Hans J. 391
Sra, Suvrit 416
Srinivasan, Soundararajan 388
Staecker, Hinrich 15
Stange, Annette 380
Steele, Charles R 262
Stegun, Irene A. 246
Steinberg, John C. 461
Steinmetz, Charles Proteus 38, 213
Stern, Richard M. 381, 388
Stevens, Stanley Smith 31, 40, 48, 81, 370
Stewart, George N. 258
Stewart, John L. 13, 28, 29, 256
Stillingfleet, Benjamin 44
Stillwell, John 38
Stippich, Christoph 391
Stockham, Thomas G. 78
Stork, David G. 417
Stowell, Dan 422
Strelow, Dennis 425
Strube, Hans W. 277
Strutt, John William 259, 365, 366, 368
Suga, Nobuo 386, 389, 392, 394, 395
Sukthankar, Rahul 465
Sullivan, W. E. 19, 45
Summerfield, Quentin 66, 344, 384, 387
Sumner, Christian J. 240, 312, 339
Sung, Po-Hsun 463
Sur, Mriganka 394
Suzuki, Yôiti 459
Swerup, Christer 180
Swets, John A. 232
Szymko, Yvonne M. 273

Taber, Larry A. 223, 262
Takahashi, Terry 45
Takasaka, Tomonori 459
Takeno, Sachio 267
Tan, Hongyang 51
Tan, Qing 243, 246, 247
Tanner, Wilson P. 232
Tapson, Jonathan 277
Tartini, Giuseppe 15
Tchorz, Jürgen 86
Teh, Yee-Whye 418
Temchin, Andrei N. 246
ter Kuile, Emile 255, 258
Terasawa, Hiroko 428
Terhardt, Ernst 53, 56, 61
Terman, Frederick E. 147
Tesi, Alberto 416
Testut, Léo 228, 251, 252, 470, 475
Tewfik, Ahmed H. 456
Therese, Shanthi 86
Theunissen, Frédéric E. 420
Thiers, Fabio A. 319
Thompson, Dennis M. 373
Thompson, Paul O. 232, 242
Thompson, Silvanus P. 367
Tibrewala, Sangita 388
Tillman, Tom W. 458
Tishby, Naftali 395
Todd, Craig C. 456
Todd, Neil P. McAngus 10
Tokita, Joshua 307

Tollin, Daniel J. 67, 377
Tolonen, Tero 80
Torre, Vincent 408
Torres-Carrasquillo, Pedro 466
Torres, David 428
Trahiotis, Constantine 27, 68, 181, 380, 381
Trautwein, Patricia 267
Treisman, Anne 385
Troland, Leonard T. 24
Trotter, Coutts 40
Truong, Khai N. 467
Tsai, Wei-Ho 442
Tsou, Brian H. 41
Tsuchitani, Chiyeko 377
Tucker, David Gordon 114, 177
Tukey, John W. 33, 41, 80
Turian, Joseph 420, 422
Turnbull, Douglas 428
Turnbull, Laurence 400
Turner, Richard E. 404
Tyndall, John 46
Tzanetakis, George 80

Unoki, Masashi 171, 173, 241, 243
Uppenkamp, Stefan 456
Upton, Hubert 463
Urrutia, Raul 273

Van Compernolle, Dirk 166, 167, 173, 234, 237, 312
van de Geer, John P. 76, 77
van de Vorst, J. J. W. 180
Van De Water, Thomas R. 15
van den Berg, Ewout 436
van den Raadt, M. P. M. G. 295
van der Heijden, Marcel 69, 187, 223, 268, 273, 286, 378
van der Pol, Balthasar 295
van Dinther, Ralph 61, 428, 463
van Hateren, J. Hans 327
Van Immerseel, Luc 173
van Netten, Sietse M. 295
van Schaik, André 238, 277, 334
Vapnik, Vladimir 416, 418, 434, 435
Vaseghi, Saeed V. 76
Veena, S. 127
Velenovsky, David S. 19
Veltkamp, Remco 465
Verhey, Jesko L. 45
Verhulst, Pierre François 406
Verhulst, Sarah 380
Vernon, Steve 456
Versteegh, Corstiaen P. C. 187, 273, 286
Vetter, Petra 395
Viergever, Maximus A. 222
Vijayanarasimhan, Sudheendra 425
Villchur, Edgar 457, 458
Vincent, Emmanuel 381
Viskov, O. V. 342
Vittoz, Eric 238
Volkmann, John 81

Wagenaars, Wil M. 53
Wainwright, Martin J. 424
Wake, Mark 267
Walker, Kerry M. M. 391
Wallach, Hans 67, 365, 379, 380
Walling, Mark N. 262, 267
Walters, Thomas C. 265, 275, 281, 302, 404, 427, 441, 443
Wang, Avery 452, 465
Wang, DeLiang 67, 68, 86, 382, 385, 388, 389, 459, 460
Wang, Hsin-Min 442
Wang, Jhing-Fa 463
Wang, Jian 267
Wang, Jieh-Neng 463
Wang, Kuansan 423, 456
Wang, Wenwu 382
Wang, Xiaoqin 391
Wang, Yuxuan 388, 459, 460
Ward, W. Dixon 14
Warr, W. Bruce 245, 272, 273
Warren, Edus H. 326
Warrier, Catherine M. 391
Watt, Henry J. 367
Watts, Lloyd 215, 221, 238, 277, 388, 392, 393, 395
Weaver, Warren 125
Webb, Brandyn J. 414, 418
Webber, Charles L. 422
Weber, Daniel L. 232, 237, 241
Webster, John C. 232, 242
Wegel, Raymond L. 51, 212, 229, 256, 258
Wei, Yin-Hsuan 420
Weiller, Cornelius 13

