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:

Herv 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 98765432 1

This book is dedicated to
our families
and our students

CONTENTS

CONTENTS

PARTI

HISTORICAL BACKGROUND

viii

x

PART I
ACOUSTICS

xii

17.6
Motor

CHAPTER 24 *DETERMINISTIC SEQUENCE RECOGNITION FOR ASR* **337**

24.1 Introduction **337**
24.2 Isolated Word Recognition **338**
24.2.1 Linear Time Warp **339**
24.2.2 Dynamic Time Warp **340**
24.2.3 Distances **344**
24.2.4 End-Point Detection **344**
24.3 Connected Word Recognition **346**
24.4 Segmental Approaches **347**
24.5 Discussion **348**

PART VII

SYNTHESIS AND CODING

PART Vii

SYNTHESIS AND CODING

Introduction **431**
Concatenative Methods **433**
30.2.1 Database **433**
30.2.2 Unit selection **434**
30.2.3 Concatenation and optional modification **435** Statistical Parametric Methods **436**
30.3.1 Vocoding: from waveforms to features and back **436** 30.3.2 Statistical modeling for speech generation **438** 30.3.3 Advanced techniques **440**
A Historical Perspective **441**
Speculation **443**
30.5.1 Physical models **444**
30.5.2 Sub-word units and the role of linguistic knowledge **445** 30.5.3 Prosody matters **445**
Tools and Evaluation **446**
30.6.1 Further reading **447**
30.7 30.8 Exercises Appendix 30.8.1 30.8.2 30.8.3 30.8.4 30.8.5 30.8.6

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

- 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 (Chap ter 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.

Speech and Audio Signal Processing: Processing and Perception of Speech and Music, Second Edition by Ben Gold, Nelson Morgan and Dan Ellis
John Wiley & Sons, Inc.

CHAPTER f

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 recog
nition (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 un
derstood. 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 transmis
sion 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 process ing. 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

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 expe
rience 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 mem
ories 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 recog
nition, 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 se quence 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 elec tronic commerce, verification of the identity of a system user is becoming increas ingly important. This chapter has largely been rewritten to reflect the signficant

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. Herv 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 In stitute 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 Uni v. 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 (2ndEd.),* Wiley-Interscience, New York, 2001.

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, Engle wood 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.

The future is a lot like the present, only it's longer.
—Dan Quisenberry

***F\N* UNDERSTANDING** of the goals and methods of the past can help us to envision the advances of the future. Ideas are often rediscovered or reinvented many times. Often the new form has much greater impact, though sometimes for fairly mundane reasons, such as greater accessibility to larger computational capabilities. It would be obvious to most that a study of the social sciences would be incomplete without inclusion of the methods and beliefs of the past. In our view, a study of current methods in speech and audio engineering without any historical context would be similarly inadequate.

For these reasons, we introduce the basic concepts of speech analysis, synthesis, and recognition in Part I, using a historical frame of reference. Later parts will provide greater technical detail in each of these areas. We begin in Chapter 2 with a brief history of synthetic audio, starting with 18th Century mechanical devices and proceeding through speech and music machines from the first half of the 20th Century. The discussion continues in Chapter 3 with a discussion of systems for analysis and synthesis, including a brief introduction to the concept of source-filter separation. Speech recognition is a 20th Century invention, and the Chapter 4 discussion of the history of research in this area is largely confined to the past 50 years. Finally, Chapter 5 introduces speech recognition technology, discussing such topics as the major components of a recognizer and the sources of difficulty in this problem. Overall, Part I is intended to provide a light overview that will give the reader motivation for the more detailed material that follows.

Processing: Processing and Perception of Speech and Music, Second Edition by Ben Gold, Nelson Morgan and Dan Ellis

2

SYNTHETIC AUDIO: A BRIEF HISTORY

2.1 VON KEMPELEN

Many years ago, von Kempelen demonstrated that the speech-production system of the human being could be modeled. He showed this by building a mechanical contrivance that "talked." The paper by Dudley and Tarnoczy [2] relates the history of von Kempelen's speaking machine. This device was built about 1780, at a time when the notion of building

automata was quite popular. Von Kempelen also wrote a book [7] that dealt with the origin of speech, the human speech-production system, and his speaking machine. Thus, for over a century, an existence proof was established that one could indeed build a machine that spoke. (Von Kempelen's work brings to mind that of another great innovator, Babbage, who also labored for many years with mechanical contrivances to try to build a computing machine.)

Figure 2.1 shows the speaking machine built by Wheatstone that was based on von Kempelen's work. The resonator of leather was manipulated by the operator to try to copy the acoustic configuration of the vocal tract during the sonorant sounds (vowels, semivowels, glides, and nasals); the bellows provided the air stream; the vibrating reed produced the periodic pressure wave; and the various small whistles and levers shown con trolled most of the consonants. (Much later, Riesz [6] built a mechanical speaking machine that was more precisely modeled after the human speech-producing mechanism. This is depicted in Fig. 2.2, shown here for comparison to the von Kempelen-Wheatstone model of Fig. 2.1).

2.2 THE VODER

Modern methods of speech processing really began in the U.S. with the development of two devices. Homer Dudley pioneered the development of the channel vocoder (voice coder) and the Voder (voice-operated demonstrator) [1]. We know from numerous news paper articles that the appearance of the Voder at the 1939 World's Fair in San Francisco and New York City was an item of intense curiosity. Figure 2.3 is a collage of some clippings from that period and reflects some of the wonder of people at the robot that spoke.

It is important to realize that the Voder did not speak without a great deal of help

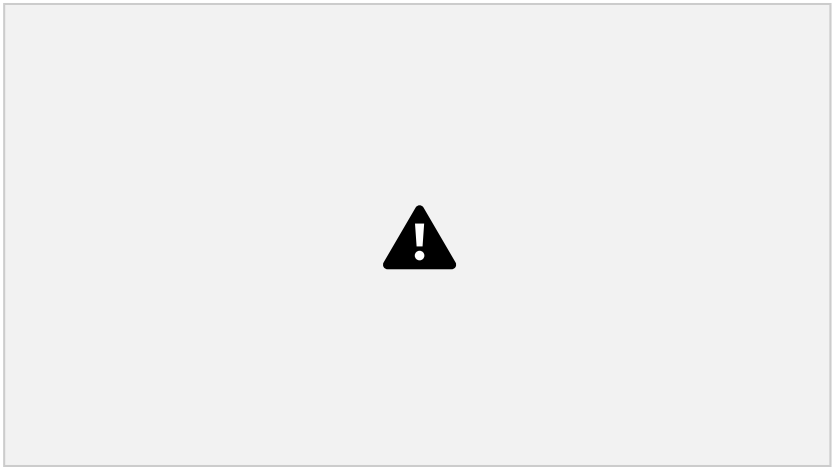


FIGURE 2.1 Wheatstone's speaking machine. From [2].

compared to a piano keyboard. In the background is the electronic device that does the speaking. Operator training proved to be a major problem. Many candidates for this job were unable to learn it, and the successful ones required training for periods of 6 months to 1 year. Figure 2.4 shows an original sketch by S. W. Watkins of the Voder console.

The keys were used to produce the various sounds; the wrist bar was a switch that determined whether the excitation function would be voiced or unvoiced, and the pitch pedal supplied intonation information. Figure 2.5 is a close-up of the controls in the console and shows how these relate to the articulators of a human vocal tract.

The keys marked 1 through 10 control the connection of the corresponding bandpass filters into the system. If two or three of the keys were depressed and the wrist bar was set

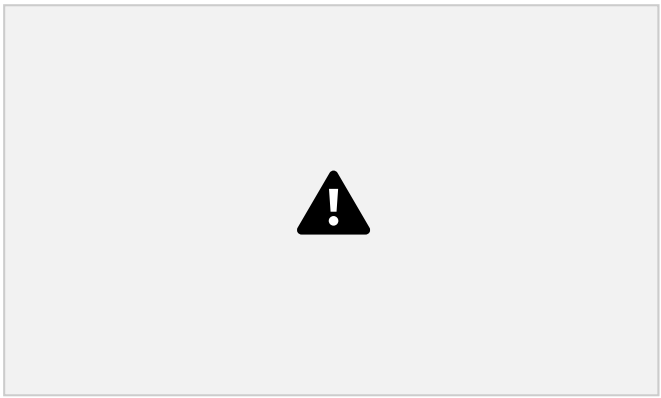


FIGURE 2.2 Riesz's speaking machine. From [3].

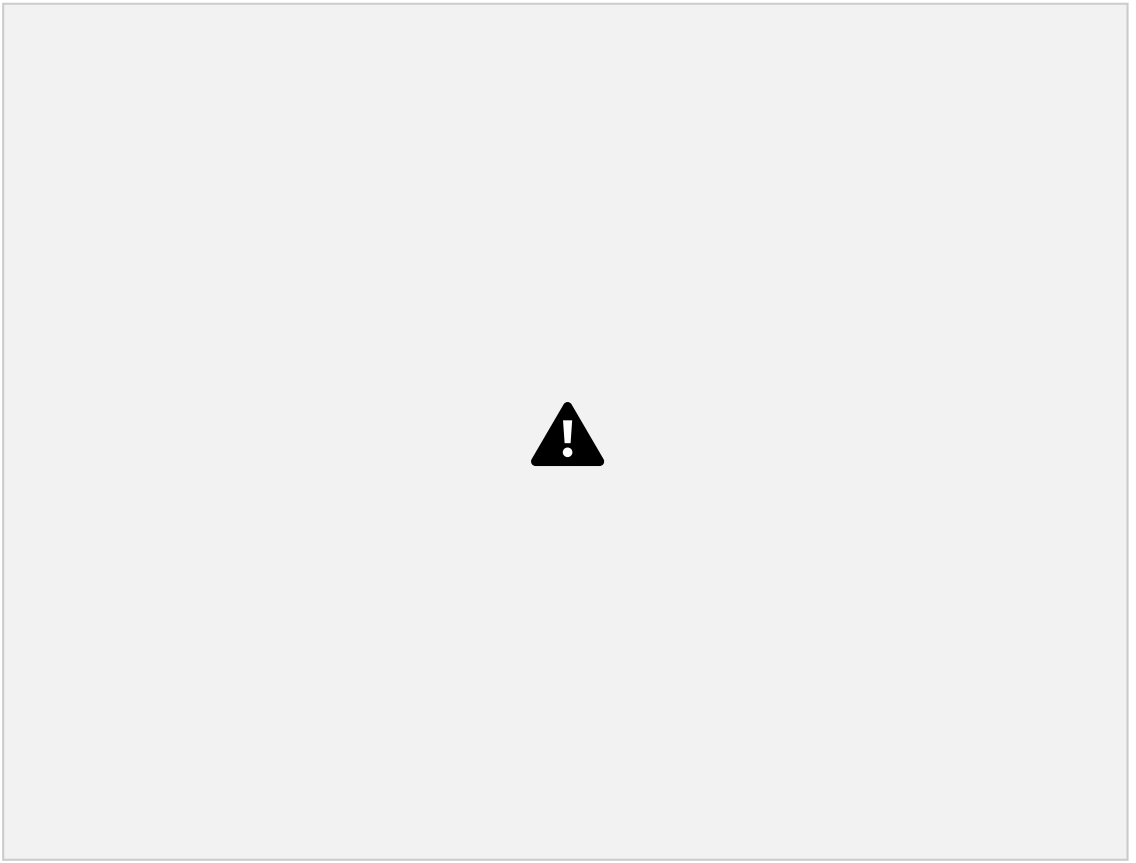


FIGURE 2.3 News clippings on the Voder.

to the buzz (voicing) condition, vowels and nasals were produced. If the wrist bar were set to hiss (voiceless), sounds such as the voiceless fricatives (e.g., f) were generated. Special keys were used to produce the plosive sounds (such as p or d) and the affricate sounds (ch as in cheese; j as in jaw).

2.3 TEACHING THE OPERATOR TO MAKE THE VODER "TALK"

The Voder was marvelous, not only because it "talked" but also because a person could be trained to "play" it. Speech synthesis today is done by real-time computer programs or

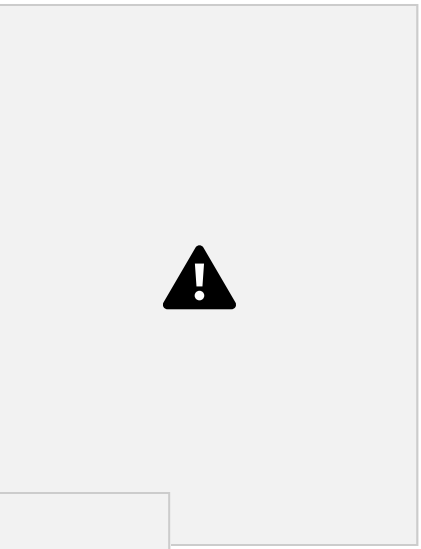
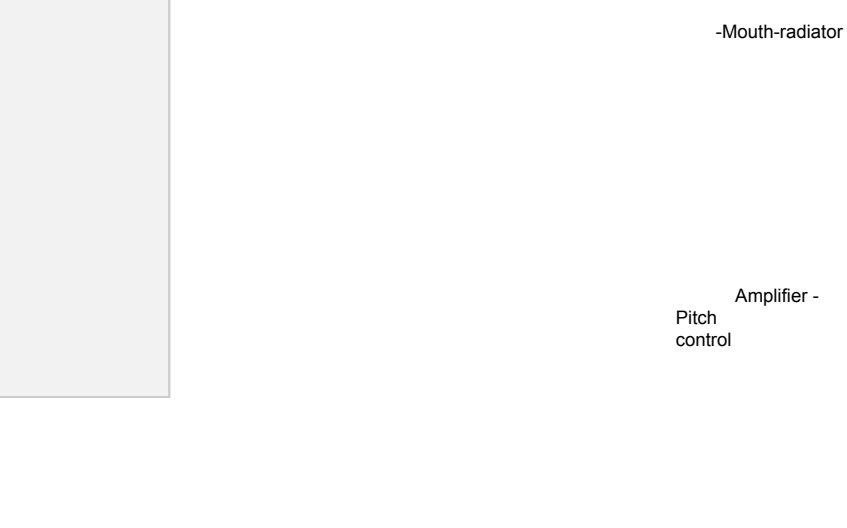


FIGURE 2.4 Sketch of the Voder.



