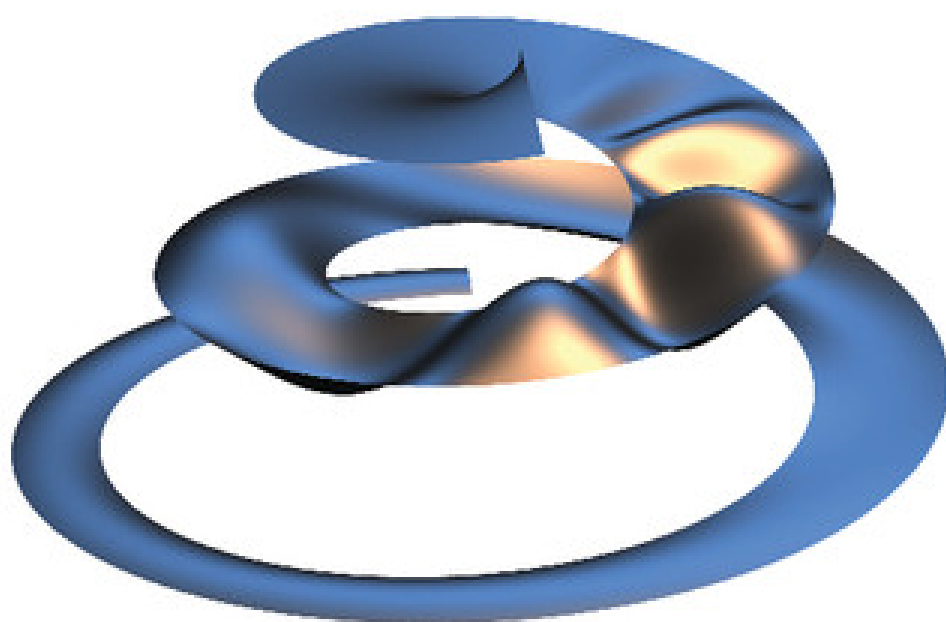


Human and Machine Hearing

Extracting Meaning from Sound



Richard F. Lyon

Human and Machine Hearing

Extracting Meaning from Sound

Human and Machine Hearing describes how human hearing works and how to build machines to analyze sounds the same way people do. The details of this approach are taught at a college engineering level, in a way designed to bring a diverse range of readers to a common technical understanding. The description of hearing as signal-processing algorithms is supported by corresponding open-source code, for which the book serves as motivating documentation. Lyon shows how to understand human hearing in terms of engineering concepts, and to make those concepts into machines that can analyze sounds the way humans do, for a wide range of modern applications. With more than 35 years invested in this approach, Lyon explains how simple concepts, such as that the ear is a Fourier analyzer, have been put behind us, so that we now build machines that approach human abilities in speech, music, and other sound-understanding domains.

Richard F. Lyon is an engineer and scientist known for his work on cochlear models and auditory correlograms for the analysis and visualization of sound, and for analog and digital VLSI implementations of these models, starting at Xerox Palo Alto Research Center, Schlumberger Palo Alto Research, and Apple Advanced Technology Group. After a decade off to develop digital cameras and image sensors at Foveon, he moved back into hearing research, and now leads Google's research and applications development in machine hearing. At Google, he concurrently led the team that developed camera systems for the Street View project. Lyon received a BS in engineering and applied science from Caltech and an MS in electrical engineering from Stanford University. He is a Fellow of the IEEE and a Fellow of the ACM, and is among the world's top 500 editors of Wikipedia. He has published widely in engineering journals of the IEEE, in the *Journal of the Acoustical Society of America*, and in book chapters in diverse fields, including hearing, VLSI design, signal processing, speech recognition, computer architecture, photographic technology, handwriting recognition, computer graphics, and slide rules. He holds 58 issued United States patents for his inventions, including the optical mouse. Although he does not have a doctorate degree, he has co-advised doctoral students and served on doctorate committees at six top universities (including Caltech, Stanford, and UC Berkeley) on three continents.

Human and Machine Hearing

Extracting Meaning from Sound

RICHARD F. LYON

Google, Inc.



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107007536

10.1017/9781139051699

© Richard F. Lyon 2017

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2017

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Lyon, Richard F., author.