Weintraub, Mitchel 60, 68, 273, 344, 346, 384, 387, 391
Weiss, Kenneth M. 66
Weiss, Ron J. 421
Weldele, Mary 420
Wellman, Barry 466
Wen, Bo 277
Wenzel, Elizabeth M. 46, 373
Werner, Stephan 339
Wertheimer, Max 368
Wertheimer, Michael 368
Weston, Jason 427, 465
Wever, Ernest G. 21, 24, 43, 258, 333
Wheeler, Harold A. 191, 193, 196, 197
White, Mark W. 459
Whitehead, Martin L. 45
Whitman, Brian 428, 433
Widrow, Bernard 403, 412
Wiegrebe, Lutz 338
Wightman, Fred L. 373
Williams, Jerome 377
Williams, Ronald J. 409
Willsky, Alan S. 113
Wilson, Harold A. 367
Wilson, John P. 25, 270
Wilson, Kevin W. 421
Wilson, Preston S. 467
Winer, Jeffery A. 341, 395
Winter, Ian M. 338
Wintz, Paul A. 419
Witkin, Andrew P. 9, 164
Wittkop, Thomas 460
Wolf, Lior 422
Wong, Patrick C. M. 391
Woo, Hyo-Chang 459
Wood, A. R. 9
Woodford, Chris 331
Woodland, Philip C. 277
Woodruff, John 389
Woods, William S. 460
Wright, Stephen J. 416
Wrightson, Thomas 22
Wu, Zunjing 86

Xiong, Ying 386

Yadav, Sandeep Kumar 464
Yaeger, Larry S. 414, 418
Yagnik, Jay 17, 425
Yang, Won Young 126
Yang, Xiaowei 456
Yates, Graeme K. 25, 297
Ye, Ye 271
Yerkes, Robert Mearns 368
Yin, Tom C. T. 28, 67, 338, 370, 374, 375, 377, 378, 380
Yoho, Sarah E. 459, 460
Yoon, Yong-Jin 261, 262
Yoshioka, Takuya 381
Yost, William A. 9, 59, 342, 381
Youn, Dae-Hee 459
Young, Eric D. 52, 59, 180, 333, 339, 395
Yu, Hsiang-Fu 420
Yu, Hung-Ming 442
Yu, Yan-Qin 386
Yund, E. William 459

Zacksenhouse, Miriam 377
Zadeh, Lotfi A. 120
Zampini, Valeria 336
Zatorre, Robert J. 391
Zavalij, Peter Y. 233
Zbilut, Joseph P. 422
Zhang, Celia 60, 167, 232, 235, 342, 443
Zhang, Jianzhi 293
Zhang, Ming 273
Zhang, Xuedong 243
Zhang, Yunfeng 386, 392
Zhang, Zhifeng 431
Zhao, Bo 307
Zhao, Hong-Bo 379
Zhao, Xiaojia 86
Zheng, Jiefu 273
Zobel, Otto J. 210
Zou, Yuan 273
Zurek, Patrick M. 25, 67, 381
Zweig, George 3, 207, 219, 221, 229, 254, 256, 263, 268
Zwicker, Eberhard 50, 256, 277, 279, 280
Zwislocki, Jozef J. 114, 191, 256, 273

Index

s plane, 108
3 dB per octave, 318
3-dB bandwidth, 136, 137, 150

abnormal growth of loudness, 457
AC, *see* alternating current
AC coupled, 138, 313
acausal system, 119
ACF, *see* autocorrelation function
acoustic approaches, 3
acoustic frequency scale, 81
acoustic phonetics, 63
acoustic startle reflex, 339
acoustic stria, 338
action potential, 25
active traveling-wave theory, 25
active undamping, 256, 267, 297
adaline, 405
address-event representation, 334
AER, *see* address-event representation
afferent, 270, 333
aliasing, 125
all-pole filter, 158
all-pole filter cascade, 238, 240
all-pole gammatone filter, 167, 233, 236–238, 240
all-pole model, 84
Allen hair cell model, 310
alternating current, 92
alveolar, 65
AM radio, 189, 196
amplifier, 94
amplitude frequency response, 101
analog VLSI cochlear model, 279
ANN, *see* artificial neural network
anteroventral cochlear nucleus, 338, 375
anti-resonance, 152
apex, 253, 255, 257, 263, 268, 291, 299, 354
APFC, *see* all-pole filter cascade
APGF, *see* all-pole gammatone filter
AR model, *see* autoregressive model
Aristotle, 7
articulatory features, 65
Artificial Intelligence and Bionics, 331
artificial neural network, 401, 409
ASA, *see* auditory scene analysis
ASR, *see* automatic speech recognition
asymmetric notched noise, 242
asymmetric resonator, 153, 158
asymmetry, 167, 169, 170
asymmetry in auditory image, 349
asymptotic phase and magnitude plots, 36
audio amplifier, 181
auditory correlogram, 53
auditory cortex, 331
auditory evoked potentials, 43
auditory filter model, 227, 229
auditory frequency scale, 41, 81
auditory image, 13
    log-lag, 359
    of music, 355
    of speech, 355
    pitch and spectrum dimensions, 353
auditory image theory, 26
auditory nerve, 227, 331
auditory nervous system, 331
auditory processing disorder, 461
auditory scene analysis, 46, 331, 383, 384
Auditory Scene Analysis: The Perceptual Organization of Sound, 10, 383
auditory stream segregation, 383
auditory streams, 46
auscultation, 463
autocorrelation function, 52–56
automatic gain control, 84, 89, 181, 191, 231, 243, 245, 315

