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

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://lccn.loc.gov/2016041119

ISBN 978-1-107-00753-6 Hardback

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

Cambridge University Press
978-1-107-00753-6 — Human and Machine Hearing
Richard F. Lyon
Frontmatter
More Information

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.