Title: Human and machine hearing : extracting meaning from sound / Richard F. Lyon (Google, Inc., Mountain View, California).

Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2017. | Includes bibliographical references and index.

Identifiers: LCCN 2016041119 | ISBN 9781107007536 (alk. paper) | ISBN 1107007534 (alk. paper)

Subjects: LCSH: Hearing. | Auditory perception--Mathematical models. | Auditory perception -- Computer simulation.

Classification: LCC QP461 .L96 2017 | DDC 612.8/5 -- dc23

LC record available at <https://lcn.loc.gov/2016041119>

ISBN 978-1-107-00753-6 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

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

<i>Foreword</i>	<i>page xv</i>
<i>Preface</i>	<i>xix</i>

Part I	Sound Analysis and Representation Overview	1
1	Introduction	5
1.1	On Vision and Hearing <i>à la</i> David Marr	8
1.2	Top-Down versus Bottom-Up Analysis	11
1.3	The Neuromimetic Approach	13
1.4	Auditory Images	14
1.5	The Ear as a Frequency Analyzer?	16
1.6	The Third Sound	18
1.7	Sound Understanding and Extraction of Meaning	18
1.8	Leveraging Techniques from Machine Vision and Machine Learning	19
1.9	Machine Hearing Systems “by the Book”	20
2	Theories of Hearing	23
2.1	A “New” Theory of Hearing	23
2.2	Newer Theories of Hearing	26
2.3	Active and Nonlinear Theories of Hearing	27
2.4	Three Auditory Theories	28
2.5	The Auditory Image Theory of Hearing	29
3	On Logarithmic and Power-Law Hearing	33
3.1	Logarithms and Power Laws	33
3.2	Log Frequency	35
3.3	Log Power	37
3.4	Bode Plots	38
3.5	Perceptual Mappings	41
3.6	Constant- Q Analysis	44
3.7	Use Logarithms with Caution	44

4	Human Hearing Overview	46
4.1	Human versus Machine	46
4.2	Auditory Physiology	46
4.3	Key Problems in Hearing	48
4.4	Loudness	50
4.5	Critical Bands, Masking, and Suppression	52
4.6	Pitch Perception	56
4.7	Timbre	65
4.8	Consonance and Dissonance	66
4.9	Speech Perception	69
4.10	Binaural Hearing	72
4.11	Auditory Streaming	74
4.12	Nonlinearity	75
4.13	A Way Forward	76
5	Acoustic Approaches and Auditory Influence	78
5.1	Sound, Speech, and Music Modeling	78
5.2	Short-Time Spectral Analysis	79
5.3	Smoothing and Transformation of Spectra	83
5.4	The Source–Filter Model and Homomorphic Signal Processing	85
5.5	Backing Away from Logarithms	88
5.6	Auditory Frequency Scales	88
5.7	Mel-Frequency Cepstrum	89
5.8	Linear Predictive Coding	91
5.9	PLP and RASTA	92
5.10	Auditory Techniques in Automatic Speech Recognition	93
5.11	Improvements Needed	94
Part II	Systems Theory for Hearing	95
6	Introduction to Linear Systems	97
6.1	Smoothing: A Good Place to Start	98
6.2	Linear Time-Invariant Systems	99
6.3	Filters and Frequencies	101
6.4	Differential Equations and Homogeneous Solutions	103
6.5	Impulse Responses	103
6.6	Causality and Stability	105
6.7	Convolution	106
6.8	Eigenfunctions and Transfer Functions	107
6.9	Frequency Response	111
6.10	Transforms and Operational Methods	113
6.11	Rational Functions, and Their Poles and Zeros	116