binaural coupling, 315
CARFAC AGC loop, 315
CARFAC loop filter, 315
CARFAC spatial response, 324
CARFAC temporal response, 319
coupled, 315
dynamics, 191, 193, 198, 201, 202, 204, 205
in the cochlea, 273
input–output compression, 191, 192, 195, 201, 205
linearized loop, 193, 198, 204
loop filter, 192–194, 198, 199, 201
speedup factor, 200, 201, 203, 204
stability, 200, 201, 203, 205
system simulation, 201–204
automatic speech recognition, 83–86, 464
autoregressive model, 84
AVCN, *see* anteroventral cochlear nucleus

backpropagation, 409
backward masking, 51
bag of patterns, 430
balanced line, 214
bandpass filter, 135
bandwidth, 41
equivalent rectangular bandwidth, 136, 137
half-power bandwidth, 3-dB bandwidth, or full width at half maximum, 136, 137, 150
bandwidth variation with gain, 182, 183
base, 253, 255–257, 262, 270, 354
basilar membrane, 23, 250
bats, 28, 293
Battle of the Currents, 92
Bayesian, 416
beam search, 385
Bessel function, 168
bifurcation, 181
bilateral Laplace transform, 105
binary logarithm, 32
binaural auditory system, 365
binaural beats, 367
binaural hearing, 331
binaural spatial processing, 331
binaural spatialization, 66
binaural stabilized auditory image, 338
binding problem, 391
BM, *see* basilar membrane
Bode plot, 104, 141, 145, 147
Bode plots, 36
bony shelf, 250, 255
bony snail, 228, 470
brainstem, 331, 389
break frequency, 34
buffer amplifier, 114
bushy cell, 338
globular, 377
spherical, 377
Butterworth lowpass filter, 139

Caltech, 3
calyx of Held, 377
canonical with respect to delay, 128
capacitance, 94
capacitor, 94
carboplatin, 267
CARFAC, 227, 275
framework, 276
open-source software, 279
physiological elements, 276
relation to PZFC, 153, 276
carrier, 223
CASA, *see* computational auditory scene analysis
cascade connection, 111
cascade filterbank, 131, 229
cascade of asymmetric resonators, 275, 281
cascade of asymmetric resonators with fast-acting compression, 153
cascade of gain stages, 197
Cauchy–Lorentz distribution, 177
causal central limit theorem, 177
causality, 98
CCRMA, *see* Center for Computer Research in Music and Acoustics
CDT, *see* cubic distortion tone
Center for Computer Research in Music and Acoustics, 3, 11, 384, 399
center frequency, 41, 131
cepstral analysis, 80
cepstral smoothing, 80
cepstrogram, 79
cepstrum, 77, 80
channel of a filterbank, 41
characteristic frequency, 179
characteristic function, 163

chirp, 169
chopper cell, 338
chopper response, 338
chroma, 60, 441, 443
chromagram, 420, 446
Churcher–King equal-loudness contours, 48
cilia, 250
circuit A, 109
classification, 405
CN, *see* cochlear nucleus
coarticulation, 65
cochleagram, 336
cochlear duct, 250
cochlear nucleus, 331, 337, 365, 375
cochlear partition, 135, 250, 255, 305
cochlear place map, 255
cochlium, 90
cocktail party problem, 46
coefficients, 108
coffee roasting, 467
coincident poles, 161
collaborative filtering, 465
comb filter, 59
combination tone, 181
common fate, 384
common logarithm, 32, 35
communication satellites, 3
compensating loudness by analyzing input-signal digital hearing aid, 459
complete basis, 103
complete characterization, 97, 101, 108
complex exponential, 101, 102
complex frequency, 101
complex gain factor, 101
complex gammatone filter, 161
complex logarithm, 36, 38
complex numbers, 37, 101
complex pole, 147
complex wave, 209
complex wavenumber, 211, 218
complex-conjugate operator, 154
complex-conjugate symmetry, 154
compliance, 213
compression, 184
compressive gammachirp, 240
compressive input–output curve, 279
compressive input–output function, 191
compressive nonlinear system, 184
compressive nonlinearity, 33
computational auditory scene analysis, 67, 384, 391
computational hearing, 11
Computer as a Communication Device, 331
computer hearing, 11
computer listening, 11
computer vision, 11
conditionally stable, 108
conductive hearing loss, 457
cone of confusion, 373
consonance, 62
consonant musical interval, 61
constant-Q filterbank, 41, 75, 76
convergent evolution, 293
convolution, 99
convolution integral, 99
convolution operator, 99
convolution theorem, 111
corner frequency, 104
correlogram, 53, 391
correlograms and the separation of sounds, 384
cortex, 331
cortical frequency map, 341
corticofugal connections, 389
cosine transform, 81
cost function, 412
coupled AGC, *see* automatic gain control, coupled
coupled automatic gain control, 459
coupled form, 283
coupled time-harmonic transmission line equations, 213
coupled-form, 157
coupled-form filter, 294
coupled-form stage, 174
coupling between neighboring channels, 315
critical band, 50, 71, 81, 232
critical bandwidth, 50
critical oscillator, 182
cross-frequency suppression, 459
CT, *see* combination tone
cubic difference tone, 180, 181
cubic distortion tone, 69, 181
cubic nonlinearity, 183