From [2].

FIGURE 2.5 Voder controls.

Sounds	Wrist Bar	Keys
S		9.
Sh	up	7.8. Light &
M	up	Smooth 1.
ë(seen)	down	1.8.
aw (dawn)	down	3.
	down	

She saw me.
See me seesaw.

Think of how they not how they are
WORDS wntten. are pronounced,

Cease Sauce See She
Me Saw Seesaw

FIGURE 2.6 Lesson 1 of the Voder instructions.

parameters. It is a pity that further work on real-time control by a human operator has not been seriously pursued.

Figures 2.6, 2.7, and 2.8 describe Lessons 1, 9, and 37 of the Voder Instruction Manual.

Relatively few of the candidate operators were successful, but one young woman (Mrs. Helen Harper) was very proficient. She performed at the 1939 New York World'sFair. Many years later (in the 1960s) a highlight of Dudley's retirement party was the Voder's speaking to Mr. Dudley, with the help of Mrs. Harper.

2.4 SPEECH SYNTHESIS AFTER THE VODER

Many speech-synthesis devices were built in the decades following the invention of the Voder, but the underlying principle, as captured in Fig. 2.5, has remained quite fixed. For many cases, there is a separation of source and filter followed by the parameterization of each. As we shall see in the following sections, the same underlying principles control the design of most music synthesizers. In later chapters, the field of speech synthesis from the past to the present is explored in some detail, including advanced systems that transform printed text into reasonable-sounding speech.

2.5 MUSIC MACHINES

Figure 2.9 shows a 17th Century drawing of a water-powered barrel organ. Spring-powered barrel organs may have existed as long ago as the 12th Century. Barrel organs work on

	Initial R	
Sounds		Make it safer.
	a (take)	She's sorry for me.
Initial R		Wrist Bar

PRACTICE SENTENCES

Your chef makes rich sauces.

Shake it off.

Is nature fair?

She wrote to Rose.

FIGURE 2.7 Lesson 9.

the same concepts as present-day music boxes; that is, once the positions of the pins are chosen, the same music will be played for each complete rotation. Keys can be depressed or strings can be plucked, depending on the overall design of the automatic instrument.

The barrel organ is a form of read-only memory, and not a very compact form at that. Furthermore, barrel organs could not record music played by a performer. In the late 18th Century, both of these problems were overcome by melography, which allowed music to

Sounds Wrist Bar Keys

Initial Bl

Final -si up—down 9—2

Final -tl up—down 6.8.-2. Very light tap on 6.8.

WORDS Think of how they are pronounced, not how they are spelt.

Ably Blow Mental Whistle

Black Blue Metal

Blade Establish Pestle

Blood Little Total

PRACTICE SENTENCES

The wind blows cold tonight.

Her hair is quite black.

Whistle and I'll come to you.

Make a mental note of it.

What a little dog it is.

How much is the total amount?

FIGURE 2.8 Lesson 37.

be both recorded and played back, using the medium of punched paper tape or cards. The idea originated for the automation of weaving and was developed fully by Joseph Marie Jacquard, who designed a device that could advance and register cards. (Punched cards were used by Babbage in the design of his computing machine and, in our time, were used by many computer manufacturers such as IBM.) Card-driven street organs made use of this technology. Card stacks were easy to duplicate; also, different stacks contained different music, so that music machines became very marketable. By the beginning of the 20th century, the concept had been applied to the player piano. A roll of paper tape could be made and the holes punched automatically while a master pianist (such as Rachmaninoff or Gershwin) played. This paper roll could then actuate the playback mechanism to produce the recorded version. Since the piano keys were air driven, extra perforations in the paper roll allowed variable amounts of air into the system, thus changing volume and attack in a way comparable to that of the human performer. Until

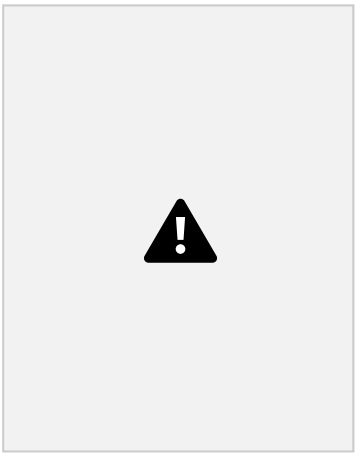


FIGURE 2.9 17th Century drawing of a water-powered barrel organ.

MIT Media Laboratory. Using this system, Fu [4] synthesized a Bosendorfer version from an old piano roll by Rachmaninoff.

At the beginning of the 20th century, a mighty device called the telharmonium was constructed by Thaddeus Cahill. Remember that this was built *before* the development of electronics; nevertheless, Cahill had the ingenuity to realize that any sound could be synthesized by the summation of suitably weighted sinusoids. He implemented each si nusoid by actuating a generator. To create interesting music, many such generators (plus much additional equipment) were needed, so the result was a monster, weighing many tons. Cahill's concept of additive synthesis is still an important feature of much of the work in electronic music synthesis. This is in contrast to many later music synthesizers that employ subtractive synthesis, in which adaptive filtering of a wideband excitation function generates the sound. (The additive synthesis concept was used by McCaulay and Quatieri [5] to design and build a speech-analysis-synthesis system; we discuss this device in later chapters.)

The player piano is only partially a music machine, since it requires a real piano to be part of the system. The telharmonium, by contrast, is a complete synthesizer, since music is made from an abstract model, that is, sine generators. Another, although totally different, complete synthesizer is the theremin, named after its inventor, the Russian Lev Termin. In this system, an antenna is a component of an electronic oscillator circuit; moving one's arm near the antenna changes the oscillator frequency by changing the capacitance of the circuit, and this variable frequency is mixed with a fixed-frequency oscillator to produce an audio tone whose frequency can be varied by arm motion. Thus the theremin generates a nearly sinusoidal sound but with a variable frequency that can produce pitch perceptions that don't exist in any standard musical scale. In the hands of a trained performer, the theremin produces rather unearthly sounds that are nevertheless identifiable as some sort of (strange) music. A trained performer could play recognizable music (e.g., Schubert's Ave Maria). Figure 2.10 shows Clara Rockmore at a theremin. Her right hand controls the

frequency of the straight antenna while her left hand controls the amplitude by changing the capacitance of a different circuit.

The theremin continues to fascinate. In 1994 a film called "Theremin: An Elec tronic Odyssey" was released, leading to the sale of more than one thousand instruments

the following year. In 2004, Moog Music, the doyen of the electronic music industry, released the Etherwave Theremin Pro. This is but the latest in a long line of theremins they have marketed, and is a consistent favorite for live performances.

2.6 EXERCISES

2.1 The Voder was which of the following:

(a) a physical model of the human vocal apparatus,
(b) an early example of subtractive synthesis,
(c) an early example of additive synthesis, or
(d) a member of the electorate with a head cold.

2.2 Shown in Fig. 2.11 is Dudley's speech-sound classification for use with Voder training. Find the Voder sequence for any of the practice sentences of Fig. 2.8 (Lesson 37). Break the sentence into a phoneme sequence, using the notation of Fig. 2.11. Note that the BK1, BK2, and BK3 keys in Fig. 2.11 are the k g, p-b, and t-d keys of Fig. 2.5. A sample is shown below for the sentence "The Voder can speak well."

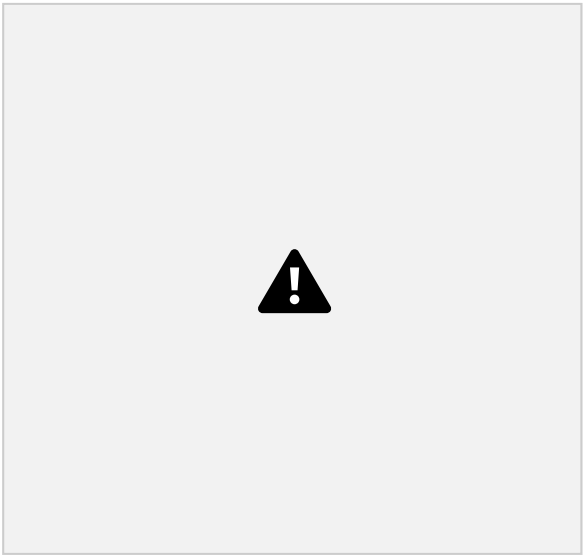


FIGURE 2.10 Clara Rockmore at the theremin.

1. Vowels

(tee) i I 8)

)

k! и (3488) (nut) Gj /
i (it! (28)
JS (apeH37"28) e (tenJ(37)
(at) (457)

(458) a (far) (458!

(Numbers in parentheses indicate key designations on VODER consoie)

FIGURE 2.11 Classification of speech sounds for Voder use.

FIGURE 2.12 Spectrogram of "greetings everybody" by an announcer.

è'K **FIGURE 2.13** Spectrogram of "greetings everybody" by the Voder.

/d/ /r/ Voiced
Voiced 10Q
Voiced 3458

/th/ /ü/ Voiced 67Q 3-2
Voiced BK3 36
/v/ /o/ Voiced

/k/ Unvoiced BK1 /ä/
Voiced 457 /n/ Voiced 1
Voiced BK2 18
/s/ /p/ /e/ Unvoiced BK1
/k/

/w/ /e/ Unvoiced 37
/1/ Voiced 2
Unvoiced 9
Unvoiced

Voder example: "The Voder can speak well."

2.3 Compare von Kempelen's speaking machine with Dudley's Voder.

(a) What are the chief differences?

(b) What are the chief similarities?

(c) How would you build a von Kempelen machine today?

2.4 Figures 2.12 and 2.13 show spectrograms of the saying "greetings everybody" by the announcer and the Voder.

(a) What do you perceive to be the main difference between the natural and the synthetic utterances? (b) Estimate the instants when the operator changes the Voder configuration.

(a) telharmonium,

(b) Wheatstone-von Kempelen speaking machine,

(c) Voder,

(d) theremin; and

(e) player piano.

BIBLIOGRAPHY

1. Dudley, H., Riesz, R., and Watkins, S., "A synthetic speaker," *J. Franklin Inst.* 227: 739, 1939.
2. Dudley, H., and Tarnoczy, T. H., "The speaking machine of Wolfgang von Kempelen," *J. Acoust. Soc. Am.* 22: 151-166, 1950.
3. Flanagan, J. L., *Speech Analysis Synthesis and Perception,* 2nd ed., Springer-Verlag, New York/Berlin, 1972.
4. Fu, A. C, "Resynthesis of acoustic piano recordings," M.S. Thesis, Massachusetts Institute of Technology, 1996.
5. McAulay, R. J., and Quatieri, T. R, "Speech analysis/synthesis based on a sinusoidal representa tion," *IEEE Trans. Acoust. Speech Signal Process.* 34(4): 744-754, Aug. 1986.
6. Riesz, R. R., personal communication to J. L. Flanagan, 1937. (Details of this work are described by Flanagan in [3, pages 207-208].)
7. Von Kempelen, W., *Le Mechanisme de lapavola, suivi de la Description d'une machine parlante.* Vienna: J.V. Degen, 1791.

Speech and Audio Signal Processing: Processing and Perception of Speech and Music, Second Edition by Ben Gold, Nelson Morgan and Dan Ellis

CHAPTER 3

SPEECH ANALYSIS AND SYNTHESIS OVERVIEW

"If I could determine what there is in the very rapidly changing complex speech wave that corresponds to the simple motion of the lips and tongue, if I could then analyze speech for these quantities, I would have a set of speech defining signals that could be handled as low frequency telegraph currents with resulting advantages of secrecy, and more
telephone channels in the same frequency space as well as a basic understanding of the carrier nature of speech by which the lip reader interprets speech from simple

motions."

-Homer Dudley, 1935

3.1 BACKGROUND

If we think, for the moment, of speech as being a mode of transmitting word messages, and telegraphy as simply another mode of performing the same action, this immediately allows us to conclude that the intrinsic information rate of speech is exactly the same as that of a telegraph signal generating words at the same average rate. Speech, however, conveys emphasis, emotion, personality, etc., and we still don't know how much bandwidth is needed to transmit these kinds of information.

In the following sections, we begin with some further historical background on speech communication.

3.1.1 Transmission of Acoustic Signals

Perhaps the earliest network for speech communication at long distances was a system that we'll call the "stentorian network," which was used by the ancient Greeks. It consisted of towers and men with very loud voices. The following excerpts were found in Homer Dudley's archives:

> *Homer has written that the warrior Stentor, who was at the siege of Troy, had such a loud voice that it made more noise than fifty men all shouting at once. Alexander the Great (356-325 B.C.) seems to have had a method whereby a stentor's voice could be heard by the whole army. Did it consist of acoustical signals which were repeated from one soldier crier to another, organized as a transmitting group ?*

for example, the massacre of the Romans which took place at Orleans at sunrise was known at nine o'clock the same evening at Auvergne, forty miles away.