6.12	Graphical Computation of Transfer Function Gain and Phase	119
6.13	Convolution Theorem	120
6.14	Interconnection of Filters in Cascade, Parallel, and Feedback	121
6.15	Summary and Next Steps	125
7	Discrete-Time and Digital Systems	126
7.1	Simulating Systems in Computers	126
7.2	Discrete-Time Linear Shift-Invariant Systems	126
7.3	Impulse Response and Convolution	127
7.4	Frequency in Discrete-Time Systems	127
7.5	Z Transform and Its Inverse	128
7.6	Unit Advance and Unit Delay Operators	129
7.7	Filters and Transfer Functions	131
7.8	Sampling and Aliasing	134
7.9	Mappings from Continuous-Time Systems	136
7.10	Filter Design	138
7.11	Digital Filters	138
7.12	Multiple Inputs and Outputs	141
7.13	Fourier Analysis and Spectrograms	142
7.14	Perspective and Further Reading	144
8	Resonators	145
8.1	Bandpass Filters	145
8.2	Four Resonant Systems	149
8.3	Resonator Frequency Responses	152
8.4	Resonator Impulse Responses	154
8.5	The Complex Resonator and the Universal Resonance Curve	157
8.6	Complex Zeros from a Parallel System	159
8.7	Keeping It Real	163
8.8	Digital Resonators	165
9	Gammatone and Related Filters	169
9.1	Compound Resonators as Auditory Models	169
9.2	Multiple Poles	170
9.3	The Complex Gammatone Filter	172
9.4	The Real Gammatone Filter	175
9.5	All-Pole Gammatone Filters	178
9.6	Gammachirp Filters	181
9.7	Variable Pole Q	184
9.8	Noncoincident Poles	184
9.9	Digital Implementations	185

10	Nonlinear Systems	189
10.1	Volterra Series and Other Descriptions	189
10.2	Essential Nonlinearity	191
10.3	Hopf Bifurcation	192
10.4	Distributed Bandpass Nonlinearity	194
10.5	Response Curves of Nonlinear Systems	195
10.6	Two-Tone Responses	198
10.7	Nonlinearity and Aliasing	199
10.8	Cautions	201
11	Automatic Gain Control	202
11.1	Input–Output Level Compression	202
11.2	Nonlinear Feedback Control	204
11.3	AGC Compression at Equilibrium	205
11.4	Multiple Cascaded Variable-Gain Stages	207
11.5	Gain Control via Damping Control in Cascaded Resonators	209
11.6	AGC Dynamics	210
11.7	AGC Loop Stability	215
11.8	Multiple-Loop AGC	218
12	Waves in Distributed Systems	219
12.1	Waves in Uniform Linear Media	221
12.2	Transfer Functions from Wavenumbers	226
12.3	Nonuniform Media	230
12.4	Nonuniform Media as Filter Cascades	234
12.5	Impulse Responses	235
12.6	Group Velocity and Group Delay	235
Part III	The Auditory Periphery	237
13	Auditory Filter Models	239
13.1	What Is an Auditory Filter?	241
13.2	From Resonance to Gaussian Filters	243
13.3	Ten Good Properties for Auditory Filter Models	244
13.4	Representative Auditory Filter Models	246
13.5	Complications: Time-Varying and Nonlinear Auditory Filters	252
13.6	Fitting Parameters of Filter Models	255
13.7	Suppression	257
13.8	Impulse Responses from Physiological Data	260
13.9	Summary and Application to Cochlear Models	264

14	Modeling the Cochlea	265
14.1	On the Structure of the Cochlea	266
14.2	The Traveling Wave	268
14.3	1D, 2D, and 3D Hydrodynamics	273
14.4	Long Waves, Short Waves, and 2D Models	276
14.5	Active Micromechanics	279
14.6	Scaling Symmetry and the Cochlear Map	280
14.7	Filter-Cascade Cochlear Models	281
14.8	Outer Hair Cells as Active Gain Elements	284
14.9	Dispersion Relations from Mechanical Models and Experiments	287
14.10	Inner Hair Cells as Detectors	288
14.11	Adaptation to Sound via Efferent Control	288
14.12	Summary and Further Reading	291
15	The CARFAC Digital Cochlear Model	293
15.1	Putting the Pieces Together	293
15.2	The CARFAC Framework	294
15.3	Physiological Elements	294
15.4	Analog and Bidirectional Models	297
15.5	Open-Source Software	298
15.6	Detailing the CARFAC	298
16	The Cascade of Asymmetric Resonators	299
16.1	The Linear Cochlear Model	299
16.2	Coupled-Form Filter Realization	300
17	The Outer Hair Cell	309
17.1	Multiple Effects in One Mechanism	309
17.2	The Nonlinear Function	311
17.3	AGC Effect of DOHC	313
17.4	Typical Distortion Response Patterns	315
17.5	Completing the Loop	319
18	The Inner Hair Cell	320
18.1	Rectification with a Sigmoid	322
18.2	Adaptive Hair-Cell Models	324
18.3	A Digital IHC Model	328
19	The AGC Loop Filter	331
19.1	The CARFAC's AGC Loop	331
19.2	AGC Filter Structure	332

