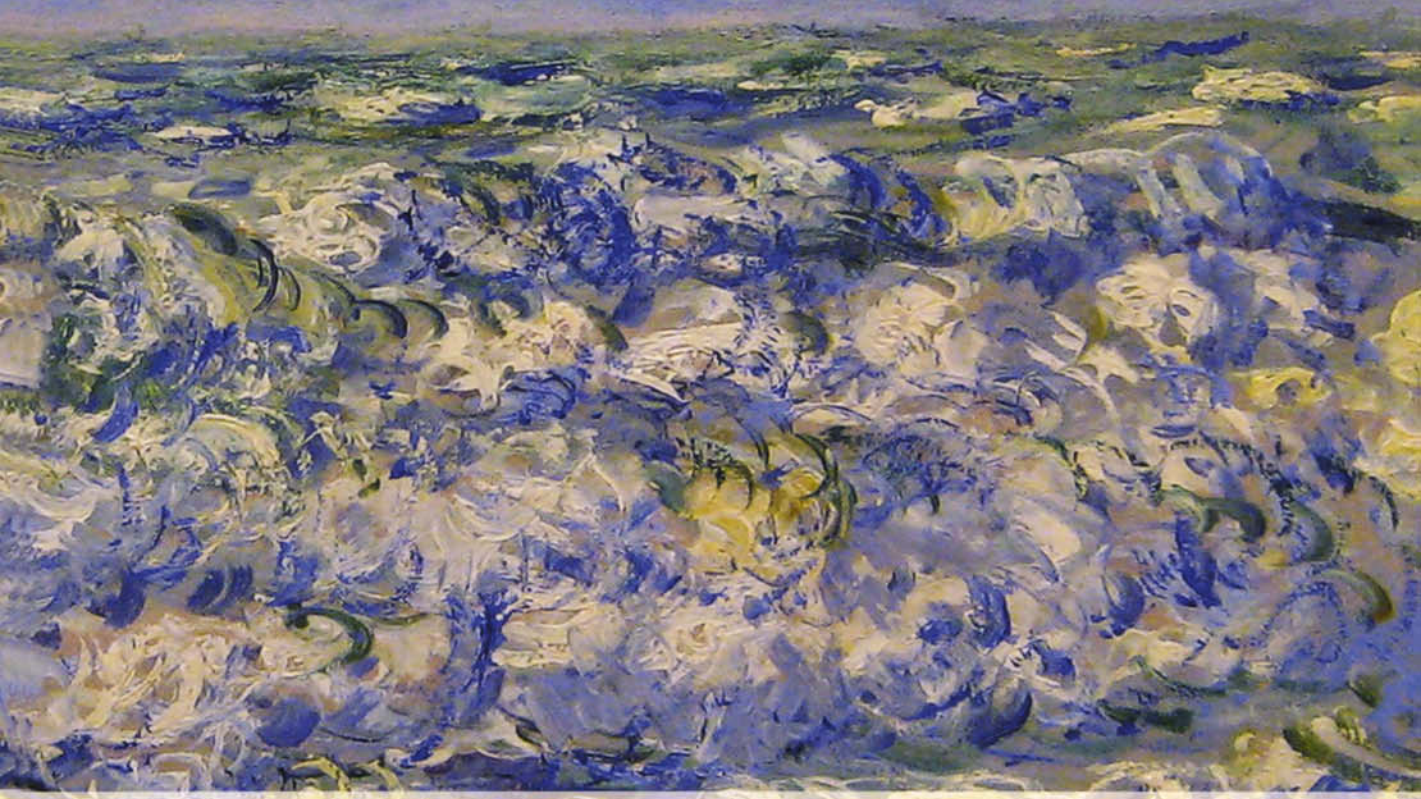
Speech and Audio Signal Processing

Processing and Perception of Speech and Music



SECOND EDITION

Ben Gold • Nelson Morgan • Dan Ellis

WILEY

This page intentionally left blank

SPEECH AND AUDIO
SIGNAL PROCESSING

This page intentionally left blank

SPEECH AND AUDIO SIGNAL PROCESSING

Processing and Perception of Speech and Music

Second Edition

BEN GOLD
Massachusetts Institute of Technology
Lincoln Laboratory

NELSON MORGAN
International Computer Science Institute
and University of California at Berkeley

DAN ELLIS
Columbia University
and International Computer Science Institute

with contributions from:

Hervé Bourlard
Eric Fosler-Lussier
Gerald Friedland
Jeff Gilbert
Simon King
David van Leeuwen
Michael Seltzer
Steven Wegman

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data is available.