damping factor, 138, 139
damping factor in cascaded resonators, 197

damping-control mechanism, 293
DC, *see* direct current
DC gain, 92, 150, 155, 285
DCN, *see* dorsal cochlear nucleus
decay rate, 138, 161
decibel, 35
decimation, 324
Dedication, ii
deep network, 417
deep-water waves, 261
dehydrated cats, 227
delay line, 208
delta function, 180
depolarization, 305
dereverberation, 339
detection nonlinearity, 312
dichotic pitch, 54
difference equation, 117, 120, 122
difference limen, 35
differential transmission line, 254
diffusion, 320
digital filter, 117
digital outer hair cell, 293–295, 297, 302
dimensionality reduction, 78
diminishing return, 33
Dirac delta function, 146
direct current, 92
direct form I, 128
direct form II, 128
discrete-time system, 117
dispersion relation, 215, 218, 221, 261
dissonance, 62
distance matrix, 449, 450
distributed bandpass nonlinearity, 183
distributed system, 207
dithering, 307
doctrine of specific nerve energies, 21
DOHC, *see* digital outer hair cell
Dolby AC-3 compression, 456
dolphins, 293
dorsal cochlear nucleus, 338, 375
duplex theory, 54, 331, 341
duplex theory of binaural localization, 366
duplex theory of pitch perception, 13, 25
dynamic range, 191

ear trumpet, 400
echolocation, 293
efferent, 270
efferent feedback, 269
eigenfunction, 100, 102, 118
electric wave filter, 210
electrical filter, 94
electrical system, 95
electronic filter, 94
emergent behavior, 12
emergent property, 181
endbulbs of Held, 338, 377
endocochlear potential, 250
endolymph, 250
energy decay, 138
energy-detection approach, 232
envelope, 53
epochs, 411
equal-loudness contours, 48
equivalent noise bandwidth, 137
equivalent rectangular bandwidth, 81, 136, 137
equivalent rectangular bandwidth scale, 266
erasures, 388
ERB, *see* equivalent rectangular bandwidth
ERB scale, *see* equivalent rectangular bandwidth scale
Erlang distribution, 163
error back-propagation, 409
essential nonlinearity, 179, 181
Euler's formula, 37
even-order distortion, 296
excitatory response, 305
expansive nonlinearity, 33
exponential, 31
exponential distribution, 163
extract meaning, 331

fast acting compression, 293
fast Fourier transform, 72
fast-acting compression, 275
feature engineering, 420
feature extraction, 419
Fechner's law, 39, 40, 49
feedback connection, 112
feedback control system, 191
fenestra ovalis, 250
fenestra rotunda, 250
fenestra tympani, 250
fenestra vestibule, 250

filter A, 109
filter cascade, 114, 131
filter cascade model, 227
filter stage, 286
filter-cascade approach, 3, 183
filter-cascade family, 235, 238
filter-cascade model, 169
filterbank, 41, 72, 131
fine temporal structure, 336
fine time structure, 336
finite-impulse-response filter, 120
FIR filter, *see* finite-impulse-response filter
first-order filter, 94
first-order Volterra kernel, 184
fish hearing, 66
Flatlanders, 370
flatness factor, 201
Fletcher–Munson equal-loudness contours, 48
Fletcher–Munson hypothesis, 51
flicker, 307
fluid mass, 213
formant, 58, 59, 65, 66
four-layer model, 18, 421
Fourier analyzer, 14
Fourier spectrum, 131
Fourier transform, 104
frequency, 99
frequency resolution, 132
frequency response, 103
frequency theory, 22
frequency–place map, 255
frequency–threshold curve, 184–186
frequency-domain view, 184
frequentist, 416
FTC, *see* frequency–threshold curve
full width at half maximum, 136, 137
full-wave rectification, 189
furosemide, 267
fusiform cell, 337
FWHM, *see* full width at half maximum

gain adjustment, 155
gamma distribution, 160
gammachirp, 136
gammachirp filter, 169, 235–237
gammatone, 146, 165
gammatone family, 89, 159, 235, 237
gammatone filter, 236
Gaussian distribution, 177
Gaussian filter, 177
GBC, *see* globular bushy cell
GCF, *see* gammachirp filter
generalized autocorrelation, 80
generating function, 120
GitHub, 279
glide, 169
glides, 289
globular bushy cell, 337
globular bushy cells, 377
glottal excitation, 66
glottal pulse, 60
glottal pulses, 66
glottal rate, 66
glottis, 78
grandmother cells, 392
Gray's Anatomy, 251
Greenwood map, 34, 263, 264, 281
group delay, 223, 286
group velocity, 222
grouping, 67
GTF, *see* gammatone filter

