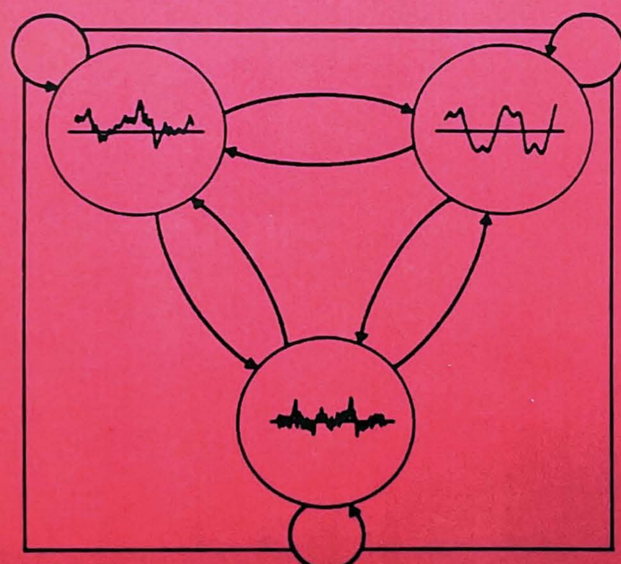
FUNDAMENTALS OF SPEECH RECOGNITION



LAWRENCE RABINER
BIING-HWANG JUANG

PRENTICE HALL SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR

FUNDAMENTALS OF SPEECH RECOGNITION

Lawrence Rabiner
Biing-Hwang Juang

Prentice Hall P T R
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Rabiner, Lawrence R., 1943–
Fundamentals of speech recognition / Lawrence Rabiner, Biing-Hwang Juang.
p. cm.
Includes bibliographical references and index.
ISBN 0–13–015157–2
1. Automatic speech recognition. 2. Speech processing systems.
I. Juang, B. H. (Biing-Hwang) II. Title.
TK7895.S65R33 1993
006.4'54—dc20 92–34093
CIP

Editorial production and interior design: *bookworks*
Cover Designer: *Ben Santora*
Acquisitions Editor: *Karen Gettman*
Manufacturing Buyer: *Mary Elizabeth McCartney*

Pearson Education

© Published by Prentice Hall PTR
Prentice-Hall, Inc.
A Pearson Education Company
Upper Saddle River, NJ 07458

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact:

Corporate Sales Department
PTR Prentice Hall

Printed in the United States of America
10 9 8

ISBN 0-13-015157-2

Prentice-Hall International (UK) Limited, London
Prentice-Hall of Australia Pty. Limited, Sydney
Prentice-Hall Canada Inc., Toronto
Prentice-Hall Hispanoamericana, S.A., Mexico
Prentice-Hall of India Private Limited, New Delhi
Prentice-Hall of Japan, Inc., Tokyo
Pearson Education Asia Pte. Ltd., Singapore
Editora Prentice-Hall do Brasil, Ltda., Rio de Janeiro

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This book is an outgrowth of an association between the authors which started over 10 years ago when one of us (BHJ) was a graduate student at the University of California at Santa Barbara and the other (LRR) was a supervisor at AT&T Bell Laboratories. We began our relationship with a mutual interest in the problem of designing and implementing vector quantization for speech processing. This association turned into a full technical partnership and strong friendship when Fred Juang joined Bell Laboratories, initially in the development area and subsequently in research. The spark that ignited formal work on this book was a series of short courses taught by one of us (LRR) on speech recognition. After several iterations of teaching, it became clear that the area of speech recognition, although still changing and growing, had matured to the point where a book that covered its theoretical underpinnings was warranted.

Once we had decided to write this book, there were several key issues that had to be resolved, including how deep to go into areas like linguistics, natural language processing, and the practical side of the problem; whether to discuss individual systems proposed by various research labs around the world; and how extensively to cover applications. Although there were no simple answers to these questions, it rapidly became obvious to us that the fundamental goal of the book would be to provide a theoretically sound, technically accurate, and reasonably complete description of the basic knowledge and ideas that constitute a modern system for speech recognition by machine. With these basic guiding principles in mind, we were able to decide consistently (and hopefully reasonably) what material had to be included, and what material would be presented in only a cursory manner. We leave it up to you, the reader, to decide if our choices have been wise ones.