ISBN 978-0-470-19536-9

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

This book is dedicated to
our families
and our students

This page intentionally left blank

CONTENTS

PART III

ACOUSTICS

PART VII
SYNTHESIS AND CODING

PREFACE TO THE 2011 EDITION

0.1 WHY WE CREATED A NEW EDITION

Technology moves at a dizzying pace; however, progress can actually seem quite slow in any area that we are deeply involved in. Conference proceedings are filled with incremental advances over previous methods, and entirely novel (and successful) approaches to speech and audio processing are rare. But a lot can happen in a decade, and it has. In addition to quite new methods, there are also many ideas that had not really been refined enough to show progress in the 1990s, but which now are in common use. For instance, Maximum Mutual Information methods, which were developed for ASR many years ago and were briefly described in the previous edition of this book, was significantly refined in the last decade, and the newer versions of this approach are now widely used. Consequently, we devoted new sections of this revision to MMI (and related methods like MPE).

These advances might have been sufficient to warrant an update of our textbook, but there were other reasons as well. A decade of teaching with the book has revealed a number of bugs and deficiencies, and a new edition affords us the opportunity to correct them. For instance, the previous version had nothing about sound source separation, an area that has received considerable attention in the last decade. Approaches to the coding, transcription, and retrieval of music are also now significant areas of audio signal processing, and were not originally covered in the book.

Last, and not least, the new edition has the benefit of a fresh look at the overall subject from our new co-author, Professor Dan Ellis from Columbia University. This hand-off is a key step in keeping the text current.

As with the previous edition, we've attempted to keep the overall style consistent, focusing on what we think is essential, and leaving many details for other publications. We hope that this choice has helped to make the text useful for many readers.

0.2 WHAT IS NEW

As noted above, we have edited and modified many of the chapters, but we also have added entirely new ones. These are:

- Acoustic model training: further topics – MAP and MLLR adaptation methods, and on MMI and MPE discriminant training (Chapter 28, by new contributor Steven Wegmann of Cisco and ICSI).

- Perceptual Audio Coding – MPEG audio and the related psychoacoustics (Chapter 35, by Dan Ellis).
- Music Signal Analysis – automatic transcription of music (Chapter 37, by Dan Ellis).
- Music Retrieval – music retrieval, including cover song detection (Chapter 38, by Dan Ellis).
- Source Separation – methods to separate different signals, including CASA and multi-microphone methods (Chapter 39, by Dan Ellis, with a section on microphone arrays by Michael Seltzer of Microsoft Research).
- Speaker Diarization – determining who spoke when (Chapter 42, by new contributor Gerald Friedland of ICSI).

Two other chapters have essentially been entirely rewritten: Speech Synthesis (Chapter 30, by Simon King from Edinburgh University), and Speaker Verification (Chapter 41, by David van Leeuwen from TNO). Also, Eric Fosler (of Ohio State University) has extensively revised his chapter on Linguistic Categories for Speech Recognition (Chapter 23).

Many other chapters have also undergone significant revisions; for instance, there are a number of significant updates to the chapters on ASR history (Chapter 4) and on feature extraction for ASR (Chapter 22), and a brief description of the Support Vector Machine (SVM) has been added to the deterministic pattern classification chapter (Chapter 8) in recognition of its increased importance. Finally, the Introduction has been modified to reflect the new distribution of chapters.