Diodorus of Sicily, a Greek historian living in the age of Augustus, said that at the order of the King of Persia, sentinels, who shouted the news which they wished to transmit to distant places were stationed at intervals throughout the land. The transmission time was 48 hours from Athens to Susa, over 1500 miles apart.

We also note that, in addition to speech, flare signals were used as a communications medium. The Appendix to this chapter illustrates this with an excerpt from the Greek play *Agamemnon* (by Aeschylus) that describes the transmission of information about the fall of Troy (also see Fig. 3.13 in the Appendix).

3.1.2 Acoustical Telegraphy before Morse Code

The Dudley archives provide some fascinating examples of pre-Morse code communica tions:

Later a group of inventors, among whom we find Kircher(1601-1680), Scheventer (1636) and the two Bernoulli brothers, sought to transmit news long distances by means of musical instruments each note representing a letter. One of the Bernoullis devised an instrument, composed of five bells, which permitted the principal letters of the alphabet to be transmitted.

It is told that the King of England was able to hear news transmitted 1.5 English miles to him by means of a trumpet. He had this trumpet taken to Deal Castle, whose commander said that this instrument permitted a person to make himself understood over a distance of three nautical miles. It was invented by the "genial mechanic" of Hammersmith, Sir Samuel Morland (1626-1696). It's [sic] mouthpiece was designed so that no sound could escape from either end. Morland published a treatise on this instrument entitled "Tube Stentorophonica" and in 1666 he wrote a report on "a new cryptographic process. "

In 1762 Benjamin Franklin experimented with transmitting sound under water. In 1785 Gauthoy and Biot transmitted words through pipes for a distance of 395 meters. But at a distance of 951 meters speech was no longer intelligible.

We can also regard the ringing of bells as acoustical telegraphy or telephony, if we consider that in certain Swiss villages the inhabitants recognize from their tone whether the person who has just died is a man or a woman, a member of a religious order, etc. Moreover, every Sunday the inhabitants of these villages follow the principal passages of the divine service with the aid of the pealing of the different bells. We have seen old people, prevented from attending the service because of their infirmities, with prayer book in hand, follow at a distance the priest's various movements.

Our story would be incomplete if we did not mention the African tom-tom, which some people consider a sort of acoustical telegraphy. The African explorer, Dr. A. R. Lindt, has written a short report on the tom-tom. We quote the following from his work: "There is no key to the acoustical telegraphy of the Africans. Since they have no written language, they are unable to divide their words into letters. The tom-tom therefore does not translate letter by letter or even word by word, but translates a series of well-defined thoughts into signals. There are different signals for all acts

interesting to the tribe: mobilization, death of the chief, and summons to a judicial convocation. However, the tom-tom also serves to transmit an order to a definite person. Thus, when a young man enters the warrior class, he receives a kind of call signal which introduces him and enables him to be recognized at a distance.

As yet, explorers have not been able to discover how intelligible the same signals

are to different tribes. It is certain, however, that friendly tribes use the same signals. A settlement receiving a signal transmits it to the next village, so that in a few minutes a communication can be sent several hundred kilometers.

Acoustical telegraphy is still used today by certain enterprises such as railroads, boats, automobiles, fire fighting services and alarm services.

This completes our quotations from the Dudley archives. We see that the concept of long-distance communication has a long history and that there is some evidence that speech communication at a distance was practiced by the ancients.

3.1.3 The Telephone

Proceeding more or less chronologically, we come to that most important development, the invention of the telephone by Alexander Graham Bell. There is no need to chronicle the well-known events leading to this invention and the enormous consequent effect on human communication; we restrict ourselves to several comments. It is interesting that Bell's primary profession was that of a speech scientist who had a keen understanding of how the human vocal apparatus worked, and, in fact, Flanagan [5] describes Bell's "harp telephone," which showed that Bell understood the rudiments of the speech spectral envelope. Nevertheless, telephone technology has been mostly concerned with transmission methods. Recently, however, with the growing use of cellular phones in which transmission rate is limited by nature, efficient methods of speech coding have become an increasingly important component of speech research at many laboratories.

3.1.4 The Channel Vocoder and Bandwidth Compression

In a *National Geographic* magazine article [2], Colton gives an engrossing account of the history of telephone transmission. Figure 3.1, taken from that article, shows the telephone wires on lower Broadway in New York City in the year 1887. It is clear that progress in telephony could easily have been brought to a halt if not for improvements, such as underground cables, multiplexing techniques, and fiber-optical transmission. Dudley pondered this traffic problem in a different way, that is, through coding to reduce the intrinsic bandwidth of the source, rather than increasing the capacity of the transmission medium.

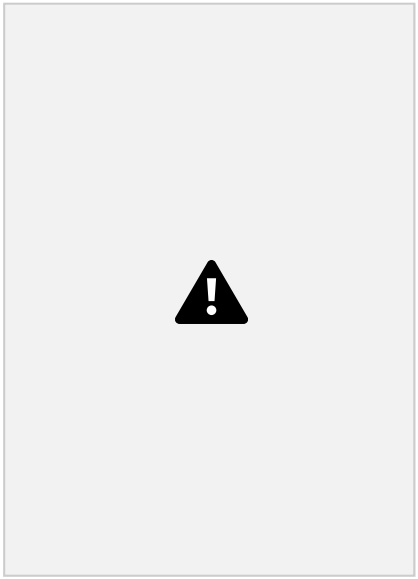


FIGURE 3.1 Lower Broadway in 1887.

could lead to the advantages of "secrecy, and more telephone channels in the same fre quency space." Both of these predictions were correct, but the precise ways in which they came to pass (or are coming to pass) probably differ somewhat from how Dudley imagined them. In 1929, there was no digital communications. When digitization became feasible, it was realized that the least-vulnerable method of secrecy was by means of digitization. However, digitization also meant the need for wider-transmission bandwidths. For exam ple, a 3-kHz path from a local telephone cannot transmit a pulse-coded modulation (PCM) speech signal coded to 64 kbits/s (the present telephone standard). The channel vocoder was thus quickly recognized as a means of reducing the speech bit rate to some number that could be handled through the average telephone channel, and this led eventually to a standard rate of 2.4 kbits/s.

With respect to the second prediction, given that the science of bandwidth compres sion is now approximately 50 years old, one might assume that "more telephone channels in the same frequency space" would by now be a completely realized concept within the public telephone system. Such, however, is not the case. Although it is our opinion that Dudley's second prediction will eventually come true, it is fair to ask why it is taking so long. With the recent boom in wireless telephony, the bandwidth is now an issue of even greater importance.

We conclude this section with a reference to an informative and entertaining paper by Bennett [1]. This paper is a historical survey of the X-System of secret telephony that was used during World War II. Now totally declassified, the X-System turns out to be a quite sophisticated version of Dudley's channel vocoder! It included such features as

PCM transmission, logarithmic encoding of the channel signal, and, of course, enciphered speech. Bennett has many interesting anecdotes concerning the use of the X-System during the war.

3.2 VOICE-CODING CONCEPTS

To understand why a device such as a vocoder reduces the information content of speech, we need to know enough about human speech production to be able to model it approximately. Then we must convince ourselves that the parameters of the model vary sufficiently slowly to permit efficient transmission. Finally, we must be able to separate the parameters so that each one is coded optimally. The implementation of these concepts is captured by the phrase "analysis-synthesis system." The analysis establishes the parameters of the model; these parameters are transmitted to the receiving end of the system and used to control a synthesizer with the goal of reproducing the original utterance as faithfully as possible.

A convenient way to understand vocoders is to begin with the synthesizer. A concise statement that helps define a model of speech is given by Fant [4]: "The speech wave is the response of the vocal tract to one or more excitation signals." This concept leads directly to engineering methods to separate the *source* (the excitation signal) from the *filter* (the time-varying vocal tract). The procedures (and there are many) for implementing this separation can be called deconvolution, thus implying that the speech wave is a linear convolution of source and filter.[1] In spectral terms, this means that the speech spectrum can be treated as the *product* of an excitation spectrum and a vocal tract spectrum. Figure 3.2 is a simplified illustration of the spectral cross section for sustained vowels. Numerous experiments have shown that such waveforms are quite periodic; this is represented in the figures by the lines. In (a) the lines are farther apart, representing a higher pitched sound; in (b) and (c) the fundamental frequency is lower.

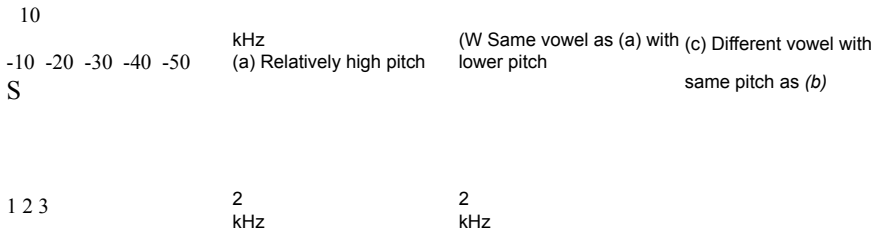


FIGURE 3.2 Fine structure and spectral envelope of sustained vowels

The spectral envelope determines the relative magnitudes of the different harmonics, and it, in turn, is determined from the specific shape of the vocal tract during the phonation of that vowel. Deconvolution is the process of physically separating the spectral envelope from the spectral fine structure, and in later chapters we describe methods of implementing such a process. Once this separation is accomplished, we can hypothesize, with some confidence, that both the spectral envelope and spectral fine structure can be efficiently parameterized, with consequent bandwidth savings during transmission.

The parameters, if appropriately obtained, must vary relatively slowly because ulti mately they depend on the articulator motions of the speech-producing mechanisms. Since these are human motions they obey the mechanical constraints imposed by the flesh-and blood properties of the pertinent human organs, which move relatively slowly compared to typical speech bandwidths of 5 kHz.

The human vocal tract has been represented as a time-variable filter excited by one or more sources. The mechanism for this production varies according to the type of speech sound. Air pressure is supplied by the lungs. For vowel production, the cyclic opening and closing of the glottis creates a sequence of pressure pulses that excite resonant modes of the vocal tract and nasal tract: the energy created is radiated from the mouth and nose to the listener.

For voiceless fricatives (e.g., s, sh, f, and th), the vocal cords are kept open and the air stream is forced through a narrow orifice in the vocal tract to produce a turbulent, noiselike excitation. For example, the constriction for "th" is between tongue and teeth; for "f" it is between lips and teeth.

For voiceless plosives (e.g., p, t, and k), there is a cross section of complete closure in the vocal tract, causing a pressure buildup. The sudden release creates a transient burst followed by a lengthier period of aspiration.

A more extensive categorization of speech sounds is given in Chapter 23, including some additional material about the articulator positions (tongue, lips, jaw, etc.) correspond ing to these categories.

Several basic methods of source-filter separation and subsequent parameterization of each have been developed over the past half-century or so. We limit our discussion to four such methods: (a) the channel vocoder, (b) linear prediction, (c) cepstral analysis, and (d) formant vocoding. Details of these methods will be examined in later chapters; for now we discuss the general problem of source-filter separation and the coding of the parameters.

One way to obtain an approximation of the spectral envelope is by means of a care fully chosen bank of bandpass filters. Looking at Fig. 3.2, we see that the complete spectrum envelope is not available; only the *samples* of this envelope at frequencies determined by the vertical lines are available. We assume that the fundamental frequency is not known so that we have no *a priori* knowledge of the sample positions. However, by passing the signal through a filter bank, where each filter straddles several harmonics, one can obtain a reasonable approximation to the spectral envelope. If the filter bandwidths are wide enough to encompass several harmonics, the resulting intensity measurements from all filters will *not* change appreciably as the fundamental frequency varies, as long as the envelope re-

mains constant. This is the method employed for spectral analysis in Dudley's channel vocoder. The array of (slowly varying) intensities from the filter bank can now be coded and transmitted.

Linear prediction is a totally different way to approximate the spectral envelope. We

hypothesize that a reasonable estimate of the nth sample of a sequence of speech samples is given by

$$*(") = Y,a_k s(n-k). \quad (3.1)$$

In Eq. 3.1, the $a^{\wedge}{}'s$ must be computed so that the error signal

$$e(n) = s(n) — s(n) \quad (3.2)$$

is as small as possible. As we will show in the Chapter 21, Eq. 3.1 and the minimizing computational structure used lead to an all-pole digital synthesizer network with a spectrum that is a good approximation to the spectral envelope of speech.

Source-filter separation can also be implemented by cepstral analysis, as illustrated in Figure 3.3. Figure 3.3(a) shows a section of a speech signal, Fig. 3.3(b) shows the spec tram of that section, and Fig. 3.3(c) shows the logarithm of the spectrum. The logarithm transforms the multiplicative relation between the envelope and fine structure into an ad ditive relation. By performing a Fourier transform on Fig. 3.3(c), one separates the slowly varying log spectral envelope from the more rapidly varying (in the frequency domain) spectral fine structure, as shown in Fig. 3.3(d). Source and filter may now be separately coded and transmitted.