The formal organization of the book is as follows. Chapter 1, called "Fundamentals of Speech Recognition," provides an overview of the entire field with a discussion of the breadth and depth of the various disciplines that are required for a deep understanding of all aspects of speech recognition. The concept of a task-oriented, speech-recognition system is introduced and it is shown that "base level" speech or sound recognition is only one step in a much larger process where higher-level task information, in the form of syntax, semantics, and pragmatics, can often play a major role. After a formal description of the material to be covered in each of the chapters, we give a brief history of speech recognition research in order to put the material presented in this book in its proper perspective.

Chapter 2, entitled the "The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization," provides a review of the theory of acoustic-phonetics in which we try to characterize basic speech sounds according to both their linguistic properties and the associated acoustic measurements. We show that although there is a solid basis for the linguistic description of sounds and a good understanding of the associated acoustics of sound production, there is, at best, a tenuous relationship between a given linguistic sound and a repeatable, reliable, measureable set of acoustic parameters. As such a wide variety of approaches to speech recognition have been proposed, including those based on the ideas of acoustic-phonetics, statistical pattern-recognition methods, and artificial intelligence (so-called expert system) ideas. We discuss the relative advantages and disadvantages of each of these approaches and show why, on balance, the pattern-recognition approach has become the method of choice for most modern systems.

In Chapter 3, entitled "Signal Processing and Analysis Methods for Speech Recognition," we discuss the fundamental techniques used to provide the speech features used in all recognition systems. In particular we discuss two well-known and widely used methods of spectrum analysis, namely the filter bank approach and the linear prediction method. We also show how the method of vector quantization can be used to code a spectral vector into one of a fixed number of discrete symbols in order to reduce the computation required in a practical system. Finally we discuss an advanced spectral analysis method that is based on processing within the human auditory system—an ear model. The ultimate goal of such a system is to increase the robustness of the signal representation and make the system relatively insensitive to noise and reverberation, in much the same way as the human ear.

Chapter 4, entitled "Pattern-Comparison Techniques," deals with the fundamental problems of defining speech feature vector patterns (from spoken input), and comparing pairs of feature vector patterns both locally (i.e., at some point in time), and globally (i.e., over the entire pattern) so as to derive a measure of similarity between patterns. To solve this pattern-comparison problem requires three types of algorithms, namely a speech-detection method (which essentially separates the speech signal from the background), a spectral vector comparison method (which compares two individual spectral vectors), and a global pattern comparison method which aligns the two patterns locally in time and compares the aligned patterns over the entire duration of the patterns. It is shown that a key issue is the way in which time alignment between patterns is achieved.

Chapter 5, entitled "Speech Recognition System Design and Implementation Issues," discusses the key issues of training a speech recognizer and adapting the recognizer parameters to different speakers, transmission conditions, and speaking environments. A key concept in most modern systems is that of learning, namely improving recognizer performance over time based on additional training provided by the user of the system. Adaptation methods provide a formalism for such learning.

In Chapter 6, "Theory and Implementation of Hidden Markov Models," we discuss a basic set of statistical modeling techniques for characterizing speech. The collection of methods, popularly called Hidden Markov Models, is a powerful set of tools for providing a statistical model of both the static properties of sounds and the dynamical changes that occur across sounds. Methods for time aligning patterns with models are discussed along with different ways of building the statistical models based on the type of representation, the sound being modeled, the class of talkers, and so forth.

Chapters 7 and 8, entitled "Speech Recognition Based on Connected Word Models" and "Large Vocabulary Continuous Speech Recognition," extend the speech-recognition problem from single word sequences to fluent speech. Modeling techniques based on whole word models are discussed in Chapter 7 where we assume that we are interested in recognizing sequences of digits, alphanumerics, and so forth. For this type of system whole-word models are most reasonable since the vocabulary is typically small and highly constrained. Hence the statistical properties of the word models, in all word contexts, can be learned from a reasonably sized training set. Modeling techniques based on subword units are discussed in Chapter 8 where we assume unlimited size vocabulary. Hence a key issue is what units are used, how context dependent the units should be, how unit models are trained reliably (and robustly to different vocabularies and tasks), and how large vocabulary recognition systems based on such units are efficiently implemented.