Haas breakdown, 380
Haas effect, 67, 379
half-power bandwidth, 142
half-wave rectification, 55
half-wave rectifier, 424
Hamming window, 72
hardware implementation, 277
harmonic sound, 61
harmonic-single-sideband encoder, 461
head-related impulse response, 373
head-related transfer function, 373
hearing aid, 400
hearing seminar, 3
height, 60
helicotrema, 254, 264, 313
Helmholtz resonators, 139
high-level passive limit, 294
high-Q resonance approximation, 148
Hilbert transform, 289
homogeneous solution, 138
homomorphic signal processing, 78
homomorphism, 78

Hopf bifurcation, 181, 295
Hopf oscillator, 182, 183, 295, 296
HRIR, *see* head-related impulse response
HRTF, *see* head-related transfer function
Huggins pitch, 54
hydrodynamic wave, 210
hydromechanical filtering, 173
hyena, 420
hyperbolic tangent, 261, 407
hyperbolic-cosine depth dependence, 261
hyperpolarization, 305

IC, *see* inferior colliculus
ICC, *see* inferior colliculus, central nucleus
ideal binary mask, 388
ideal observer, 232
ideal ratio mask, 388
IF filter, *see* intermediate-frequency filter
IIR filter, *see* infinite-impulse-response filter
ILD, *see* interaural level difference
ill-posed problem, 408
imaginary-part operator, 154
impulse invariance, 127
impulse response, 96, 118
impulse-invariance digital filter design method, 174
incomplete gamma functions, 177
incus, 250
independent identically distributed random variables, 163
inductor, 94
inferior colliculus, 331, 375, 389
inferior colliculus, central nucleus, 28
infinite-impulse-response filter, 120
inharmonic partial, 63
inhibition, 184, 305
inner hair cell, 305
input–output level curve, 196
instantaneous frequencies, 286
instantaneous input–output function, 180
integrator, 108
intensity level, 193
interaural coherence, 381
interaural level difference, 66, 366
interaural phase difference, 366
interaural time difference, 28, 66, 366
interaural-polar coordinate system, 371
interclick interval, 55, 56
intermediate frequency, 189
intermediate-frequency filter, 189
intermodulation product frequencies, 189
Internet, 331
interval histogram, 334
intervalgram, 420, 441, 446
IPD, *see* interaural phase difference
ipsilateral, 333
iso-frequency curve, 185
iso-intensity curve, 185
iso-level curve, 185
iso-response curve, 185
isofrequency unit, 272
ITD, *see* interaural time difference

Jet Propulsion Laboratory, 3
jnd, *see* just noticeable difference
just tuning, 61
just-noticeable difference, 31, 35, 49

kanamycin, 267
Karhunen–Loève transform, 76
key invariance, 442
Kirchhoff's current law, 94
KLT, *see* Karhunen–Loève transform
kurtosis, 137
Kurzweil, Ray, 401

labial, 65
laboratory instrument computer, 89
ladder filter, 210
language, written and spoken, 7
Laplace transform, 100, 104, 117, 162
large-signal linear limit, 293
lasso regression, 416
lateral geniculate body, 392
lateral geniculate nucleus, 392
lateral inhibition, 324
lateral inhibitory connection, 338
lateral lemniscus, 375
lateral line system, 66
lateral nucleus of the trapezoid body, 374, 375
lateral superior olive, 66, 374
lateral suppression, 338
law of the first wavefront, 67, 379
least mean square, 403
level, 193
level-dependent linear filter, 240

level-dependent model, 229
LG method, 219
LGN, *see* lateral geniculate nucleus
Licklider's duplex theory, 54
liftered mel spectrogram, 82
liftered spectrogram, 79
LINC, 89
linear mechanical system, 95
linear predictive coding, 83
linear superposition, 179
linear systems theory, 89
linear time-invariant system, 92, 153
linear–compressive–linear input–output response, 299
Liouville–Green method, 219
live cochlea, 184
LL, *see* lateral lemniscus
LMS, *see* least mean square
LNTB, *see* lateral nucleus of the trapezoid body
local gain control, 319
locality-sensitive hash, 441, 449, 453
log base, 32
logarithm, 31
logistic function, 405, 407, 409
logit, 407
long-term adaptation, 319
long-wave region, 261
Lorentzian function, 147
loss function, 405
lossless medium, 209
loudness level, 193
loudness recruitment, 457
low harmonics, 61
low-bit-rate communication of speech, 83
lowpass filter, 94
LPC, *see* linear predictive coding
LSO, *see* lateral superior olive
LTI system, *see* linear time-invariant system
lumped element, 95