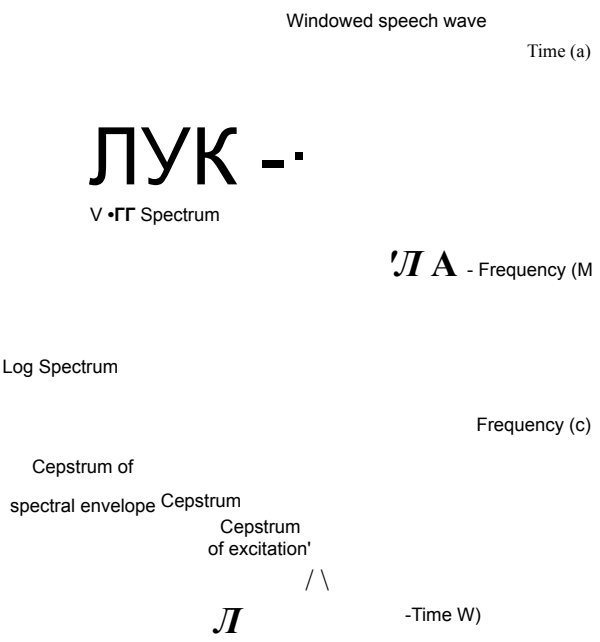


FIGURE 3.3 Illustration of source-filter separation by cepstral analysis.

(abscissa), frequency (ordinate), and intensity (darkness). Much can be said about the interpretation of spectrograms; here we restrict our discussion to the highly visible resonances or *formants* and to the difference between Fig. 3.4 (wideband spectrogram) and Fig. 3.5 (narrow-band spectrogram).

We see from Fig. 3.4 that during the vowel sounds, most of the energy is concentrated in three or four formants. Thus, for vowel sounds, an analysis could entail tracking of the frequency regions of these formants as they change with time. Many devices have been invented to perform this operation and also to parameterize the speech for other sounds, such as fricatives (s, th, sh, f) or plosives (p, k, t); again we defer detailed descriptions for later.

Formant tracks are also visible in Fig. 3.5, but there is a significant difference between

Time (sec) **FIGURE 3.4 Wideband spectrogram.**

Finally, formant analysis can be used for source-filter separation. In Chapters 10 and 11 (Wave Basics and Speech Production), the theory of vocal-tract resonance modes is developed. However, we can to some extent anticipate the result by studying the speech spectrograms of Figs. 3.4 and 3.5. These figures are three-dimensional representations of time

7 -
6 -
_ 5 - IM
x
£ 4 - >,
S 3 - £ 2 -

W E P L E DGETYÖ S 0 ME H EAV У TTREÄSUR E **1 2**

Time (sec)

FIGURE 3.5 Narrow-band spectrogram.

the two figures. Whereas Fig. 3.4 displays the periodicity of the signal during vowels as vertical striations, Fig. 3.5 displays the periodicity horizontally. An explanation of this difference is left as an exercise.

3.3 HOMER DUDLEY (1898-1981)

Homer Dudley's inventions of the channel vocoder and Voder triggered a scientific and engineering effort that is still in progress. On my first visit to the Bell Laboratories in 1961[2] I was hosted by Dr. Ed David, who then managed the speech-processing group. As we passed an office, David whispered to me, "that's Homer Dudley." I was not intro duced to Mr. Dudley and on subsequent visits did not see him. At that time he was near retirement age and, I suppose, not in the mainstream of Bell Laboratories' work. Quite a few years later (the late 1960s), Lincoln Laboratory was privileged to have the then retired inventor as a consultant. We mention several items of interest from his brief stay there.

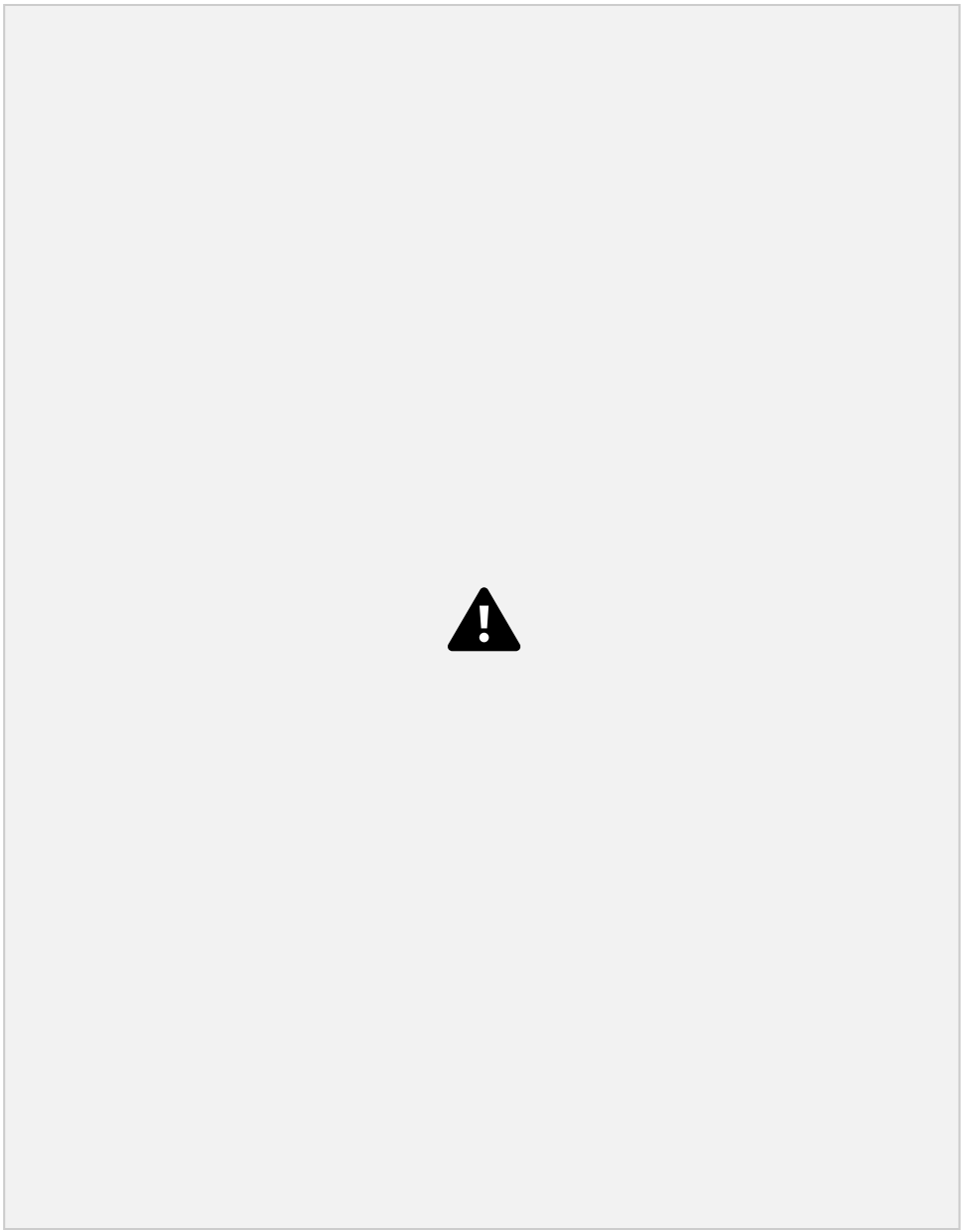
Dudley had a strong feeling that we should study speech waveforms as much, and perhaps more, than speech spectrograms. He felt that with practice, one could learn to read these waveforms. Dudley's speculation remains unproven. However, in an effort to augment his claim, Dudley, with the help of Everett Aho, produced photographs that are very informative and aesthetically pleasing. They are reproduced here as Figs. 3.6-3.11. Observing these waveforms, one develops a good feeling for the relative duration and amplitude of the vowels versus the consonants. In addition, we see the precise timing of the burst and voice-onset time of the voiced plosive sounds, b, d, and g. An inspection of the vowel sound I as in "thin" or "fish" illustrates the high frequency of the second resonance and the low frequency of the first resonance. We also note that the energy of the sh sound in "fish" is much stronger than the "f" sound in fish. Many other relationships among the acoustic properties of the phonemes can be found by careful observation of good-quality speech waveforms.

In 1967, a vocoder conference was organized under the auspices of the U.S. Air Force Cambridge Research Laboratory (AFCRL). Dudley was honored at this conference. Figure 3.12 shows Dudley displaying the plaque to the audience.

In 1969, when Dudley discontinued his consultancy at Lincoln Laboratory, he en trusted one of the authors (Gold) with two boxes filled with various technical information, plus a large number of newspaper clippings on his inventions. These have been used freely in this chapter, and they have been donated to the archives of the Massachusetts Institute of Technology.

In 1981, we received the news of his death at age 83. Tributes to him were written by Manfred Schroeder and James L. Flanagan, who worked with Dudley at Bell

orn.



Time centiseconds FIGURE 3.6 Dudley's waveform display.

Time-centiseconds

FIGURE 3.7 Continuation of Dudley's waveform display.

SPEECH

ANALYSIS AND

SYNTHESIS

OVERVIEW

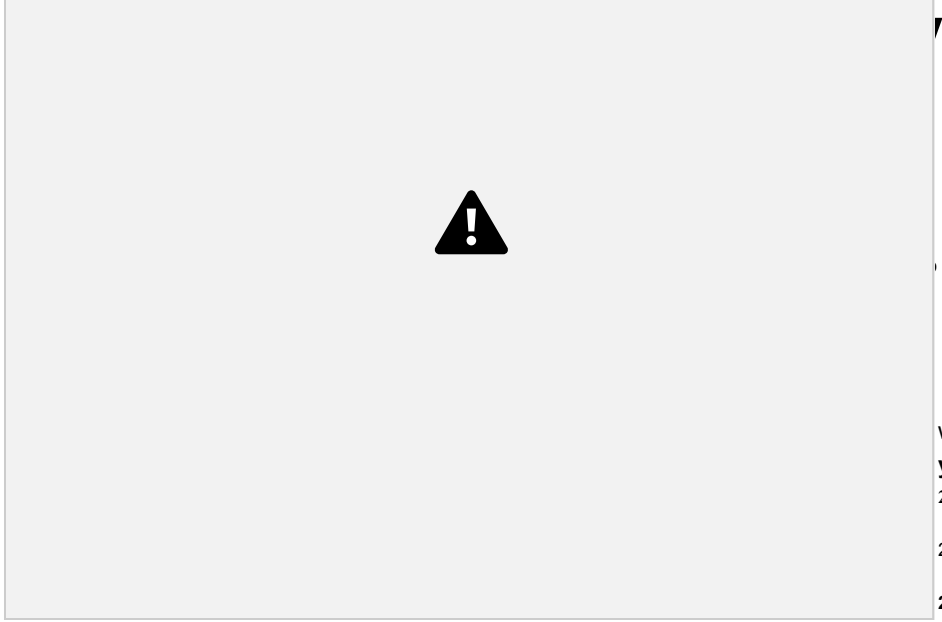
Time-centiseconds

200 Щ » *чтвЩшШш я* I
Т ™ 201 »*—^|.«i»«|>,l|iWJ

-\ г __г 202
20:3 204 205 Ц

205 210

«■—* wyiWfA *тШ*Ф*** 4- 210



waveform display.

FIGURE 3.8 Conclusion of Dudley's

360 365 370

^^\|\/*л лл*-^ /У\л^^^---- ^ 365

FIGURE 3.10 Continuation of Dudley's second waveform display.

505 510

525

520

525

515

530

520

530 535 V44/4.

535 540 **«-Jy í^Λ^^ΛΛ^^ ^Vv^^/v^^^^**

540 5^5 0JW\. Vw^JVty V W^»Av

545 5ʳÜ -s^v^vJt ^*ifimfKu**.**^ ^./шX/чи"*«—_ *

550 555 − + 555 560 560 565

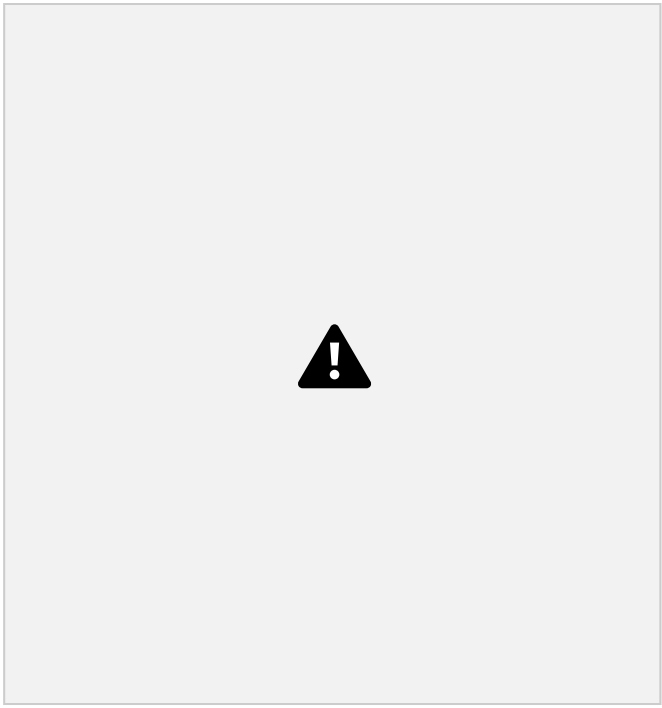


FIGURE 3.12 Dudley receiving an award.

3.4 EXERCISES

3.1 Explain why wideband spectrograms show periodicity in time whereas narrow-band spectrograms show periodicity in frequency.

