Speech and Audio Processing

With this comprehensive and accessible introduction to the field, you will gain all the skills and knowledge needed to work with current and future audio, speech, and hearing processing technologies.

Topics covered include mobile telephony, human–computer interfacing through speech, medical applications of speech and hearing technology, electronic music, audio compression and reproduction, big data audio systems and the analysis of sounds in the environment. All of this is supported by numerous practical illustrations, exercises, and hands-on MATLAB examples on topics as diverse as psychoacoustics (including some auditory illusions), voice changers, speech compression, signal analysis and visualisation, stereo processing, low-frequency ultrasonic scanning, and machine learning techniques for big data.

With its pragmatic and application driven focus, and concise explanations, this is an essential resource for anyone who wants to rapidly gain a practical understanding of speech and audio processing and technology.

Ian Vince McLoughlin has worked with speech and audio for almost three decades in both industry and academia, creating signal processing systems for speech compression, enhancement and analysis, authoring over 200 publications in this domain. Professor McLoughlin pioneered Bionic Voice research, invented super-audible silent speech technology and was the first to apply the power of deep neural networks to machine hearing, endowing computers with the ability to comprehend a diverse range of sounds.

Speech and Audio Processing

A MATLAB®-based Approach

IAN VINCE MCLOUGHLIN
University of Kent

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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/9781107085466

First published 2016

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

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

Library of Congress Cataloguing in Publication data
Names: McLoughlin, Ian, author.
Title: Speech and audio processing : a Matlab-based approach / Ian Vince
McLoughlin, University of Kent.
Description: New York, NY : Cambridge University Press, 2016. | © 2016 |
Includes bibliographical references and index.
Identifiers: LCCN 2015035032 | ISBN 9781107085466 (Hardcopy : alk. paper) |
ISBN 1107085462 (Hardcopy : alk. paper)
Subjects: LCSH: Speech processing systems. | Computer sound processing. | MATLAB.
Classification: LCC TK7882.S65 M396 2016 | DDC 006.4/5–dc23
LC record available at http://lccn.loc.gov/2015035032

ISBN 978-1-107-08546-6 Hardback

Additional resources for this publication at www.cambridge.org/mcloughlin and www.mcloughlin.eu

Contents

Cambridge University Press
978-1-107-08546-6 - Speech and Audio Processing: A Matlab®-Based Approach
Ian Vince Mcloughlin
Frontmatter
More information

Cambridge University Press
978-1-107-08546-6 - Speech and Audio Processing: A Matlab®-Based Approach
Ian Vince Mcloughlin
Frontmatter
More information

Preface

Humans are social creatures by nature – we are made to interact with family, neighbours and friends. Modern advances in social media notwithstanding, that interaction is best accomplished in person, using the senses of sound, sight and touch.

Despite the fact that many people would name sight as their primary sense, and the fact that it is undoubtedly important for human communications, it is our sense of hearing that we rely upon most for social interaction. Most of us need to talk to people face-to-face to really communicate, and most of us find it to be a much more efficient communications mechanism than writing, as well as being more personal. Readers who prefer email to telephone (as does the author) might also realise that their preference stems in part from being better able to regulate or control the flow of information. In fact this is a tacit agreement that verbal communications can allow a higher rate of information flow, so much so that they (we) prefer to restrict or at least manage that flow.

Human speech and hearing are also very well matched: the frequency and amplitude range of normal human speech lies well within the capabilities of our hearing system. While the hearing system has other uses apart from just listening to speech, the output of the human sound production system is very much designed to be heard by other humans. It is therefore a more specialised subsystem than is hearing. However, despite the frequency and amplitude range of speech being much smaller than our hearing system is capable of, and the precision of the speech system being lower, the symbolic nature of language and communications layers a tremendous amount of complexity on top of that limited and imperfect auditory output. To describe this another way, the human sound production mechanism is quite complex, but the speech communications system is massively more so. The difference is that the sound production mechanism is mainly handled as a motor (movement) task by the brain, whereas speech is handled at a higher conceptual level, which ties closely with our thoughts. Perhaps that also goes some way towards explaining why thoughts can sometimes be ‘heard’ as a voice or voices inside our heads?

For decades, researchers have been attempting to understand and model both the speech production system and the human auditory system (HAS), with partial success in both cases. Models of our physical hearing ability are good, as are models of the types of sounds that we can produce. However, once we consider either speech or the inter-relationship between perceived sounds in the HAS, the situation becomes far more complex. Speech carries with it the difficulties inherent in the natural language processing (NLP) field, as well as the obvious fact that we often do not clearly say what we mean or mean what we (literally) say.