Mössbauer technique, 269
machine hearing, 11
machine hearing applications, 399
machine hearing field, 3
machine learning, 401
machine listening, 11
machine vision, 11
magnitude frequency response, 101, 185
malleus, 250
Man–Computer Symbiosis, 331
manner of articulation, 65
manometric flame, 461
masking, 50, 245
mass–spring system, 109
matched Z transform method, 127
matching pursuit, 431
maximally informative dimension, 424
McGurk effect, 65
mean opinion score, 388
mean-rate threshold, 307
mean-response threshold, 307
mechanical system, 95
mechano-electrical transducer, 306, 307
Meddis hair cell model, 312
medial geniculate body, 389
medial nucleus of the trapezoid body, 374, 375
medial olivo-cochlear efferents, 315
medial superior olive, 28, 66, 374
medulla oblongata, 389
mel cepstrogram, 82
mel scale, 34, 35, 81
mel spectrogram, 82
mel-frequency cepstral coefficients, 81, 419
mel-frequency cepstrum, 81
mel-scale log spectrum, 81
melody, 441
memory, 94
memoryless compressive nonlinearities, 188
memoryless nonlinearity, 180
mercury delay line, 208
MET, *see* mechano-electrical transducer
MGB, *see* medial geniculate body
micromechanics, 255
midbrain, 331, 389
minimum phase, 108, 129
minimum-radius parameter, 294
missing data, 388
ML, *see* machine learning
MLP, *see* multilayer perceptron
MNTB, *see* medial nucleus of the trapezoid body
MOC efferents, *see* medial olivo-cochlear efferents
mode lock, 339
mode-coupling Liouville–Green approximation, 221
modulating the damping, 183
moment-generating function, 163

momentum, 95
motor protein, 261
moving average, 91, 92
moving target indicator, 208
moving-average filter, 208
MP, *see* matching pursuit
MP3 compression, 456
MSO, *see* medial superior olive
multiband compression, 457, 459
multilayer perceptron, 408
multipolar cell, 337
musical interval, 61, 441

NA, *see* nucleus angularis
narrowband spectrogram, 132
natural frequency, 138, 139
natural logarithm, 32, 37
near miss to Weber's law, 49
negative binomial distribution, 164
negative frequency, 102
neural network, 401
neural networks, 414
neuromimetic, 12
neuromorphic, 12
neurotransmitter, 227
NL, *see* nucleus laminaris
NLF, *see* nonlinear function
NM, *see* nucleus magnocellularis
noise power spectral density, 137
nonlinear basilar membrane response, 245
nonlinear cascade filterbank, 276
nonlinear damping, 241
nonlinear filter system, 184
nonlinear filter-cascade model, 183
nonlinear function
    outer hair cell, 293, 295–297, 301, 302
    rectifying inner hair cell, 313
nonlinear system, 89
nonlinear-oscillator hair-cell model, 182
nonlinear-system characterizations, 184
nonrecursive filter, 120
notch, 152
notched noise, 242
nucleus angularis, 376
nucleus laminaris, 375
nucleus magnocellularis, 375
Nyquist criterion, 125
Nyquist frequency, 125
Nyquist rate, 125
Nyquist–Shannon sampling theorem, 122

octave, 34
octopus cell, 337
odd-order distortion, 296
Ohm's law, 94, 107
olivary complex, 66, 331, 365
olivocochlear bundle, 269
one-pole complex system, 147
one-third-octave filterbank, 41, 61, 63
one-zero gammatone filter, 167, 229, 236–238, 240
online training, 403
onset detection, 339
onset response, 338
onset-trigger event, 338
operational calculus, 105
operator, 103
operator notation, 105
organ of Corti, 4, 249–252, 254, 255, 259, 261, 262, 267, 272, 277, 305, 306, 341, 475
orthonormal basis, 77
ossicles, 250
otoacoustic emission, 236
output level, 191
overcomplete basis, 103
OZGF, *see* one-zero gammatone filter, 244

pairs of sinusoids, 184, 187
parallel combination of filters, 138
parallel connection, 112
parallel filterbank, 131
parallel-of-cascades filter, 317
parametric linear system, 188
part tones, 63
partials, 63
Pascal distribution, 164
passband, 135
patent ductus arteriosus, 464
Pattern Classification and Scene Analysis, 383
pauser cell, 338
pauser response, 338
PCA, *see* principal components analysis
PDF, *see* probability density function
peak frequency, 135
peak gain, 135
peak shape, 135