3.2 Invent a display that shows periodicity in both time and frequency.

3.3 Can you think of a reason why spectrograms are preferable visual displays to direct oscillographic waveforms?

3.4 Which sounds are more likely to be better understood from waveforms? From spectrograms?

3.5 Construct a table for the phonemes of the phrase "we pledge you some heavy treasure." The leftmost column should list the phonemes alternating with the transition regions; the next column should list your best estimate of the beginning; and the third column should list the end of the speech section. Base your estimates on Figs. 3.4 and 3.5.

3.6 Construct a syllable table in the same manner as in the previous exercise.

3.7 During World War II, Roosevelt and Churchill conversed by telephone between London and Washington, using a channel vocoder. Explain why the vocoder was an important component of the communications link.

3.8 The phrase "carrier nature of speech" was proposed by Dudley as a way of explaining how a vocoder could represent speech with fewer bits (or less bandwidth). Explain how channel vocoders, linear predictive vocoders, and cepstral vocoders implement this concept and, as a result, represent the speech signal more efficiently than a standard telephone or PCM system.

3.5 APPENDIX: HEARING OF THE FALL OF TROY

LEADER OF CHORUS:
I come to do you reverence, Clytemnestra.
For it is right to give the king's wife honor,
A woman on a throne a man left empty
But if you know of good or only hope
to hear of good and so do sacrifice,
I pray you speak. Yet if you will, keep silence.
CLYTEMNESTRA:
With glad good tidings, so the proverb runs,
may dawn arise from the kind mother night.
For you shall learn a joy beyond all hope:
the Trojan town has fallen to the Greeks.
LEADER:
You say? I cannot hear-I cannot trust—
CLYTEMNESTRA:
I say the Greeks hold Troy. Do I speak clear?
LEADER:
Joy that is close to tears steals over me.
CLYTEMNESTRA:
Quite right. Such tears give proof of loyalty.
LEADER:
What warrant for mese words? Some surety have you?
CLYTEMNESTRA:
I have. How not-unless the gods play tricks.
LEADER:
A fair persuasive dream has won your credence?
CLYTEMNESTRA:
I am not one to trust a mind asleep.
LEADER:
A wingless rumor then has fed your fancy?
CLYTEMNESTRA:
Am I some little child that you would mock at?

LEADER:
But when, *when,* tell us, was the city
sacked? CLYTEMNESTRA:
This night, I say, that now gives birth

to dawn. LEADER:
And what the messenger that came so swift?
CLYTEMNESTRA:
A god! The fire-god flashing from Mount Ida. Beacon sped beacon on, couriers of flame.
First, Ida signaled to the island peak of Lemnos, Hermes' rock, and swift from there Athos, God's mountain, fired the great torch.
It leaped, it skimmed the sea, a might of moving light. joy-bringing, golden shining, like a sun,
and sent the fiery message to Macistus.
Whose towers, then, in haste, not heedlessly
or like some drowsy watchman caught by sleep, sped on the herald's task and flashed the beacon afar, beyond the waters of Euripus
to sentinels high on Messapius' hillside,
who fired in turn and sent the tidings onward, touching with flame a heap of withered heather. So, never dimmed but gathering strength, the splendor over the levels of Asopus sprang,
lighting Cithaeron like the shining moon,
rousing a relay there of travelling flame.
Brighter beyond their orders given, the guards kindled a blaze and flung afar the light.
It shot across the mere of Gorgopis.
It shone on Aegiplanctus' mountain height,
swift speeding on the ordinance of fire,
where watchers, heaping high the tinder wood, sent darting onward a great beard of flame
that passed the steeps of the Saronic Gulf
and blazing leaped aloft to Arachnaeus,
the point of lookout neighbor to our town.
Whence it was flashed here to the palace roof, a fire fathered by the flame on Ida.
Thus did the they hand the torch on,
one to other, in swift succession finishing the course.
And he who ran both first and last is victor.
Such is my warrant and my proof to you:
my lord himself has sent me word from Troy.

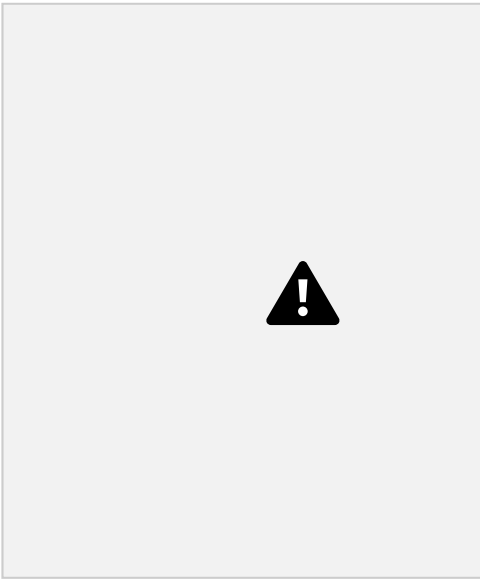


FIGURE 3.13 Map, showing the communications path described in *Agamemnon.*

BIBLIOGRAPHY

1. Bennett, W. R., "Secret telephony as a historical example of spread-spectrum communication," *IEEE Trans. Commun.* **COM-31:** 98-104, 1983.
2. Colton, F. B., "The miracle of talking by telephone," *National Geographic* 70(4): 395^133, 1937.
3. Dudley, H., "The vocoder," *Bell Labs Record* **17:** 122-126, 1939.

, *Acoustic Theory of Speech Production,* Morton, S-Gravenhage, 1960.

. J. L., *Speech Analysis Synthesis and Perception,* 2nd ed., Springer-Verlag, New lin, 1972.

CHAPTER 3

BRIEF HISTORY

OF AUTOMATIC SPEECH RECOGNITION

^CONCEPTUALL Y , the development of speech recognition is closely tied with other developments in speech science and engineering, and as such can be viewed as having roots in studies going back to the Greeks (as with synthesis). However, the history of speech recognition[1] per se in the 20th Century began with the invention of a small toy, Radio Rex.

4.1 RADIO REX

The first machine to recognize speech to any significant degree may have been a commercial toy named Radio Rex, which was manufactured in the 1920s. Here is a description from a 1962 review paper [18]:

> *It consisted of a celluloid dog with an iron base held within its house by an electromagnet against the force of a spring. Current energizing the magnet flowed through a metal bar which was arranged to form a bridge with 2 supporting members. This bridge was sensitive to 500 cps acoustic energy which vibrated it, interrupting the current and releasing the dog. The energy around 500 cps contained in the vowel of the word Rex was sufficient to trigger the device when the dog's name was called.*

It is likely that the toy responded to many words other than "Rex," or even to many nonspeech sounds that had sufficient 500-Hz energy. However, this inability to reject out-of vocabulary sounds is a weakness shared by most recognizers that followed it. Furthermore, the toy was in some sense useful, since it fulfilled a practical purpose (amusing a child or playful adult), which was not often accomplished by many of the laboratory systems that followed. Although quite simple, it embodied a fundamental principle of speech recognizers for many years: store some representation of a distinguishing characteristic of the desired sound and implement a mechanism to match this characteristic to incoming speech.

'As with any such brief historical review, we have been limited to discussing a small fraction of the many contributions and contributors to this extremely active field.

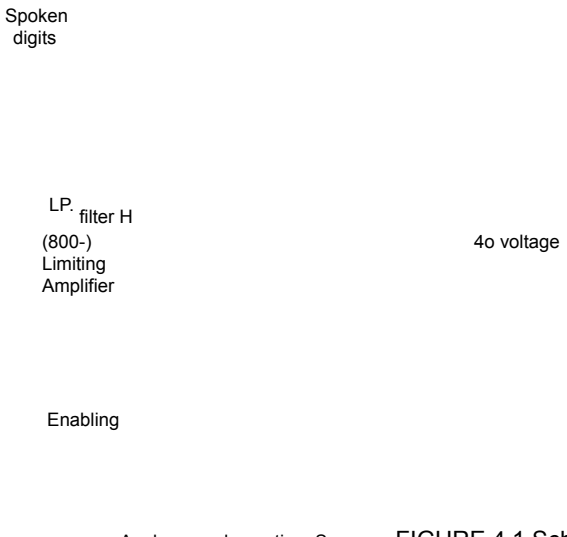


FIGURE 4.1 Schematic for 1952 Bell Labs digit recognizer [19].

Radio Rex was later referred to in a famous letter to the Acoustical Society by John Pierce of Bell Labs [57], in which he strongly criticized the speech recognition research of that time (1969):

> *What about the possibility of directing a machine by spoken instructions? In any practical way, this art seems to have gone downhill ever since the limited commercial success of Radio Rex.*

4.2 DIGIT RECOGNITION

A system built at Bell Labs and described in [19] may have been the first true word recognizer, as it could be trained to recognize digits from a single speaker. It measured a simple function of the spectral energy over time in two wide bands, roughly approximating the first two resonances of the vocal tract (i.e., formants). Although the system's analysis was crude, its estimate of a word-long spectral quantity may well have been more robust to speech variability than some of the later common approaches to estimating the time-varying speech spectrum. It tracked a rough estimate of formant positions instead of the spectrum itself. This is potentially resistant to irrelevant modifications of the overall speech spectrum. For instance, a simple turn of the talker's head away from a direct path to the listener often produces marked changes in the spectrum of the received speech (in particular, a relative reduction in the amplitude of the higher spectral components). The Bell Labs system's spectral estimation technique was, however, quite crude, histogramming low- and high frequency spectral moments over an entire utterance, and thus timing information was lost. Although the idea was good, there was insufficient technology to develop it very far by modern standards; it used analog electrical components and must have been difficult to modify. Still, the inventors claimed that it worked very well, achieving a 2% error for a single speaker uttering digits that were isolated by pauses [19].

The system (see Fig. 4.1) worked generally as follows: incoming speech was filtered into low- and high-frequency components and each component strongly saturated so that its amplitude was roughly independent of signal strength. The cutoff frequency in each case was roughly 900 Hz, which is a reasonable boundary between first and second formants for adult males.[2] Zero crossings were counted for each of the two bands, and the system used this value to estimate a central frequency for each band. The low-frequency number was quantized to one of six 100-Hz subbands (between 200 and 800 Hz), and the high frequency number was quantized to be one of five 500-Hz subbands, beginning at 500 Hz. Together, these two quantized values correspond to one of 30 possible frequency pairs (in practice, only 28 were used, as the other two were rarely applicable). During a training period, capacitors were used to store charges associated with the time that the signal was mapped to a particular pair of frequencies. This distribution was learned for each digit. The resulting distributions were then used to choose conductances for RC circuits that would be used during recognition. When a new digit was uttered, a new distribution was determined in a similar way and compared to the stored distributions by switching between RC circuits corresponding to all possible digits (where the conductance corresponded to the template, the capacitances and charge time were all equal, and where the charging voltage for each frequency pair was determined by the new utterance). This procedure essentially implemented correlations between each stored distribution and the new distribution. The digits had distinguishable frequency-pair distributions and so could usually be discriminated from one another (See [19], Fig. 2, p. 639).

[2]Children and adult women often have first formants above this frequency, and speakers of either gender can have second formants that are below 900 Hz for some sounds. Still, 900 Hz is a reasonable dividing point between major energy components in speech.

Note that even in 1952, researchers were reporting a speech recognizer that was 98% accurate! An examination of modern press releases suggests that this figure may be a constant for speech-recognition systems (those that are reported, anyway).

4.3 SPEECH RECOGNITION IN THE 1950s

In 1958, Dudley made a classifier that continuously evaluated spectra, rather than approx imations to formants. This new paradigm was commonly used afterward; in fact, broadly speaking, the current dominant paradigm for speech recognition uses some function of a local spectral estimate varying over time as the representation of the incoming speech. In 1959, Denes, from the College of London, added grammar probabilities in addition to acoustic information. In other words, he pointed out that the probability of a particular linguistic unit being uttered can also be dependent on the previous linguistic unit, so that the probability of a word need not be solely dependent on the acoustic input.

In 1962, David and Selfridge put together Table 4.1, which compared a number of speech-recognition experiments in the preceding decade [18] including the two recognizers mentioned above. In general, researchers performed spectral tracking, detected a few words and sounds, and performed tests on a small number of people.

4.4 THE 1960S

Throughout much of the 1960s, automatic speech-recognition research continued along similar lines. Martin deployed neural networks for phoneme recognition in 1964. Digit recognizers became better in the 1960s, achieving good accuracy for multiple speakers. Widrow trained neural networks to recognize digits in 1963 [81]. Phonetic features were also explored by a number of researchers. However, as noted earlier, in 1969 John Pierce wrote a caustic letter entitled "Whither Speech Recognition?" In it he argued that scientists were wasting time with simple signal-processing experiments because people did not do speech recognition, but rather speech understanding. He also pointed out the lack of scien tific rigor in the experimentation at that time and he suggested that arbitrary manipulation of recognizer parameters to find the best performance was like the work of a "mad scientist," rather than that of a serious researcher. At the time, Pierce headed the Communications Sciences Division at Bell Labs, and his remarks were quite influential.

Although there may have been much that was correct about Pierce's criticism, there were a number of major breakthroughs in the 1960s that became important for speech recognition research in the 1970s. First, as noted previously, prior to this period the pri mary approach to estimating the short-term spectrum was a filter bank. In the 1960s, three spectralestimation techniques were developed that were later of great significance for recognition, although their early applications to speech were for vocoding: the Fast Fourier transform (FFT), cepstral (or homomorphic) analysis, and linear predictive coding (LPC). Additionally, new methods for the pattern matching of sequences were developed: a deterministic approach called dynamic time warp (DTW), and a statistical one called the hidden Markov model (HMM).