Speech processing itself is usually concerned with the output of the human speech system, rather than the human interpretation of speech and sounds. In fact, whenever we talk of speech or sounds we probably should clarify whether we are concerned with the physical characteristics of the signal (such as frequency and amplitude), the perceived characteristics (such as rhythm, tone, timbre), or the underlying meaning (such as the message conveyed by words, or the emotions). Each of these aspects is a separate but overlapping research field in its own right.

NLP research considers natural language in all its beauty, linguistic, dialectal and speaker-dependent variation and maddening imperfect complexity. This is primarily a computation field that manipulates symbolic information like phonemes, rather than the actual sounds of speech. It overlaps with linguistics and grammar at one extreme, and speech processing at the other.

Psychoacoustics links the words *psycho* and *acoustics* together (from the Greek ψυχή and ἀκούω respectively) to mean the human interpretation of sounds – specifically as this might differ from a purely physical measurement of the same sounds. This encompasses the idea of auditory illusions, which are analogous to optical illusions for our ears, and form a fascinating and interesting area of research. A little more mundane, but far more impactful, is the computer processing of physical sounds to determine how a human would hear them. Such techniques form the mainstay of almost all recordings and reproductions of music on portable, personal and computational devices.

Automatic speech recognition, or ASR, is also quietly impacting the world to an increasing extent as we talk to our mobile devices and interact with automated systems by telephone. Whilst we cannot yet hold a meaningful conversation with such systems (although this does rather depend upon one's interpretation of the word 'meaningful'), at the time of writing they are on the cusp of actually becoming useful. Unfortunately I realise now that I had written almost the same words five years ago, and perhaps I will be able to repeat them five years from now. However, despite sometimes seemingly glacially slow performance improvements in ASR technology from a user's perspective, the adoption of so-called 'big data' techniques has enabled a recent quantum leap in capabilities.

Broadly speaking, 'big data' is the use of vast amounts of information to improve computational systems. It generally ties closely to the field of machine learning. Naturally, if researchers can enable computers to learn effectively, then they require material from which to learn. It also follows that the better and more extensive the learning material, the better the final result. In the speech field, the raw material for analysis is usually recordings of the spoken word.

Nowhere is the 'big data' approach being followed more enthusiastically than in China, which allies together the world's largest population with the ability to centralise research, data capture and analysis efforts. A research-friendly (and controversial) balance between questions of privacy and scientific research completes the picture. As an illustration, consider the world's second-biggest speech-related company, named

Cambridge University Press
978-1-107-08546-6 - Speech and Audio Processing: A Matlab®-Based Approach
Ian Vince Mcloughlin
Frontmatter
More information