peak width, 135
Pearson distribution, 137
Pearson type III distribution, 163
Pearson type IV distribution, 177
Pearson type VII distribution, 177
pendulum, 109
perceptron, 402
perceptron rule, 405
perceptual coder, 456
perceptual linear prediction, 84
perfect fifth, 34
perfect fourth, 34
perilymph, 250
period histogram, 334
periodic limit cycle, 182, 295
periodicity analysis, 339
periodicity pitch, 53
periodicity theory, 22
phalangeal cells, 250
phase, 61
phase delay, 223
phase difference, 366
phase frequency response, 101
phase shift, 101
phase velocity, 209
phon, 48
phonation, 65
phonautograph, 461
phoneme, 65
phonocardiogram, 463
piano keyboard, 34
pillar cells, 255
pitch, 23
pitch chroma, 60
pitch class, 60
pitch height, 60
pitch map, 338
pitch of the residue, 53
pitch perception, 3, 52
pitchogram, 357, 420, 441
pivot of phase, 286
place coding of pitch, 23
place of articulation, 65
place theory, 22
place theory of sound localization, 13, 25
plane wave, 212
PLP, *see* perceptual linear prediction
Poincaré–Andronov–Hopf bifurcation, 182, 295
point nonlinearity, 180
poised at a bifurcation, 297
pole radius, 294
pole–zero filter cascade, 153, 238, 240
pole–zero mapping method, 126
pole–zero plot, 141, 149, 154
poles, 108
poles and zeros, 159
pons, 389
pooling operator, 422
post-stimulus-time histogram, 334, 377
posteroventral cochlear nucleus, 338, 375
power frequency response, 137
power law, 31, 32, 40
power spectrum, 72
power-law function, 296
precedence effect, 67, 338, 379, 380
precision, 430
precision–recall curve, 441, 450, 452
prestin, 261, 293
primal sketch, 9
primary-like response, 338
principal components analysis, 76
prior distribution, 412
probability density function, 163
probability generating function, 164
probability mass function, 164
pure audition, 387
PVCN, *see* posteroventral cochlear nucleus
pyramidal cell, 337, 338
Pythagoras, 44
PZFC, *see* pole–zero filter cascade, *see* pole–zero filter cascade, 244
PZFC+, 238

Q, 41
Q of a filter, 137
QDT, *see* quadratic difference tone
quadratic difference tone, 180
quadratic distortion tone, 296, 297, 299, 302, 303
quadratic features, 422
quadratic formula, 138
quadratic nonlinearity, 296
quality factor, 137, 142
quantization error, 456
quasi-linear filter, 230, 240

quefrency, 78

raised cosine, 72
ranking, 427
rapid adaptation, 319
rapid and short-term adaptation, 338
RASTA, *see* relative spectral processing
rate–place spectral representation, 338
rational function, 108, 162
rational transfer function, 122, 137, 154
rational-function sigmoid, 312
Rayleigh oscillator, 183
Rayleigh's duplex theory, 54, 365
Rayleigh–Van der Pol oscillator, 295
RC filter, 94
RC lowpass, 94
RC smoothing circuit, 94
RC time constant, 96
reactive element, 107
real gammatone filter, 165
real sinusoid, 101
real system, 153
real-part operator, 154
receptor potential, 254
rectified linear unit, 409
rectifying nonlinear function, 313
recurrence quantification analysis, 422
recurrence relation, 121
recursive filter, 120
regression, 405
regularization, 408
regularized least squares, 408, 412
Reissner's membrane, 250
relative spectral processing, 84
relative undamping, 294
ReLU, *see* rectified linear unit
repeated excitation, 60
reservoirs, 309
residue pitch, 53
resistance, 94
resistive voltage divider, 107
resistor, 94
resonance, 249
resonance theory, 22
resonance–place theory, 25
resonant filter, 89
resonant system, 109
resonator, 137
response of a nonlinear system, 185
response threshold, 185
reticular lamina, 250
retrieval, 427
revcor function, 235
reverse correlation, 245
ridge regression, 416
ringing, 107
ringing frequency, 142, 246
RL, *see* recticular lamina
RLS, *see* regularized least squares
Robinson–Dadson equal-loudness contours, 48
robustness, 66
rods of Corti, 255
roex, 235
roex family, *see* rounded exponential family
root pitch, 62
roots, 108
roots of the denominator, 138
roughness, 61
round window, 254
rounded exponential family, 235, 236

sampled complex exponential, 118
samples, 117
sampling and aliasing, 179
sampling theorem, 122
saturating function, 295
SBC, *see* spherical bushy cell
scala height, 261
scala media, 250
scala tympani, 250
scala vestibuli, 250
scalae, 250
scale-space analysis, 164
scene analysis, 67
Schroeder–Hall hair cell model, 310
scores matrix, 451
second dimension of the cortical sheet, 341
second filter, 184
second-filter theory, 25
second-order digital filter, 124
second-order filter, 108, 109
second-order section, 129, 137
second-order Volterra kernel, 188
semitone, 34