19.3	Smoothing Filter Pole–Zero Analysis	332
19.4	AGC Filter Temporal Response	335
19.5	AGC Filter Spatial Response	337
19.6	Time–Space Smoothing with Decimation	338
19.7	Adapted Behavior	341
19.8	Binaural or Multi-Ear Operation	341
19.9	Coupled and Multistage AGC in CARFAC and Other Systems	342
Part IV	The Auditory Nervous System	345
20	Auditory Nerve and Cochlear Nucleus	347
20.1	From Hair Cells to Nerve Firings	347
20.2	Tonotopic Organization	350
20.3	Fine Time Structure in Cochleagrams	351
20.4	Cell Types in the Cochlear Nucleus	352
20.5	Inhibition and Other Computation	353
20.6	Spike Timing Codes	354
21	The Auditory Image	355
21.1	Movies of Sound	355
21.2	History	356
21.3	Stabilizing the Image	357
21.4	Triggered Temporal Integration	360
21.5	Conventional Short-Time Autocorrelation	365
21.6	Asymmetry	367
21.7	Computing the SAI	367
21.8	Pitch and Spectrum	369
21.9	Auditory Images of Music	369
21.10	Auditory Images of Speech	369
21.11	Summary SAI Tracks: Pitchograms	371
21.12	Cochleagram from SAI	373
21.13	The Log-Lag SAI	376
22	Binaural Spatial Hearing	379
22.1	Rayleigh’s Duplex Theory: Interaural Level and Phase	379
22.2	Interaural Time and Level Differences	385
22.3	The Head-Related Transfer Function	386
22.4	Neural Extraction of Interaural Differences	389
22.5	The Role of the Cochlear Nucleus and the Trapezoid Body	392
22.6	Binaural Acoustic Reflex and Gain Control	394
22.7	The Precedence Effect	395

22.8	Completing the Model	397
22.9	Interaural Coherence	397
22.10	Binaural Applications	398
23	The Auditory Brain	400
23.1	Scene Analysis: ASA and CASA	400
23.2	Attention and Stream Segregation	402
23.3	Stages in the Brain	407
23.4	Higher Auditory Pathways	410
23.5	Prospects	415
Part V	Learning and Applications	417
24	Neural Networks for Machine Learning	419
24.1	Learning from Data	419
24.2	The Perceptron	420
24.3	The Training Phase	421
24.4	Nonlinearities at the Output	423
24.5	Nonlinearities at the Input	426
24.6	Multiple Layers	428
24.7	Neural Units and Neural Networks	428
24.8	Training by Error Back-Propagation	429
24.9	Cost Functions and Regularization	432
24.10	Multiclass Classifiers	434
24.11	Neural Network Successes and Failures	436
24.12	Statistical Learning Theory	437
24.13	Summary and Perspective	439
25	Feature Spaces	441
25.1	Feature Engineering	442
25.2	Automatic Feature Optimization by Deep Networks	443
25.3	Bandpass Power and Quadratic Features	444
25.4	Quadratic Features of Cochlear Filterbank Outputs	445
25.5	Nonlinearities and Gain Control in Feature Extraction	446
25.6	Neurally Inspired Feature Extraction	448
25.7	Sparsification and Winner-Take-All Features	448
25.8	Which Approach Will Win?	449
26	Sound Search	450
26.1	Modeling Sounds	451
26.2	Ranking Sounds Given Text Queries	457
26.3	Experiments	461

26.4	Results	463
26.5	Conclusions and Followup	465
27	Musical Melody Matching	467
27.1	Algorithm	469
27.2	Experiments	475
27.3	Discussion	478
27.4	Summary and Conclusions	480
28	Other Applications	481
28.1	Auditory Physiology and Psychoacoustics	481
28.2	Audio Coding and Compression	482
28.3	Hearing Aids and Cochlear Implants	483
28.4	Visible Sound	489
28.5	Diagnosis	491
28.6	Speech and Speaker Recognition	493
28.7	Music Information Retrieval	493
28.8	Security, Surveillance, and Alarms	494
28.9	Diarization, Summarization, and Indexing	495
28.10	Have Fun	495
	<i>Bibliography</i>	497
	<i>Author Index</i>	545
	<i>Subject Index</i>	557

Foreword

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.