TABLE 4.1 Pre-1962 Systems' Speech

Speech-Recognition Investigators			No. of talkers	Appox. error rate	Additional facts
Kersta	Davis, Biddulph, and Balashek				
		Fry and Denes			
				Belar	Denes
				Dudley and Balashek	Hughes
				Mathews and	
			Olson and		

Shultz	*Af-At* matrix		3
Petrick and Willett	Formants 1 and 2		<1.0
Forgie and Forgie	**Vocabulary** 10		7
Keith-Smith and Klem	digits	1 talker	6
		2 men	
	10 digits	6 men	
Sebestyen Suzuki and Nakata			<1 Ä;20
	14 speech sounds in 139 words	4 men, 3 women	
From [18]. **representation**		25 men, 25 women 1 talker	
Selected entries from Δ/-Δí matrix (200 cps x 67 ms)	10 words or syllables	11 men, 10 women 11 men, 10 women	
Formants 1 and 2 as a function of time	10 digits		
Selected entries from *Af-At* matrix	11 sound categories in 100 words	10 speakers 5 speakers **(%)**	
	10 digits	0.2	
	10 digits		
Af-At matrix	10 vowels 10 vowels	2.0	
Af-At matrix		28.0 (sounds) 56.0 (words)	
	10 digits		
Spectral features	5 vowels in consonant contexts **tested**	2.0	
Spectral features	9 men, 5 women	5	
Af-At matrix		6	
Spectral features	1 talker		
Δ/-Δ? matrix		30	
	1 talker		

and comments	Correlation
Spectrograms	
quantized into 2 levels	metric

Phoneme

vowels in bisyllable words and short sentences yield higher error rates

digram frequencies used to supplement primitive recognition from acoustics

sequence disregarded

Temporal

Spectral pattern time and amplitude normalized

Feature-selection based linguistic analysis

Spectral patterns time normalized

Statistical decision procedure

used to select relevant spectral features

Additional experiments on

4.4.1 Short-Term Spectral Analysis

As discussed in Chapter 7, Cooley and Tukey introduced the FFT [17]. This is a com putationally efficient form of the discrete Fourier transform (DFT), which in turn can be interpreted as a filter bank. However, its efficiency was important for speech-recognition research, as it was for many other disciplines.

An alternative to filter banks and their equivalent FFT implementation was cepstral processing, which was originally developed by Bogert for seismic analysis [10] and applied later to speech and audio signals by Oppenheim, Schafer, and Stockham [53]. Cepstral pro cessing will be discussed later (primarily in Chapter 20), but its significance for speech recognition is primarily as an approach to estimating a smooth spectral envelope. It ulti mately became widely used for recognition, particularly in combination with other analysis techniques (see Chapter 22).

LPC is a mathematical approach to speech modeling that has a strong relation to the

acoustic tube model for the vocal tract. Fundamentally, it refers to the use of an autoregressive (pole only) model to represent the generation of speech; each time point in sampled speech is predicted by a weighted linear sum of a fixed number of previous samples. In Chapter 21 we will provide a more rigorous definition, but for now the significance of LPC is that it provides an efficient way of finding a short-term spectral envelope estimate that has many desirable properties for the representation of speech, in particular the emphasis on the peak spectral values that characterize voiced sounds. Some of the early writings on this topic include [32], [2], and [44]. An excellent tutorial on the topic was written by Makhoul [42].

4.4.2 Pattern Matching

Dynamic programming is a sequential optimization scheme that has been applied to many problems [9]. In the case of speech analysis for recognition, it was proposed as a method of time normalization - different utterances of the same word or sentence will have differing durations for the sounds, and this will lead to a potential mismatch with the stored represen
tations that are developed from training materials. DTW applies dynamic programming to this problem. It was proposed by Sakoe around 1970 (but published in an English-language journal in 1978 [66]). Vintsyuk was among the first to develop the theory, and he also applied it to continuous speech [75]. DTW for connected word recognition was described by Bridle [13] and Ney [51]. Excellent review articles on the subject were written by White [80] and by Rabiner and Levinson [62].

DTW is a deterministic approach to the matching of the time sequence of short term spectral estimates to stored patterns that are representative of the words that are being modeled [50]. Alternatively, one could imagine a statistical approach, in which the incoming time sequence is used to assess the likelihood of probabilistic models rather than speech examples or prototypes. The mathematic foundations for such an approach were developed in the 1960s, and they were built on the statistical characterization of the noisy commu nications channel as described in 1948 by Shannon [69]. Most notably, the work of Baum and colleagues at the Institute for Defense Analysis established many of

Towards the end of the 1960s, a number of researchers became interested in devel oping these ideas further for the case of a naturally occurring sequence, and in particular for speech recognition. Many of these ultimately joined a research group at IBM, which pioneered many aspects of HMM-based speech recognition in the 1970s. An early IBM re port that influenced this work was [74], and a range of other publications followed through the early to mid-1970s, for example, [7], [3], [35], and [34]. The group developed an early HMM-based automatic speech-recognition system that was used for a continuous speech recognition task referred to as New Raleigh Language. Baker independently developed an HMM-based system called Dragon while still a graduate student at Carnegie Mellon University (CMU) [4]. Many other researchers were working with this class of approaches by the mid-1980s (e.g., [67]).

4.5 1971-1976 ARPA PROJECT

As noted earlier, one of Pierce's criticisms of earlier efforts was that there was insufficient attention given to the study of speech *understanding,* as opposed to recognition. In the 1970s the Advanced Research Projects Agency (ARPA)[3] funded a large speech-understanding project. The main work was done at three sites: System Development Corporation, CMU, and Bolt, Beranek & Newman (BBN). Other work was done at Lincoln, SRI International, and University of California at Berkeley. The goal was to perform 1000-word automatic speech recognition by using a few speakers, connected speech, and constrained grammar with less than a 10% semantic error. The funding was reported to be $15 million. According to Klatt, who wrote an interesting critique of this program [36], only a system called Harpy, built by a CMU graduate student (Bruce Lowerre), fulfilled the goals. He used LPC segments, incorporated high-level knowledge, and modified techniques from Baker's Dragon system, as well as from another CMU system, Hearsay.

4.6 ACHIEVED BY 1976

By 1976, researchers were using spectral feature vectors, LPC, and phonetic features in their recognizers. They were incorporating syntax and semantic information. Approaches incorporating neural networks, DTW, and HMMs were developed. A number of systems were built. Efforts on reducing search cost were explored. Techniques from artificial intelli
gence were often used, particularly for the ARPA program. HMM theory had been applied to automatic speech recognition, and HMM-based systems had been built. In short, many of the fundamentals were in place for the systems that followed.

[3]This U.S. government agency was originally known as ARPA but later became known as DARPA (the D standing for Defense), but after a few years it reverted back to ARPA; as of this writing it is DARPA again.

4.7 THE 1980S IN AUTOMATIC SPEECH RECOGNITION

In the 1980s, most efforts were concentrated in scaling existing techniques (e.g., LPC and

HMMs) to more difficult problems. New front-end processing techniques were also developed in this time period. For the most part, however, the structure of speech-recognition systems did not change; they were trained on a larger quantity of data and extended to more difficult tasks. This extension did require extensive engineering developments, which were made possible by a concerted effort in the community. In particular, there was a major effort to develop standard research corpora.

4.7.1 Large Corpora Collection

Prior to 1986 or so, the speech-recognition community did not have any widely accepted common databases for training recognition systems. This made comparisons between labs difficult, since few researchers trained or tested on the same acoustic data. Many speech researchers were concerned with this problem. Industrial scientists (e.g., those with Texas Instruments and Dragon Systems) worked with NIST (National Institute of Standards and Technology)[4] and compiled large standard corpora.

In 1986, collection began on the TIMIT[5] corpus [52], which was to become the first widely used standard corpus. A 61-phone alphabet was chosen to represent phonetic distinctions. The sentences in TIMIT were chosen to be phonetically balanced, meaning that a good representation of each phone was available within the training set. There were 630 speakers that each said 10 sentences, including two that were the same for each speaker. The data were recorded at Texas Instruments and phonetically segmented at MIT, first by use of an automatic segmenter [40], followed by manual inspection and repair of the alignments by graduate students. This resulted in a database in which the time boundaries of the phone in the speech signal are marked for every phone uttered by a speaker. Even though errors still undoubtedly exist in the TIMIT database, it remains one of the largest and most widely used hand-labeled phonetic corpora.

With the advent of the second major ARPA speech program in the mid-1980s, a new task called Resource Management (RM) was defined, with a new database [60] of speech. RM had much in common with the task from the first ARPA program in the 1970s. The major differences were that the grammar had a greater perplexity,[6] and the recordings were made of read speech. Sentences were constructed from a 1000-word language model, so that no out-of-vocabulary words were encountered during testing. The corpus contained 21,000 utterances from 160 speakers. One important characteristic of the RM task was that it included speaker-independent recognition; that is, some systems were trained on many speakers, and they were tested on speakers not in the training set.

Later on in the program, the focus shifted to the Wall Street Journal Task - rec

ognizing read speech from the Wall Street Journal.[7] The first test was constrained to be a 5000-word vocabulary test with no out-of-vocabulary words; later, a 20,000-word task with out-of-vocabulary words was developed. More recent tests used an essentially unlimited vocabulary, and researchers often used 60,000-word decoders for system evaluations.

Another task that was developed in parallel with the read speech program was Air Travel Information System (ATIS), which was based on spontaneous query in the airline-reservation domain. ATIS is a speech-understanding task (as opposed to a speech recognition task). Systems not only had to produce word strings, but they also had to attempt to derive some semantic meaning from these word strings and perform an appropriate function. For instance, if the user said "show me the flights from Boston to San Francisco," the system should respond by showing a list of flights. Interaction continued with the system in order to reach some goal; in this case, ordering airline tickets. This domain was more practical than the Wall Street Journal task, but the vocabulary size was smaller. Systems today are now quite good at this task.

DARPA funded the collections of these corpora, and the collection processes were managed by NIST. NIST subcontracted much of the collection work to sites such as SRI and Texas Instruments. These and other corpora are now distributed through the Linguistic Data Consortium (LDC), which is based at the University of Pennsylvania in Philadelphia.

4.7.2 Front Ends

A number of new front ends, that is, subsystems that extract features from the speech signal, were developed in the 1980s. Of particular note are mei cepstrum [20], perceptual linear prediction [29], delta cepstral coefficients [22], and other work in auditory-inspired signal-processing techniques, for example, [68] and [27]. (See Chapter 22 for a discussion of many of these approaches.)

4.7.3 Hidden Markov Models

As noted previously, the fundamentals of HMM methodology were developed in the late 1960s, with applications to speech recognition in the 1970s. In the 1980s, interest in these approaches spread to the larger community. Research and development in this area led to system enhancements from researchers in many laboratories, for example, BBN [67] and Philips [12]. By the mid-late 1980s, HMMs became the dominant recognition paradigm, with, for example, systems at SRI [48], MIT-Lincoln [55], and CMU. The CMU system was quite representative of the others developed at this time, and [38] provides an extended description.

Much of this activity focused on tasks defined in a new ARPA program. As in the 1970s, IBM researchers primarily worked with their own internal tasks, although ultimately they too participated in DARPA evaluations. See [61] for descriptions of the wide range of work done at Bell on HMMs for telephone speech, as well as on many other aspects of automatic speech recognition.

[7]The task was later called CSRNAB (Continuous Speech Recognition of North American Business News), which included data from other news sources.

4.7.4 The Second (D)ARPA Speech-Recognition Program

In 1984, ARPA began funding a second program. The first major speech-recognition task in this program was the Resource Management task mentioned earlier. This task involved reading sentences derived from a 1000-word vocabulary. The sentences were questions and commands designed to manipulate a naval information database, although the systems did not actually have to interface with any database; ratings were based on word recognition. Sample sentences from the corpus [60] include the following:

- Is Dixon's length greater than that of Ranger?
- What is the date and hour of arrival in port for Gitaro?
- Find Independence's alerts.
- Nevermind.

Evaluations of participating systems were held one to two times per year. Sites would receive a CD-ROM with test data, and send NIST the sentences produced by their recognizer, where the results would be officially evaluated.

The competition tended to make systems converge on good, similar systems, with each lab attempting to incorporate improvements that had been noted by the others. Al though this led to a rapid set of improvements, this also led to a convergence of approaches for many systems.

The ARPA project fueled many engineering advances. As of 1998, many research systems can recognize read speech from new speakers (without speaker-specific train ing) with a 60,000-word vocabulary in real time, with less than a 10% word error.[8] The competition also inspired other sites that were not funded by the project, including lab oratories in Europe. For example, Cambridge University in England participated in the evaluations, and developed HT or HMM ToolKit, which has been widely distributed [82]. It is now possible to use HT to get large vocabulary-recognition results close to those achieved by the major sites.

It could be argued that the fundamentals of speech-recognition science have not greatly changed in many years; at least it is not clear that any major mechanisms (of the significance of dynamic programming, HMMs, or LPC) were developed during the last decade or two. However, there have been many developments that may ultimately prove to have been important, particularly in the 1990s - examples include front-end developments (mei or bark-scaled cepstral estimates, delta features, channel normalization schemes, and vocal tract normalization) and probabilistic estimation (e.g., maximum likelihood linear

regression (MLLR) - see Chapter 28) to adapt to new speakers or acoustics, schemes to improve discrimination with neural networks, or training paradigms to maximize the mutual information between the data and the models). Still, it is fair to say that the field has matured to the point that the efforts of many workers in the field are more oriented toward improving the engineering effectiveness of existing ideas rather than generating radically different ones. It is a matter of current controversy as to whether such an engineering orientation is sufficient to make major progress in the future, or whether radically different approaches will actually be required [11].