sensorineural hearing loss, 457
series impedance, 107, 109, 213
shallow-water gravity waves, 261
shape parameter, 137, 153
sharpness-enhancing second filter, 184
shear motion, 254
shift invariance, 93
shifting property, 162
shifting property of the Laplace transform, 246
shock absorber, 92
short circuit, 313
short-term adaptation, 319
short-term power spectrum, 75
short-time autocorrelation function, 75, 84, 349
short-time Fourier transform, 72, 132
short-time power, 131
short-time segment, 132
short-time spectral representation, 85
short-wave region, 261
shunt admittance, 213
shunt impedance, 107, 109
shunting inhibition, 377
sigmoid, 308, 407
signal-flow diagram, 120
sine wave, 93
sine-wave speech, 66
single-ended line, 214
single-layer perceptron, 402
single-tuned resonator, 109
siren, acoustic, 8, 14
sketch, 13
skew, 137
skirt of a bandpass filter, 135
SLP, *see* single-layer perceptron
small-signal linear limit, 293
smoothing, 91
smoothing filter, 94, 160, 315
softmax, 425
software implementation, 277
sone, 40, 47, 48
sound effects, 436
sound pressure level, 45
sound separation, 387
sound understanding, 387
source–filter model, 78
space vector, 212
space–time pattern, 28
space–time pattern theory, 24
sparse coding, 428
sparse features, 420, 425
sparse weights, 412
spatial frequency, 209
spatial gradient, 320
spatial smoothing, 324
spatial spreading, 324
spectral estimation, 131
spectral fine structure, 78
spectral tilt, 77
spectrogram, 13
spectrum analyzers, 131
spherical bushy cell, 337
spherical bushy cells, 377
spike timing pattern, 339
spike-timing model, 339
spiral ganglion, 4
spiral ligament, 250
SPL, *see* sound pressure level
square-law detector, 189
square-law rectifier, 424
stability, 98
stability condition, 203
stability of zero crossings, 246
stabilization, 337
stabilized auditory image, 53, 331, 338, 428, 455
stabilized logarithm, 34, 40, 42
stabilized power law, 42
stable system, 98, 119
standing wave, 209, 216
stapedius, 379
stapes, 250
state, 94
state variable, 94
stateless element, 94
statistical learning, 416
stellate cell, 337
stereocilium, 307
Stevens power law, 40
STFT, *see* short-time Fourier transform
stimulus-energy features, 423
stochastic gradient descent, 403
stochastic resonance, 307
stopband, 135
straight-through path, 138, 150
streams, 67

stria vascularis, 250
strobed temporal integration, 428, 430
structure invariance, 442
Student's *t*-distribution, 177
Studies in Auditory and Visual Space Perception, 366
supervised learning, 401
support-vector machines, 416
suppression, 50, 244, 286
suppression areas, 187
suppressor tone, 244
syllabic-time-scale compression, 458
syllable, 65
synchrony suppression, 245
synchrony to waveform events, 338

tail of a bandpass filter, 135
tanh, *see* hyperbolic tangent
Taylor series, 180
tectorial membrane, 250
telegrapher's equations, 213
telephone theory, 22
tempo invariance, 442
temporal theory, 23
tensor tympani, 379
thalamus, 331, 389
Theorica Musice, 44
thermal agitation, 307
third-order distortion product, 181
threshold, 179
timbre, 10, 46, 60
time constant, 96, 161
time invariance, 93
time pattern theory, 24
time resolution, 132
tip link, 307
TM, *see* tectorial membrane
tonality, 60
tone chroma, 60
tone color, 60
Tonempfindungen, 21
tonotopic, 12
tonotopic organization, 333, 336
transfer function, 100, 101
transistor, 3
transmission line, 94
transmission-line model, 232
transposition, 442
transversal filter, 120
transverse cut, 228, 470
traveling wave, 114, 159, 227, 249
traveling wave delay, 108
traveling-wave tube, 3
triangular filter, 236
trigger event, 338
triggered temporal integration, 342, 344
triplex theory, 54
TTI, *see* triggered temporal integration
tuning fork, 6
two-alternative forced-choice experiment, 242
two-layer perceptron, 402
two-pole resonator with unity gain at DC, 155
two-pole–two-zero filter, 137, 281
two-tone suppression, 187, 245
two-voice pitch tracking, 391
tympanum, 250

uniform medium, 209
unit advance, 119
unit delay, 120
unit impulse, 96, 118
unity gain at DC, 92
UNIVAC, 208
universal resonance approximation, 148
universal resonance curve, 147, 164
unstable system, 98, 119
unsupervised learning, 401

Van der Pol oscillator, 183, 295
Van der Pol resonator, 295
variable-gain amplifier, 196
variance, 137
VCN, *see* ventral cochlear nucleus
vector quantization, 425, 431
velar, 65
velocity signal, 294
ventral cochlear nucleus, 375
vertical cell, 338
virtual pitch, 53
Vision, 7
vocal tract, 78
voice activity detection, 467
volley principle, 24
volley theory, 24, 367
voltage divider, 107, 109
Volterra kernel, 179, 182

Volterra series, 179
vowel, 66
VQ, *see* vector quantization

wave digital filter, 277
wave propagation in distributed systems, 89
wave vector, 212
waveform, 6
wavefront plane, 212
wavelength, 212
wavelet modulus operator, 422
wavenumber, 209
Weber fraction, 40, 49
Weber's law, 39, 40
Weber–Fechner psychophysical law, 39
wefts, 391
weight decay, 412
weights, 402
Wentzel–Kramers–Brillouin method, 219, 229
Wertbostel, 368
whispered speech, 66
white noise, 137
wideband spectrogram, 132
Wiener series, 179
windowed segment, 72
windowing, 132
Winnie-the-Pooh, 16
WKB method, *see* Wentzel–Kramers–Brillouin method

Z transform, 117, 118
zero-crossing times, 246, 286
zeros, 108