0.3 A FINAL THOUGHT

Ben Gold was the key inspiration and co-author for the first edition; there clearly would have been no book without him. He also was an inspiration and role model for me (Morgan) personally. It saddens me that he cannot be here for the new edition, but I know that his generous spirit would have welcomed the new contributions from Dan Ellis and others.

CHAPTER 1

INTRODUCTION

We are confronted with insurmountable opportunities.
—Walt Kelly

1.1 WHY WE WROTE THIS BOOK

Speech and music are the most basic means of adult human communication. As technology advances and increasingly sophisticated tools become available to use with speech and music signals, scientists can study these sounds more effectively and invent new ways of applying them for the benefit of humankind. Such research has led to the development of speech and music synthesizers, speech transmission systems, and automatic speech recognition (ASR) systems. Hand in hand with this progress has come an enhanced understanding of how people produce and perceive speech and music. In fact, the processing of speech and music by devices and the perception of these sounds by humans are areas that inherently interact with and enhance each other.

Despite significant progress in this field, there is still much that is not well understood. Speech and music technology could be greatly improved. For instance, in the presence of unexpected acoustic variability, ASR systems often perform much worse than human listeners (still!). Speech that is synthesized from arbitrary text still sounds artificial. Speech-coding techniques remain far from optimal, and the goal of transparent transmission of speech and music with minimal bandwidth is still distant. All fields associated with the processing and perception of speech and music stand to benefit greatly from continued research efforts. Finally, the growing availability of computer applications incorporating audio (particularly over the Internet and in portable devices) has increased the need for an ever-wider group of engineers and computer scientists to understand audio signal processing. For all of these reasons, as well as our own need to standardize a text for our graduate course at UC Berkeley, we wrote this book; and for the reasons noted in the Preface, we have updated it for the current edition.

The notes on which this book is based proved beneficial to graduate students for close to a decade; during this time, of course, the material evolved, including a problem set for each chapter. The material includes coverage of the physiology and psychoacoustics of hearing as well as the results from research on pitch and speech perception, vocoding methods, and information on many aspects of ASR. To this end, the authors have made use of their own research in these fields, as well as the methods and results of many other

contributors. And as noted in the Preface, this edition includes contributions from new authors as well, in order to broaden the coverage and bring it up to date.

In many chapters, the material is written in a historical framework. In some cases, this is done for motivation's sake; the material is part of the historical record, and we hope that the reader will be interested. In other cases, the historical methods provide a convenient introduction to a topic, since they often are simpler versions of more current approaches. Overall, we have tried to take a long-term perspective on technology developments, which in our view requires incorporating a historical context. The fact that otherwise excellent books on this topic have typically avoided this perspective was one of our major motivations for writing this book.

1.2 HOW TO USE THIS BOOK

This text covers a large number of topics in speech and audio signal processing. While we felt that such a wide range was necessary, we also needed to present a level of detail that is appropriate for a graduate text. Therefore, we have elected to focus on basic material with advanced discussion in selected subtopics. We have assumed that readers have prior experience with core mathematical concepts such as difference equations or probability density functions, but we do not assume that the reader is an expert in their use. Consequently, we will often provide a brief and selected introduction to these concepts to refresh the memories of students who have studied the background material at one time but who have not used it recently. The background topics are selected with a particular focus, namely, to be useful to both students and working professionals in the fields of ASR and speaker recognition, speech bandwidth compression, speech analysis and synthesis, and music analysis and synthesis. Topics from the areas of digital signal processing, pattern recognition, and ear physiology and psychoacoustics are chosen so as to be helpful in understanding the basic approaches for speech and audio applications.

The remainder of this book comprises 41 chapters, grouped into eight sections. Each section or part consists of three to seven chapters that are conceptually linked. Each part begins with a short description of its contents and purpose. These parts are as follows:

I. **Historical Background.** In Chapters 2 through 5 we lay the groundwork for key concepts to be explored later in the book, providing a top-level summary of speech and music processing from a historical perspective. Topics include speech and music analysis, synthesis, and speech recognition.

II. **Mathematical Background.** The basic elements of digital signal processing (Chapters 6 and 7) and pattern recognition (Chapters 8 and 9) comprise the core engineering mathematics needed to understand the application areas described in this book.