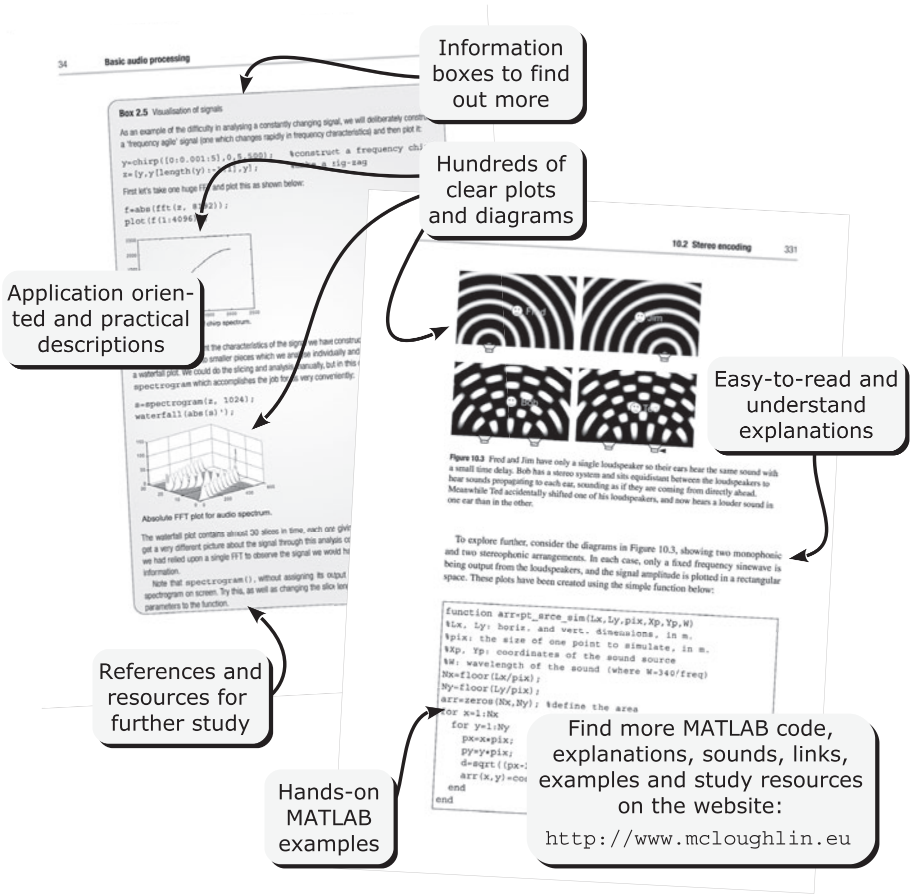
iFlytek, which we will discuss in Chapter 8. Although largely unknown outside China, the flagship product of this impressive company is a smartphone application that understands both English and Chinese speech, acting as a kind of digital personal assistant. This cloud-based system is used today by more than 130 million people, who find it a useful, usable and perhaps invaluable tool. During operation, the company tracks both correct and incorrect responses, and incorporates the feedback into their machine learning model to continuously improve performance. So if, for example, many users input some speech that is not recognised by the system, this information will be used to automatically modify and update their recognition engine. After the system is updated – which happens periodically – it will probably have learned the ability to understand the speech that was not recognised previously. In this way the system can continuously improve as well as track trends and evolutions in speech patterns such as new words and phrases. Launched publicly a few years ago with around 75% recognition accuracy for unconstrained speech without excessive background noise, it now achieves over 95% accuracy, and still continues to improve.

It is approaches like that which are driving the speech business forward today, and which will ensure a solid place in the future for such technologies. Once computers can reliably understand us through speech, speech will quickly become the primary human–computer interfacing method – at least until thought-based (mind-reading) interfaces appear.

This book appears against the backdrop of all of these advances. In general, it builds significantly upon the foundation of the author's previous work, *Applied Speech and Audio Processing with MATLAB Examples*, which was written before big data and machine learning had achieved such significant impact in these fields. The previous book also predated wearable computers with speech interfaces (such as Google's Glass), and cloud-based speech assistants such as Apple's Siri or iFlytek's multilingual system. However, the hands-on nature and practical focus of the previous book are retained, as is the aim to present speech, hearing and audio research in a way that is inspiring, fun and fresh. This really is a good field to become involved in right now. Readers will find that they can type in the MATLAB examples in the book to rapidly explore the topics being presented, and quickly understand how things work.

Also in common with the previous work, this text does not labour over meaningless mathematics, does not present overly extensive equations and does not discuss dreary theory, all of which can readily be obtained elsewhere if required. In fact, any reader wishing to delve a little deeper may refer to the list of references to scientific papers and general per-chapter bibliographies that are provided. The references are related to specific items within the text, whereas the bibliography tends to present useful books and websites that are recommended for further study.

Book features



Information boxes to find out more

Hundreds of clear plots and diagrams

Application oriented and practical descriptions

Easy-to-read and understand explanations

References and resources for further study

Hands-on MATLAB examples

Find more MATLAB code, explanations, sounds, links, examples and study resources on the website:
http://www.mcloughlin.eu

Box 0.1 What is an **information box**?

Self-contained items of further interest, useful tips and reference items that are not within the flow of the main text are presented inside boxes similar to this one.

Each chapter begins with an **introduction** explaining the thrust and objectives being explored in the subsequent sections. MATLAB **examples** are presented and explained throughout to illustrate the points being discussed, and provide a core for further self-directed exploration. Numbered **citations** are provided to support the text (e.g. [1]) where appropriate. The source documents for these can be found listed sequentially from page 370. A **bibliography** is also provided at the end of each chapter, giving a few selected reference texts and resources that readers might find useful for further exploration of the main topics discussed in the text.