4.7.5 The Return of Neural Nets

The field of neural networks suffered a large blow when Minsky and Papert wrote their 1969 book *Perceptrons,* proving that the perceptron, which was one of the popular net architec tures of the time,[9] could not even represent the simple exclusive or (XOR) function.[10] With the advent of backpropagation, a training technique for multilayer perceptrons (MLPs), in the early 1980s, the neural network field experienced a resurgence.

One application of neural networks to speech classification in the early 1980s was the use of a committee machine to judge whether a section of speech was voiced or unvoiced [26]. In 1983 Makino reported using a simple time-delayed neural network (a close cousin to a MLP in which the input layer includes a delayed version of itself in order to provide a simple context-delay mechanism) to perform consonant recognition [43]. This technique was later expanded by other researchers to add these delayed versions at multiple layers in the net [78]. Other researchers in the mid-1980s used Hopfield nets to classify both vowels and consonants [41].

By the late 1980s, many labs were experimenting with neural networks, both in isolated and continuous contexts. Only a few labs attacked large problems in automatic speech recognition with neural networks during this period; discrete probability estimators and mixtures of Gaussians were used in HMM recognizers for the majority of systems. Some sites have been using hybrid HMM-artificial neural network techniques, in which the neural network is used as a phonetic probability estimator, and the HMM is used to search through the possible space of word strings comprising the phones from the artificial neural network [46], [64]. In recent years, neural networks have also been used for feature transformation as part of a discriminatively trained front end for use in a Gaussian-mixture based recognition system [30].

4.7.6 Knowledge-Based Approaches

As noted previously, much of the work in the first ARPA speech project was strongly influenced by an artificial intelligence perspective. In the late 1970s and early 1980s, ap

[9]Although other network architectures were (and still are) available, including the perceptron's cousin, the MLP, the perceptron had properties that made it relatively easy to train.

[10]The XOR is a two-input logic function that returns true for inputs that are different (only one or the other is true) and false if the inputs are the same (either both true or both false).

proaches based on the codification of human knowledge, typically in the form of rales, became widely used in a number of disciplines. Some speech researchers developed recog nition systems that used acoustic-phonetic knowledge to develop classification rules for speech sounds; for instance, in [79], the consonants "k" and "g" following a vowel were

discriminanted on the basis of the proximity of the second and third resonances at the end of the vowel. This style of recognition was explained very well in [84]. One of the potential advantages of such an approach was that the speech characteristics used for discrimination were not limited to the acoustics of a single frame. Some of these points were explained in [16]. This reference, which is reprinted in [77], is also interesting because it includes a commentary from two BBN researchers (Makhoul and Schwartz), who took issue with the idea of focusing on the weak knowledge that we have about the utility of features chosen by experts. In this commentary, they suggested that systems should instead be focused on representing the ignorance that we have. In this case, they were really pointing to HMM
based approaches.[11] This dialog, and the personal interactions surrounding it at various meetings around this time, were extremely influential. By 1988 nearly every research site had turned to statistical methods. In the long term, however, the dichotomy might be viewed as elusive, since all of the researchers employing statistical methods continued to search for ways to include different knowledge sources, and the systems that attempted to use knowledge-based approaches also used statistical models.

4.8 MORE RECENT WORK

Since the early 1990s, there have been many events and advances in the field of speech recognition, though, arguably, few have had the fundamental impact of such things as the use of common databases and evaluations, and the core statistical modeling approach. However, the cumulative effect of these more recent efforts has been considerable. Here is a sampling of what we view to be the more significant components of work in ASR since the early 1990s.

1. The DARPA program continued, and moved on to tasks such as Broadcast News. This is a significantly more realistic task than the Wall Street Journal transcription, since it includes a range of speaking styles (from read to spontaneous) and acoustic conditions (e.g., quiet studio to noisy street). It also is a *real* task, in the sense that the automatic transcription of broadcast data is closely related to several potential commercial applications.

2. The U.S. Defense Department also funded an effort to transcribe conversational speech. Two databases collected for this work were Switchboard and Call Home; in the first case, talkers were asked to converse on the telephone on a selected topic (e.g., credit cards). In the second, callers were asked to telephone family members and discuss anything they wanted. These were, and are, extremely difficult tasks, though by 2005 the best systems achieved word error rates on Switchboard (and Fisher, a related task) in the

(Call Home test sets remained very challenging, possibly due to the extremely relaxed and informal conversational style used in talking to family members).

3. Beginning in the mid to late 1990s, there was an increased effort by a number of American and European laboratories to study conversational speech in meetings [45] [1]. This extended both the technical problem set and the potential utility of successful systems by including scenarios with distant microphones (thus generating noisy and reverberant speech signals) and extremely natural conversational phenomena. Data sets were collected and made available from the U.S. [33] and Europe [14]. For a number of years NIST conducted evaluations of ASR and speaker diarization for such test sets.

4. Beginning in 1993, there has been an annual 6-week summer workshop that is focused on recognizing conversational speech. It was held for 2 years at Rutgers, and then each summer at Johns Hopkins, although the latter workshop has since broadened its scope, looking at many problems in speech and natural language processing.

5. Many of the first speech recognizers were segment based; that is, the recognizer hypothesized the boundaries of phone segments in the speech signal and then tried to do recognition based on this segmented speech. By the 1970s, most researchers turned to a more frame-based system, in which the base acoustic analysis regions were small, constant-duration sections, or frames, of speech. However, some researchers continue to work with segment-based systems, e.g., the MIT SUMMIT system [85], [56]. These systems developed ways of using statistical models [28], much as the frame-oriented systems had. Additionally, a number of researchers developed ways of extending HMM based approaches to include segment statistics; see, for example, [54]. More recently, there has been increased experimentation with hybrids of HMM-based and memory-based (episodic) approaches, such as in [76].

6. Discriminative training methods are now widely used, particularly in large re search systems that are trained with many hours (often more than 1000) of speech. The availability of these very large data sets also permitted much more detailed models, which were a major factor in the reduction in word error rates since the early 1990s. Discriminative approaches to model training such as those developed in [58] will be discussed in Chapters 27 and 28; similar methods have been developed for feature transformation, as described in [59][30][47], and different discriminative methods have also been successfully combined, for instance as reported in [83].

7. Through the 1980s, essentially every recognition system was extremely suscep tible to a linear filtering operation (as one might experience from a telephone channel with a different frequency response than the one that was used to collect training data). In the 1990s there was significant work to improve recognition robustness to different channels, as well as to variability in the microphone, and to acoustic noise [31], [72], [25], [37]. This was continued in the following decade with work on the Aurora task, based on speech recognition with additive noise, and developed by a working group for the Euro pean Telecommunications Standards Institute (ETSI) as part of the selection process for a standard front end for "Distributed Speech Recognition" (DSR), which is discussed further in Chapter 22.

8. Language models have been extended by using many billions of words from available text, including incorporating language from the Web. Additionally, while n-gram

remain predominant, additional methods that incorporate more language structure have increasingly been incorporated as a supplement.

9. The use of multiple systems or subsystems became widespread for large research systems, with combination at the feature, model, or system output levels, as described

in[21],[71],[70], and [73].

10. There has been an increased emphasis on issues of pronunciation [63], dialog modeling [15], model adaptation schemes [39], and long-distance dependencies within word sequences [65], to mention just a few major topics.

11. There has been a rapid expansion of research in other classification tasks related to automatic speech recognition. For instance, methods and systems were developed for speaker identification and verification ([23], [24] and Chapter 41), as well as for language identification [49]. Speaker diarization (determining who spoke when) has also become a significant topic.

Some additional topics are discussed in a 2009 review paper, published in two parts [5][6].

4.9 SOME LESSONS

Researchers often return to the same themes decade after decade - frame-based measures versus segment-based ones, statistical estimation of acoustic and language probabilities, incorporation of speech knowledge, and so on. With each return, the technology is more sophisticated. For instance, consumers can now purchase a dictation system that can rec ognize tens of thousands of words in continuous speech with a moderate error rate (after adaptation to the speaker), and the computers that can accomplish this are widely avail able.

However, the problems in speech recognition remain deep. Even five-word rec ognizers operate with significant errors under common natural conditions (e.g., mod erate background noise and room reverberation, accent, and out-of-vocabulary words). In contrast, human performance is often far more stable under the same conditions, as discussed further in Chapter 18. We expect the general problem of the recognition and interpretation of spoken language to remain a challenging problem for some time to come.

EXERCISES

4.1 How was the 1952 Bell Labs automatic-speech recognition system limited in comparison with a modern system? Is there any way in which it could potentially be better, while keeping the same basic structure?

4.2 A new speech-recognition company is advertising their wonderful product. What percentage accuracy would you expect them to ascribe to their system? Describe some ways in which performance could be benchmarked in more realistic ways.

4.3 Find a newspaper, magazine, or Web announcement about some speech-recognition system, either commercial or academic. Can you conclude anything about the structure and capabilities of these systems? If there is any content in the release information, try to associate your best guesses about the systems with any of the historical developments described in this chapter.

4.4 In what way could Radio Rex be a better system than a recognizer trained to understand read versions of the Wall Street Journal?

BIBLIOGRAPHY

1. www.amiproject.org
2. Atal, B., and Hanauer, S., "Speech analysis and synthesis by prediction of the speech wave," *J. Acoust. Soc. Am.* 50: 637-655, 1971.
3. Bahl, L., and Jelinek, F., "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition," *IEEE Trans. Inform. Theory* IT-21: 404-411, 1975. 4. Baker, J., "The DRAGON system - an overview," *IEEE Trans. Acoust. Speech, Signal Process.* 23: 24-29, 1975.
5. Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., and O'Shaughnessy, D. "Developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Process. Mag.* 26(3): 75-80, May 2009.
6. Baker, J., Deng, L., Khudanpur, S., Lee, C.H., Glass, J., Morgan, N., and O'Shaughnessy, D. "Updated MINDS report on speech recognition and understanding, Part 2," *IEEE Signal Process. Mag.* 26(4): 78-85, July 2009.
7. Bakis, R., "Continuous-speech word spotting via centisecond acoustic states," IBM Res. Rep. RC 4788, Yorktown Heights, New York, 1974; abstract in *J. Acoust. Soc. Am.* 59 (Supp. 1): S 97, 1976.
8. Baum, L. E., and Petrie, T., "Statistical inference for probabilistic functions of finite state Markov chains," Аил. *Mathemat. Stat.* 37: 1554-1563, 1966.
9. Bellman, R., "On the theory of dynamic programming," *Proc. Nat. Acad. Sci.* 38:716-719,1952.
10. Bogert, B., Healy, M., and Tukey, J., "The quefrency analysis of time series for echos," in M. Rosenblatt, ed., *Proc. Symp. on Time Series Analysis,* Chap. 15, Wiley, New York, pp. 209-243, 1963.
11. Bourlard, H., Hermansky, H., and Morgan, N., "Towards increasing speech recognition error rates," *Speech Commun.* 18: 205-231, 1996.
12. Bourlard, H., Kamp, Y, Ney, H., and Wellekens, C. J., "Speaker-dependent connected speech recognition via dynamic programming and statistical methods," in M. R. Schroeder, ed., *Speech and Speaker Recognition,* Karger, Basel, 1985.

13. Bridle, J., Chamberlain, R., and Brown, M., "An algorithm for connected word recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Paris, pp. 899-902, 1982. 14. Cadetta, J. "Announcing the AMI Meeting Corpus.," *The ELRA Newsletter* 11(1): 3-5, January-March 2006.
15. Cohen, P., "Dialogue modeling," in R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampoli, and V. Zue, eds. *Survey of the State of the Art in Human Language Technology,* Cambridge Univ. Press, London/New York, 1997.
16. Cole, R., Stern, R., and Lasry, M., "Performing fine phonetic distinctions: templates versus features," in J. S. Perkell and D. M. Klatt, eds., *Variability and Invariance in Speech Processes,* Erlbaum, Hillsdale, N.J., 1986.