III. **Acoustics.** The topics in this section (Chapters 10–13) range from acoustic wave theory to simple models for acoustics in human vocal tracts, tubes, strings, and rooms. All of these aspects of acoustics are significant for an understanding of speech and audio signal processing.

IV. Auditory Perception. This section (Chapters 14–18) begins with descriptions of how the outer ear, middle ear, and inner ear work; most of the available information comes from experiments on small mammals, such as cats. Insights into human hearing are derived from experimental psychoacoustics. These fundamentals then lead to the study of human pitch perception as applied to speech and music, as well as to studies of human speech perception and recognition. Some of these topics are further developed in Chapters 34 and 35 in the context of perceptual audio coding.

V. Speech Features. Systems for ASR and vocoding have nearly always incorporated filter banks, cepstral analysis, linear predictive coding, or some combination of these basic methods. Each of these approaches has been given a full chapter (19–21).

VI. Automatic Speech Recognition. Eight chapters (22–29) are devoted to this study of ASR. Topics range from feature extraction to statistical and deterministic sequence analysis, with coverage of both standard and discriminant training of hidden Markov models (including neural network approaches). A new chapter (Chapter 28) updates the book to include now-standard adaptation techniques, as well as further explanation of discriminant training techniques that are commonly used. Part VI concludes with an overview of a complete ASR system.

VII. Synthesis and Coding. Speech synthesis (culminating in text-to-speech systems) is first presented in Chapter 30, a chapter that has largely been rewritten to emphasize concatenative and HMM-based techniques that have become dominant in recent years. Chapter 31 is devoted to pitch detection, which applies to both speech and music devices. Many aspects of vocoding systems are then described in Chapters 32–34, ranging from very-high-quality systems working at relatively high bit rates to extremely low-rate systems. Finally, Chapter 35 provides a description of perceptual audio coding, now used for consumer music systems.

VIII. Other Applications. In Chapters 36–42 we present several application areas that were not covered in the bulk of the book. Chapter 36 is a review of major issues in music synthesis. Chapter 37 introduces the transcription of music through several kinds of signal analysis. Chapter 38 is focused on methods for identifying and selecting musical selections. Chapter 39 introduces the topic of source separation, which ultimately could be the critical step in bringing many other applications to a human level of performance, since most desired sounds in the real world exist in the context of other sounds occurring simultaneously. Modifications of the time scale, pitch, and spectral envelope can transform speech and music in ways that are increasingly finding common applications (Chapter 40).

Chapter 41 is an overview of speaker recognition, with an emphasis on speaker verification. With increasing access to electronic information and expansion of electronic commerce, verification of the identity of a system user is becoming increasingly important. This chapter has largely been rewritten to reflect the signficant progress that has occurred in this field since 1999. A related area, speaker diarization, is the topic of the final chapter (42). An area of significant commercial and public interest is the labeling of multiparty conversations (such as a technical meeting) with what is sometimes called a rich transcription, which includes not only the

sequence of words, but also other automatic annotations such as the attribution of which speaker is speaking when; this latter capability is often referred to as speaker diarization.

Readers with sufficient background may choose to focus on the application areas described in Parts V–VIII, as the first four parts primarily give preparatory material. However, in our experience, readers at a graduate or senior undergraduate level in electrical engineering or computer science will benefit from the earlier parts as well. In teaching this course, we have also found the problem sets to be helpful in clarifying understanding, and we suspect that they would have similar value for industrial researchers. Another useful study aid is provided by a collection of audio examples that we have used in our course. These examples have been made freely available via the book's World-Wide Web site which can be found at http://catalog.wiley.com/. This Web site may also be augmented over time to include links to errata and addenda for the book.

Other books on a similar topic but with a different emphasis can also be used to complement the material here; in particular, we recommend [9] or [11]; a more recent book with significant detail on current methods is [4].