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.

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.

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

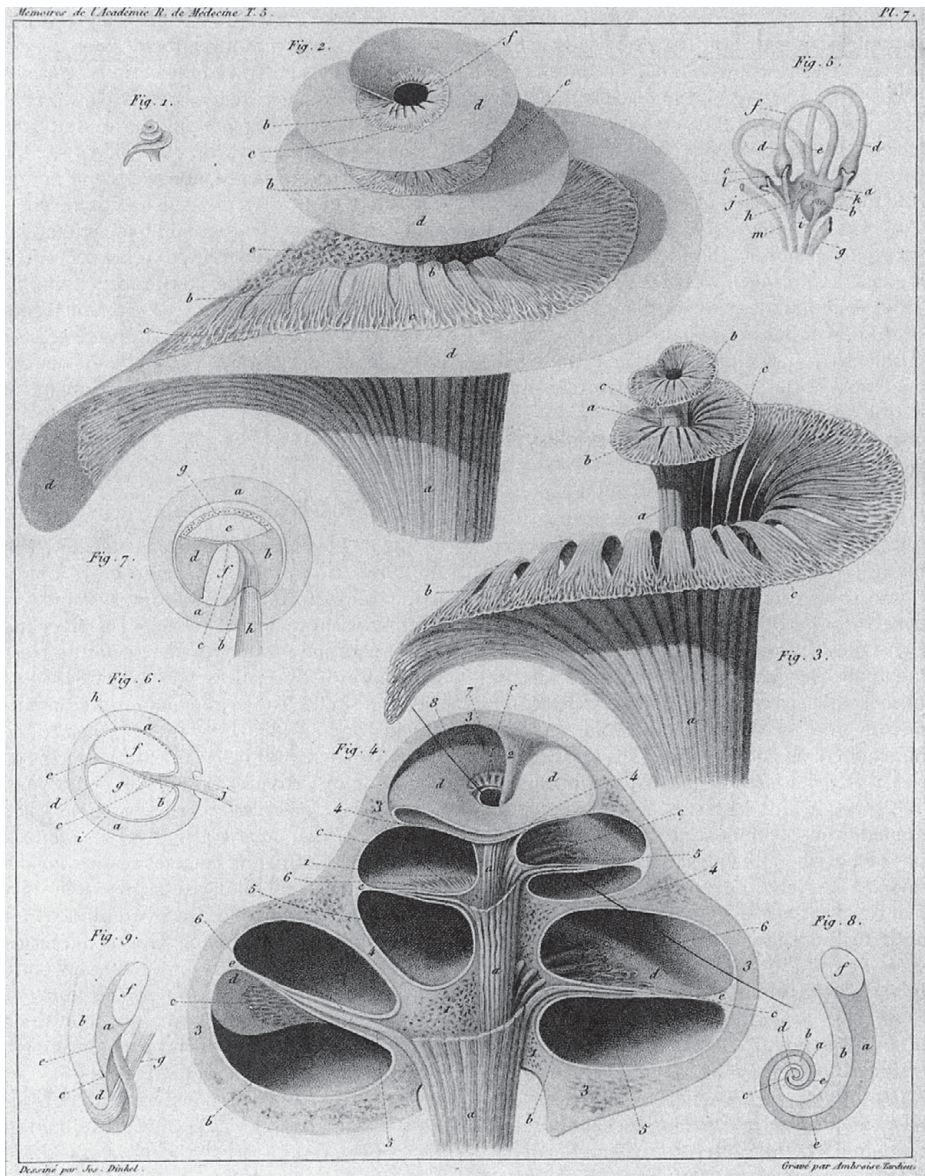
This part is dedicated to the memory of John Robinson Pierce (1910–2002). John was a dear friend and mentor for many years, beginning in my undergraduate years at Caltech. He gave me a summer job doing lab work on electronic musical instruments, and then on digital codecs that led to my first journal 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.