Note that commands for MATLAB or computer entry are written in a `computer font` and code listings are presented using a separate highlighted listing arrangement:

```
This is Matlab code.
You can type these commands into MATLAB.
```

All of the commands and listings are designed to be typed at the command prompt in the MATLAB command window. They can also be included and saved as part of an m-file program (this is a sequence of MATLAB commands in a text file having a name ending in .m). These can be loaded into MATLAB and executed – usually by double clicking them. This book does not use Simulink for any of the examples since it would obscure some of the technical details of the underlying processes, but all code can be used in Simulink if required. Note also that the examples only use the basic inbuilt MATLAB syntax and functions wherever possible. However, new releases of MATLAB tend to move some functions from the basic command set into specialised toolboxes (which are then available at additional cost). Hence a small number of examples may require the Signal Processing or other toolboxes in future releases of MATLAB, but if that happens a Google search will usually uncover free or open source functions that can be downloaded to perform an equivalent task.

Companion website

A companion website at http://mcloughlin.eu has been created to link closely with the text. The table at the top of the next page summarises a few ways of accessing the site using different URLs.

An integrated search function allows readers to quickly access topics by name. All code given in the book (and much more) is available for download.

URL	Content
mcloughlin.eu/speech	Main book portal
mcloughlin.eu?s=Xxx	Jump to information on topic Xxx
mcloughlin.eu/chapterN	Chapter *N* information
mcloughlin.eu/listings	Directory of code listings
mcloughlin.eu/secure	Secure area for lecturers and instructors
mcloughlin.eu/errata	Errata to published book

Book preparation

This book has been written and typeset with LaTeX using **TeXShop** and **TeXstudio** front-end packages on Linux Ubuntu and OS-X computers. All code examples have been developed on MATLAB, and most also tested using **Octave**, both running on Linux and OS-X. Line diagrams have all been drawn using the OpenOffice/LibreOffice drawing tool, and all graphics conversions have made use of the extensive graphics processing tools that are freely available on Linux. Audio samples have either been obtained from named research databases or recorded directly using Zoom H4n and H2 audio recorders, and processed using Audacity.

MATLAB® and Simulink are registered trademarks of MathWorks, Inc. All references to MATLAB throughout this work should be taken as referring to MATLAB®.

Acknowledgements

Anyone who has written a book of this length will know the amount of effort involved. Not just in the writing, but in shuffling various elements around to ensure the sequence is optimal, in double checking the details, proofreading, testing and planning. Should a section receive its own diagram or diagrams? How should they be drawn and what should they show? Can a succinct and self-contained MATLAB example be written – and will it be useful? Just how much detail should be presented in the text? What is the best balance between theory and practice, and how much background information is required? All of these questions need to be asked, and answered, numerous times during the writing process. Hopefully the answers to these questions, that have resulted in this book, are right more often than they are wrong.

The writing process certainly takes an inordinate amount of time. During the period spent writing this book, I have seen my children Wesley and Vanessa grow from being pre-teens to young adults, who now have a healthy knowledge of basic audio and speech technology, of course. Unfortunately, time spent writing this book meant spending less time with my family, although I did have the great privilege to be assisted by them: Wesley provided the cover image for the book,[1] and Vanessa created the book index for me.

Apart from my family, there are many people who deserve my thanks, including those who shaped my research career and my education. Obviously this acknowledgement begins chronologically with my parents, who did more than anything to nurture the idea of an academic engineering career, and to encourage my writing. Thanks are also due to my former colleagues at The University of Science and Technology of China, Nanyang Technological University, School of Computer Engineering (Singapore), Tait Electronics Ltd (Christchurch, New Zealand), The University of Birmingham, Simoco Telecommunications (Cambridge), Her Majesty's Government Communications Centre and GEC Hirst Research Centre. Also to my many speech-related students, some of whom are now established as academics in their own right. I do not want to single out names here, because once I start I may not be able to finish without listing everyone, but I do remember you all, and thank you sincerely.

However, it would not be fair to neglect to mention my publishers. Particularly the contributions of Publishing Director Philip Meyler at Cambridge University Press,

[1] The cover background is a photograph of the famous Huangshan (Yellow Mountains) in Anhui Province, China, taken by Wesley McLoughlin using a Sony Alpha-290 DSLR.

Editor Sarah Marsh, and others such as the ever-patient Assistant Editor Heather Brolly. They were responsive, encouraging, supportive and courteous throughout. It has been a great pleasure and privilege to have worked with such a professional and experienced publishing team again.

Finally, this book is dedicated to the original designer of the complex machinery that we call a human: the architect of the amazing and incomparable speech production and hearing apparatus that we often take for granted. All glory and honour be to God.