Additionally, a more complete exposition on the background material introduced in Parts II–IV can be found in such texts as the following:

- [8] for digital signal processing
- [1] or [3] for pattern recognition (note that this is the revised edition of the classic [2])
- [6] for acoustics
- [10] for auditory physiology
- [7] for psychoacoustics

Finally, an excellent book already in its second edition is [5], which focuses much more on the language-related aspects of speech processing.

1.3 A CONFESSION

The authors have chosen to spend much of their lives studying speech and audio signals and systems. Although we would like to say that we have done this to benefit society, much of the reason for our vocational path is a combination of happenstance and hedonism; in other words, dumb luck and a desire to have fun. We have enjoyed ourselves in this work, and we continue to do so. Speech and audio processing has become a fulfilling obsession for us, and we hope that some of our readers will adopt and enjoy this obsession too.

1.4 ACKNOWLEDGMENTS

Many other people contributed to this book. Students in our graduate class at Berkeley contributed greatly over the years, both by their need for a text and by their original scribe notes that inspired us to write the book. Two (former) students in particular, Eric Fosler-Lussier and Jeff Gilbert, ultimately wrote material that was the basis of Chapters 22 and 33, respectively. Hervé Bourlard came through very quickly with the original core of the Speaker Verification chapter when we realized that we had no material on speaker identification or verification, and David van Leeuwen provided the updates for the current version. Simon King wrote a new chapter on Speech Synthesis, and Alan Black provided useful comments and criticism. Steven Wegmann wrote new material on adaptation and discriminant training. John Lazzaro provided useful comments on the new perceptual audio and music chapters, and Mike Seltzer added important material on microphone arrays for the source separation chapter. Finally Martin Cooke helped with his remarks on our draft of the CASA description.

For the original version, Su-Lin Wu was an extremely helpful critic, both on the speech recognition sections and on some of the psychoacoustics material. Anita Bounds-Morgan provided very useful editorial assistance. The anonymous reviewers that our publisher used at that time were also quite helpful.

We certainly appreciate the indulgence of the International Computer Science Institute and Lincoln Laboratory for permitting us to develop the original manuscript on the equipment of these two labs, and for Columbia University for similarly providing the necessary resources for Dan Ellis's extensive efforts to update the current version. Devra Polack and Elizabeth Weinstein of ICSI also provided a range of secretarial and artistic support that was very important to the success of the earlier project. We also thank Bill Zobrist of Wiley for his interest in the original book, and George Telecki for his support for our developing the second edition.

Finally, we extend our special thanks to our wives, Sylvia, Anita, and Sarah, for putting up with our intrusion on family time as a result of all the days and evenings that were spent on this book.

BIBLIOGRAPHY

1. Bishop, C., *Neural Networks for Pattern Recognition*, Oxford Univ. Press, London/New York, 1996.
2. Duda, D., and Hart, P., *Pattern Classification and Scene Analysis*, Wiley–Interscience, New York, 1973.
3. Duda, D., Hart, P., and Stork, D., *Pattern Classification (2nd Ed.)*, Wiley–Interscience, New York, 2001.
4. Huang, X., Acero, A., and Hon, H.-W., *Spoken Language Processing*, Prentice-Hall, Englewood Cliffs, N.J., 2001.
5. Jurafsky, D., and Martin, J., *Speech and Language Processing*, Prentice-Hall, Upper Saddle River, N.J., 2009.

6. Kinsler, L., and Frey, A., *Fundamentals of Acoustics*, Wiley, New York, 1962.
7. Moore, B. C. J., *An Introduction to the Psychology of Hearing*, 5th ed. Academic Press, New York/London, 2003.
8. Oppenheim, A., and Schafer, R., *Discrete-Time Signal Processing (3rd Ed.,* Prentice–Hall, Englewood Cliffs, N.J., 2009.
9. O'Shaughnessy, D., *Speech Communication*, Addison–Wesley, Reading, Mass., 1987.
10. Pickles, J., *An Introduction to the Physiology of Hearing*, Academic Press, New York, 1982.
11. Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition,* Prentice–Hall, Englewood Cliffs, N.J., 1993.