17. Cooley, J. W., and Tukey, J. W., "An algorithm for the machine computation of complex Fourier series," *Math. Comput.* 19: 297-301, 1965.
18. David, E., and Selfridge, O., "Eyes and ears for computers," *Proc. IRE* 50: 1093-1101, 1962.
19. Davis, K., Biddulph, R., and Balashek, S., "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.* 24: 637-642, 1952.
20. Davis, S., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.* 28: 357-366, 1980.
21. Fiscus, J. G., "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. Auto. Speech Recog. linderst.,* Santa Barbara, pp. 347-354, 1997.
22. Furui, S., "Speaker independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. Acoust. Speech Signal Process.* 34: 52-59, 1986.
23. Furui, S., "An overview of speaker recognition technology," in C. H. Lee, F. K. Soong, and K. K. Paliwal, eds., *Automatic Speech and Speaker Recognition,* Kluwer, Boston, Mass., 1996. 24. Furui, S., "40 Years of Progress in Automatic Speaker Recognition," in *Adv. Biometrics,* pp. 1050-1059, 2009.
25. Gales, M., and Young, S., "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comput. Speech Lang.* 9: 289-307, 1995.
26. Gevins, A., and Morgan, N., "Ignorance-based systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* San Diego, pp. 39A.5.1-39A.5.4., 1984.
27. Ghitza, O., "Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment," *J. Phonet.* 16: 109-124, 1988. 28. Glass, J.R., "A probabilistic framework for segment- based speech recognition," *Comput, Speech Lang.* 17(2-3): 137-152,2003.
29. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.* 87: 1738-52, 1990.
30. Hermansky, H., Ellis, D., and Sharma, S., "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE Int. Confi Acoust. Speech Signal Process.,* Istanbul, pp. Ill-1635-1638, 2000.
31. Hermansky, H., and Morgan, N., "RASTA processing of speech," *IEEE Trans. Speech Audio Process.* 2: 578-589, 1994.
32. Itakura, F., and Saito, S., "Analysis-synthesis telephone based on the maximum-likelihood

35. Jelinek, F., Bahl, L., and Mercer, R., "The design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Inform. Theory* **IT-21:** 250-256, 1975.

36. Klatt, D., "Review of the ARPA speech understanding project," *J. Acoust. Soc. Am.* 62: 1345-1366, 1977.

37. Lee, C.-H., "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.* 25: 29-48, 1998.

38. Lee, K.-R, *Automatic Speech Recognition - the Development of the Sphinx System,* Kluwer, Norwell, Mass., 1989.

39. Leggetter, C, and Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.* 9: 171-185, 1995.

40. Leung, H., and Zue, V., "A procedure for automatic alignment of phonetic transcriptions with continuous speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* San Diego, pp. 2.7.1-2.7.4, 1984.

41. Lippmann, R., and Gold, B., "Neural classifiers useful for speech recognition," in *Proc. IEEE First Int. Conf. Neural Net.,* San Diego, pp. 417^22, 1987.

42. Makhoul, J., "Linear prediction: a tutorial review," *Proc. IEEE* 63: 561-580, 1975.

43. Makino, S., Kawabata, T., and Kido, K., "Recognition of consonants based on the perceptron model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Boston, Mass., pp. 738-741, 1983.

44. Markel, J., and Gray, A., *Linear Prediction of Speech,* Springer-Verlag, New York/Berlin, 1976.

45. Morgan, N., Baron, D., Bhagat, S. Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C, "Meetings about meetings: research at ICSI on speech in multiparty conversations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Hong Kong, pp. IV 740-743, 2003.

46. Morgan, N., and Bourlard, H., "Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Process. Mag.* 12: 25^12, 1995.

47. Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Bourlard, H., and Athineos, M., "Pushing the envelope-aside," *IEEE Signal Process. Mag.* 22(5): 81-88, 2005.

48. Murveit, H., Cohen, M., Price, P., Baldwin, G., Weintraub, M., and Bernstein, J., "SRI's DECI PHER system," in *Proc. Speech Natural Lang. Workshop,* Philadelphia, pp. 238-242, 1989. 49. Muthusamy, Y. K., Barnard, E., and Cole, R. A., "Reviewing automatic language identification," *IEEE Signal Process. Mag.* **11:** 33^11, 1994.

50. Myers, C, Rabiner, L., and Rosenberg, L., "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust. Speech Signal Process.* 28: 623-635, 1980.

51. Ney, H., "The use of a one stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust. Speech Signal Process.* 32: 263-271, 1984.

52. National Institute of Standards and Technology, *TIMIT Acoustic-Phonetic Continuous Speech Corpus,* Speech Disc 1-1.1, NIST Order No. PB91-505065, 1990.

53. Oppenheim, A. V, Schäfer, R. W., and Stockham, T. G. Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE* 56: 1264-1291, 1968.

54. Ostendorf, M., Bechwati, I., and Kimball, O., "Context modeling with the stochastic segment model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* San Francisco, pp. 389-392, 1992.

55. Paul, D., "The Lincoln continuous speech recognition system: recent developments and results," in *Proc. Speech Natural Lang. Workshop,* Philadelphia, pp. 160-165, 1989.

56. Phillips, M., Glass, J., and Zue, V., "Automatic learning of lexical representations for sub-word unit based speech recognition systems," *Proc. Eurospeech,* Genova, pp. 577-580, 1991.

57. Pierce, J., "Whither speech recognition," *J. Acoust. Soc. Am.* 46: 1049-1051, 1969. 58. Povey, D., *Discriminative Training for Large Vocabulary Speech Recognition,* Ph. D. Thesis, Cambridge University, 2004.

59. Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., NFMPE: Discrimina tively trained features for speech recognition, in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* Philadelphia, pp. 96le 964, 2005.

60. Price, P., Fisher, W., Bernstein, J., and Pallett, D., "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* New York, S.13.21, pp. 651-654, 1988.

61. Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition,* Prentice-Hall, Englewood Cliffs, N.J., 1993.

62. Rabiner, L., and Levinson, S., "Isolated and connected word recognition: theory and selected applications," *IEEE Trans. Commun.* 29: 621-659, 1981.

63. Riley, M., and Ljolje, A., "Automatic generation of detailed pronunciation lexicons," in C. H. Lee, F. K. Soong, and K. K. Paliwal, eds., *Automatic Speech and Speaker Recognition,* Kluwer, Boston, Mass., 1996.

64. Robinson, T., Hochberg, M., and Renals, S., "The use of recurrent neural networks in continuous speech recognition," in C. H. Lee, F. K. Soong, and K. K. Paliwal, eds., *Automatic Speech and Speaker Recognition,* Kluwer, Boston, Mass., 1996.

65. Rosenfeld, R., "A maximum entropy approach to adaptive statistical language modeling," *Comput. Speech Lang.* 10: 187-228, 1996.

66. Sakoe, H., and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech Signal Process.* 26: 43^49, 1978. 67. Schwartz, R., Chow, Y, Kimball, O., Roucos S., Krasner, M., and Makhoul, J., "Context dependent modeling for acoustic-phonetic recognition of continuous speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Tampa, pp. 1205-1208, 1985.

68. Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonet.* 16: 55-76, 1988.

69. Shannon, C, "A mathematical theory of communication," *Bell Sys. Tech. J.* 27: 379^123, 623-656, 1948.

70. Sinha, R., Gales, M., Kim, D. Y., Liu, X. A., Sim, K. C, Woodland, P. C, "The CU-HTK Mandarin Broadcast News Transcription System," in *Proc. IEEE Int. Confi Acoust. Speech Signal Process.,* Toulouse, pp. 1-1077-1080, 2006.

71. Siohan, O., Ramabhadran, B., and Kingsbury, B., "Constructing Ensembles of ASR Systems Using Randomized Decision Trees," in *Proc. IEEE Int. Confi Acoust. Speech Signal Process.,* Philadelphia, pp. 1-197-200, 2005.

72. Stern, R., Acero, A., Liu, F.-H., and Ohshima, Y, "Signal processing for robust speech recognition," in C. H. Lee, F. K. Soong, and K. K. Paliwal, eds., *Automatic Speech and Speaker Recognition,* Kluwer, Boston, Mass., 1996.

73. Stolcke, A., Chen, B., Franco, H., Gadde, V.R.R., Graciarena, M., Hwang, M.-Y, Kirchhoff, K., Morgan, N., Lin, X., Ng, T., Ostendorf, M., SÄnmez, K., Venkataraman, A., Vergyri, D., Wang, W, Zheng, J., and Zhu, Q., "Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW" *IEEE Trans. Audio Speech Lang. Process.,* 14(5): 1729-1744, 2006.

74. Tappert, C, Dixon, N., Rabinowitz, A., and Chapman, W., "Automatic recognition of continuous

based continuous speech recognition," in *Proc. Eurospeech,* Geneva, pp. 1133-1136, 2003. 77. Waibel, A., and Lee, K., eds., *Readings in Speech Recognition,* Morgan Kaufmann, San Mateo, Calif., 1990.

78. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., "Phoneme recognition: neural networks vs. hidden Markov models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* New York, pp. 107-110, 1988.

79. Weinstein, C, McCandless, S., Mondshein, L., and Zue, V., "A system for acoustic-phonetic analysis of continuous speech," *IEEE Trans. Acoust. Speech Signal Process.* 23: 54—67, 1975. 80. White, G., "Speech classification using linear time stretching or dynamic programming," *IEEE Trans. Acoust. Speech Signal Process.* 24(2): 183-188, 1976.

81. Widrow, B., Personal Communication, Phoenix, Az., 1999.

82. Woodland, P., Odell, J., Valtchev, V., and Young, S., "Large vocabulary continuous speech recognition using HTK," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Adelaide, pp. 11-125-128, 1994.

83. Zheng,, J., Cetin, O., Huang, M.-Y, Lei, X., Stolcke, A., and Morgan, N., "Combining Discrim inative Feature, Transform, and Model Training for Large Vocabulary Speech Recognition" in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Honolulu, pp. 633-636, 2007.

84. Zue, V., "The use of speech knowledge in automatic speech recognition," *Proc. IEEE* 73: 1602-1615, 1985.

85. Zue, V., Glass, J., Phillips, M., and Seneff, S., "The MIT SUMMIT speech recognition system: a progress report," in *Proc. Speech Natural Lang. Workshop,* Philadelphia, pp. 179-189, 1989.

CHAPTER 3

SPEECH-RECOGNITION OVERVIEW

5.1 WHY STUDY AUTOMATIC SPEECH RECOGNITION?

Why do we study automatic speech recognition (ASR)? For one thing, there is a lot of money at stake: speech recognition is potentially a multi-billion-dollar industry in the near future. As of 2011, earnings (and savings) from simple telephone applications are reputed to be billions of dollars per year.

There are many aspects of speech recognition that are already well understood. However, it is also clear that there is much that we still don't know. We don't have human quality speech recognition; performance degrades rapidly when small changes are made to the speech signal, such as those that can be caused from switching microphones.

Speech recognition is potentially very useful. Sample applications include the following.

Telephone applications: For many current voice-mail systems, one has to follow a

series of touch-tone button presses to navigate through a hierarchical menu. Speech recognition has the potential to cut through the menu hierarchy, although simple "press or say one" speech applications do not do this. Many "smart phones" now also incorporate speech recognition, for instance to simplify dialing a number in the phone's contact list.

Hands-free operation: There are many situations in which hands are not available to issue commands to a device. Using a car phone and controlling the microscope position in an operating room are two examples for which some limited vocabulary systems already exist.

Applications for the physically handicapped: Speech recognition is a natural alter native interface to computers for people with limited mobility in their arms and hands, or for those with sight limitations.

For some aspects of computer applications, speech may be a more natural interface than a keyboard or mouse.

Dictation: General dictation is an advanced application, requiring a much larger vocabulary than, for instance, replacing a menu system. Dictation systems currently accept continuous, large vocabulary input, and work well for may people when trained specifically for that person.

Translation: Another advanced application is translation from one language to another. The Verbmobil project in Germany was both a collaborative and competitive effort to provide language-to-language translation. The goal was to facilitate a conversa tion between native speakers of German and Japanese, using English as an intermediate

and phrases as needed from German into English (the speaker is assumed to be moderately competent in English). A number of U.S. projects such as Transtac are also have attempted to provide two-way spoken translation.

5.2 WHY IS AUTOMATIC SPEECH RECOGNITION HARD?

There are many reasons why speech recognition is often quite difficult.

First, natural speech is continuous; it often doesn't have pauses between the words. This makes it difficult to determine where the word boundaries are, among other things. Also, natural speech contains disfluencies. Speakers change their mind in midsentence about what they want to say, will often accidentally switch phones (as in the phrase "teep кape," which means "keep a tape"), and utter filled pauses (e.g., "uh" and "urn") while they are thinking of their next message.

Second, natural speech can also change with differences in global or local rates of speech, pronunciations of words within and across speakers, and phonemes in different contexts. As a result, we can't just say that X is the spectral representation that corresponds to "uh." The spectrum will change, often quite dramatically, if any of these conditions are changed.

Third, large vocabularies are often confusable. A 20,000-word vocabulary is more likely to have more words that sound like each other than a 10-word vocabulary. There is also the issue of out-of-vocabulary words; for some tasks, no matter what words are in a vocabulary, recognition will always encounter words that have not been seen before. How to model these unknown words is an important unsolved problem.

Fourth, as noted previously, recorded speech is variable over room acoustics, channel characteristics, microphone characteristics, and background noise. In telephone speech, the channel used by the telephone company on any particular call (especially for analog seg ments) will have spectral and temporal effects on the transmitted speech signal. Background noise and acoustics in the environment that a telephone speaker is in will also have tangible effects on the signal. Different handsets, or in general different microphones, have different frequency responses; tilting a microphone at different angles will also change the frequency response. Nonlinear effects are particularly significant in carbon-granule microphones, but in general they can complicate the effects of using a particular handset. Some effects will be phone dependent; for instance, nasal sounds may be louder if the microphone is closer to the nose.

All of these factors can change the characteristics of the speech signal - a difference that humans can often compensate for, but that current recognition systems often cannot. The algorithms for training recognition systems must be chosen carefully, for large training times are not practical for research purposes. Algorithms that take a year to run on available hardware may be of great theoretical interest, but since most programs have bugs, such a choice does not really permit the development of an experimental approach. To replace other input modes with speech recognition, a high level of performance must be obtained. This does not necessarily mean near-perfect accuracy (although certainly too many errors can be very frustrating); perhaps it is just as important that recognition