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

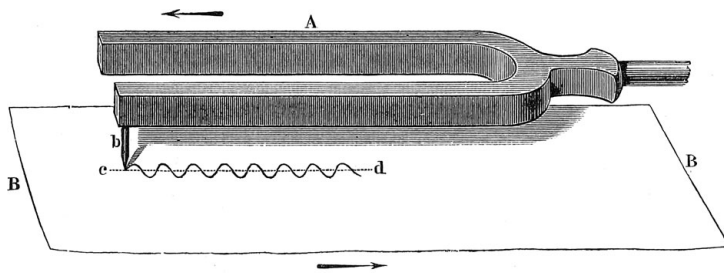


Figure 1.1 Helmholtz explained the idea of a sound's waveform via this diagram of a tuning fork with a stylus point attached, drawing its vibration on a moving piece of paper.

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. 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.



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.

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 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

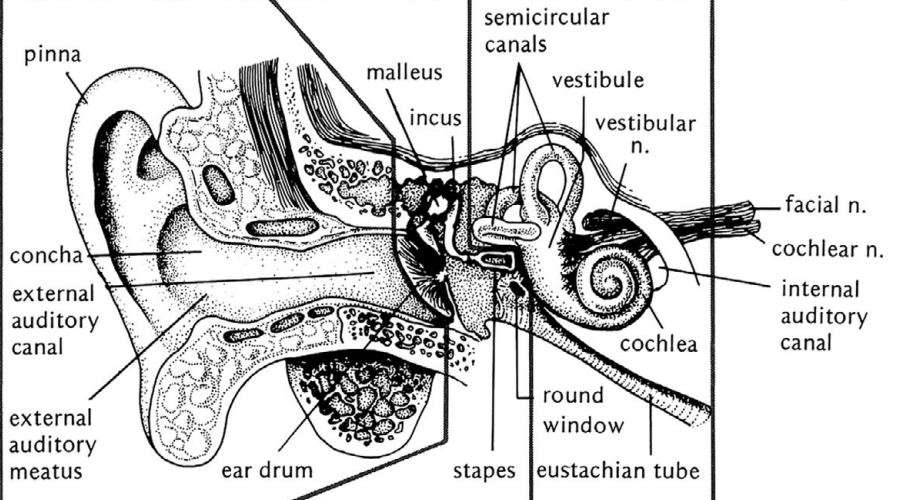
Gross division	Outer ear	Middle ear	Inner ear	Central auditory nervous system
Anatomy				
Mode of operation	Air vibration	Mechanical vibration	Mechanical, Hydrodynamic, Electrochemical	Electrochemical
Function	Protection, Amplification, Localization	Impedance matching, Selective oval window stimulation, Pressure equalization	Filtering distribution, Transduction	Information processing

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.]

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.

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),

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*.

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.



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 $2f_1 - f_2$.

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 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 semiological 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 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;

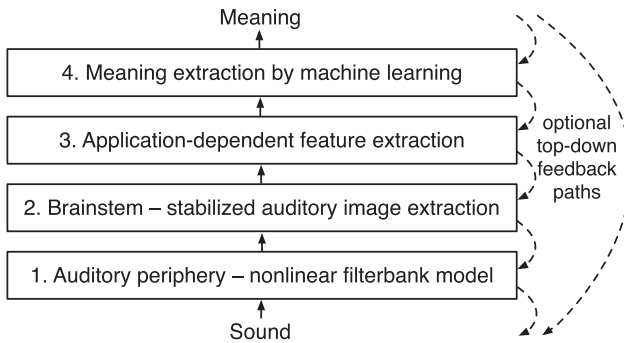


Figure 1.5 The *four-layer model* of machine hearing systems developed in this book—from sound to meaning, and sometimes back the other way. The big feedback loop from meaning to sound is for a system that can make sound and hear itself, for example, a speech conversation system.

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.

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 in section 2.5 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.

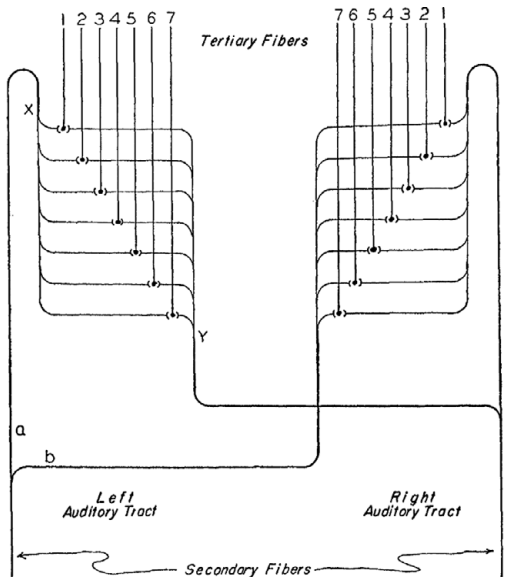


FIGURE 1. HYPOTHETICAL MID-BRAIN MECHANISM FOR THE LOCALIZATION OF LOW FREQUENCY TONES

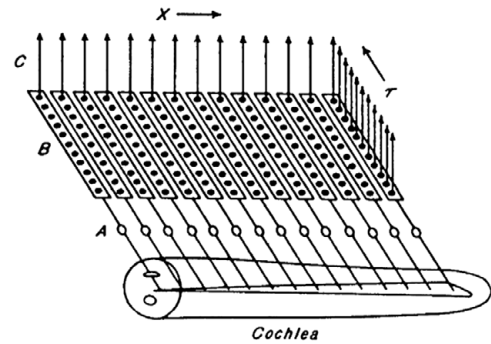


Fig. 2. – Schematic diagram of overall analyzer. At the bottom is the uncoiled cochlea. Its lengthwise dimension and the corresponding dimension in the neural tissue above it is designated the x -dimension. The cochlea performs a crude frequency analysis of the stimulus time function, distributing different 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.]

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.

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).

Sept. 23, 1969

J. L. STEWART

3,469,034

NEURAL-LIKE ANALYZING SYSTEM

Filed May 23, 1966

3 Sheets-Sheet 1

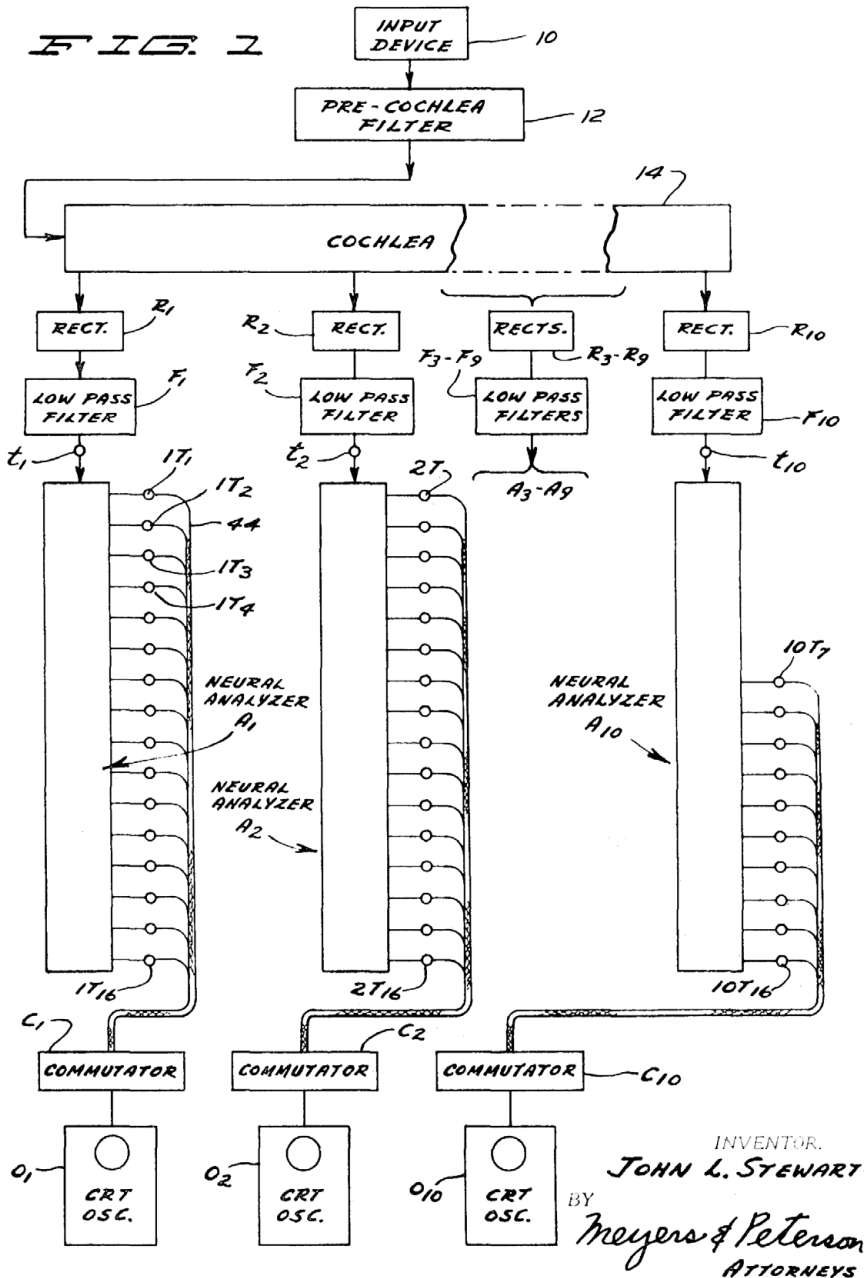


Figure 2.2 A patent drawing from the John L. Stewart (1966) “neural-like analyzer.” The “neural analyzer” stages at each rectified output of the “cochlea” generate a second dimension, mapping the cochlea’s space-time pattern to a space-space pattern, an image-like map, much as in Jeffress’s and Licklider’s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

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.

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 N th power and N th 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* logarithms, 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:

$$\begin{aligned} y &= x^\alpha \\ x &= y^{1/\alpha} \end{aligned}$$

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 $x = 0$:

$$\begin{aligned} y &= (x^\alpha - 1) / \alpha + 1 \\ x &= (\alpha y - \alpha + 1)^{1/\alpha} \end{aligned}$$

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:

$$\begin{aligned} y &= \log_e(x) + 1 \\ x &= \exp(y - 1) \end{aligned}$$

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

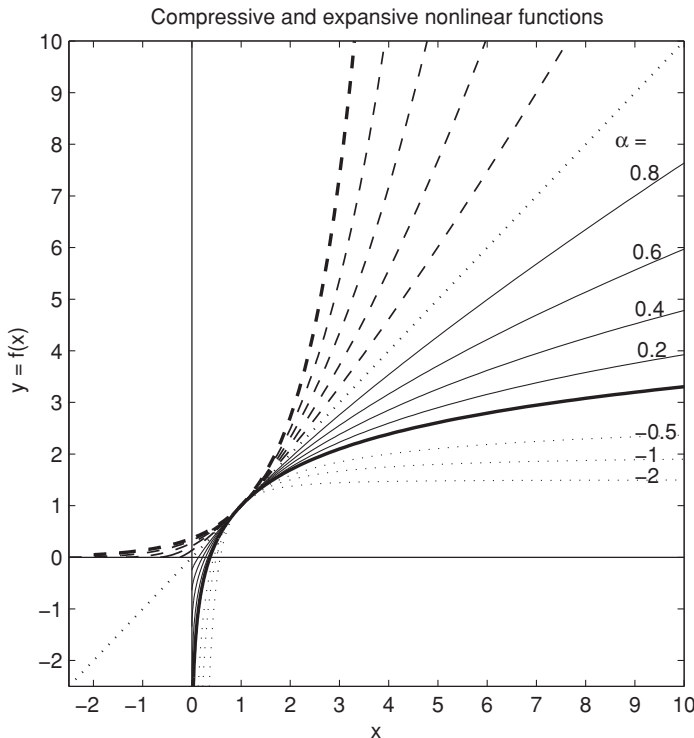


Figure 3.1 Some compressive nonlinearities (solid curves) and the expansive nonlinearities (dashed curves) that are their inverses (compressive means with “diminishing return,” that is, slope decreasing as input increases, while expansive means the opposite). The heavy solid curve is a logarithmic compression, 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 transformations 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.

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 A_m . 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: $x = \text{PitchRatio} = f/27.5$ 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.