Finally, in Chapter 9, entitled "Task-Oriented Applications of Automatic Speech Recognition," we come full circle and return to the concept of a task-oriented system. We discuss the basic principles that make some tasks successful while others fail. By way of example we discuss, in fairly general terms, a couple of task-oriented recognizers and show how they perform in practice.

The material in this book is primarily intended for the practicing engineer, scientist, linguist, programmer, and so forth, who wants to learn more about this fascinating field. We assume a basic knowledge of signal processing and linear system theory as provided in an entry level course in digital signal processing. Although not intended as a formal university course, the material in this book is indeed suitable for a one-semester course at the graduate or high undergraduate level. Within almost every chapter we have provided "exercises" for the student to assess how well he or she understands the material. Solutions to the exercises are provided immediately following the exercise. Hence, for maximum effectiveness, each student must exercise self-discipline to work through an answer before comparing it with the published solution.

In order to truly understand the fundamentals of speech recognition, a person needs hands-on experience with the software, hardware, and platforms. Hence we strongly encourage all serious readers of this book to program the algorithms, implement the systems, and literally build applications. Without such practical experience the words in this book will not come alive for most people.

ACKNOWLEDGMENTS

Although the authors take full responsibility for the material presented in this book, we owe a great debt to our colleagues, both with AT&T Bell Laboratories and outside, for their many technical contributions which underly the material presented. In particular the authors owe a debt of gratitude to Dr. James L. Flanagan (currently director of the CAIP Institute at Rutgers University) for his roles in guiding and shaping both our careers and the field of speech processing. Without Jim's understanding and inspiration, this book would never have existed.

The number of people who have made substantial contributions to speech recognition are too numerous to mention. However there are three individuals who have had a profound influence on the field and they deserve special mention. The first is Professor Raj Reddy of Carnegie-Mellon University who was essentially the first person to realize the vast potential of speech recognition and has devoted over 25 years as a leader, innovator, and educator in this field. The second individual of note is Dr. Jack Ferguson (retired from Institute for Defense Analyses in Princeton) who is the person most responsible for development of the theory of the Hidden Markov Model as applied to speech recognition. Dr. Ferguson, as editor of the key textbook in this area and lecturer, par excellence, has spread the word on Hidden Markov Models so that this technology has rapidly risen from technical obscurity to become the preeminent method of speech recognition today. Finally, the third individual of note is Dr. Fred Jelinek of IBM, who has led the world's largest speech-recognition research group for almost two decades and has been responsible for a large number of major innovations in large vocabulary speech recognition. These three individuals have played major roles in nurturing the technology and enabling it to reach the state of maturity it has achieved today.

Within Bell Laboratories the authors have drawn freely from the research of our former and current colleagues. We would like to acknowledge the direct support and contributions of the following individuals: Prof. Ronald Schafer (currently at Georgia Tech.), Dr. Steve Levinson, Dr. Bishnu Atal, Dr. Esther Levin, Dr. Tali Tishby (currently at the Hebrew University in Jerusalem), Dr. Oded Ghitza, Jay Wilpon, Dr. Frank Soong, Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Cory Myers (currently at Atlantic Aerospace), Dr. Aaron Rosenberg, Dr. Chin Lee and Dr. Roberto Pieraccini. We thank these colleagues for their research contributions and for their friendship and guidance over the years.

Several individuals provided technical comments on a preliminary version of the manuscript. These comments were invariably insightful and provided valuable feedback on ways to improve the book. We would like to acknowledge and thank the following individuals for their help: Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Esther Levin, Dr. Oded Ghitza, Dr. Roberto Pieraccini, Dr. Chin Lee, Dr. Wu Chou and Dr. Sadaoki Furui (NTT Corporation, JAPAN).

The production of a book is an essentially infinite task which is seemingly an endless one. However all good things do come to an end and this one was no exception. The authors owe a great deal to Ms. Martina Sharp of the Bell Labs Word Processing group, who entered all the text material for the book using the LaTeX system. Tina worked on three

generations of the manuscript with a degree of skill that is essentially unmatched among her peers. It was a pleasure working with Tina and we look forward to future projects with her. Most of the art work for the book was produced by Ms. Danuta Sowinska-Khan of the Bell Labs art department. Danuta was a pleasure to work with and is a professional in her trade. The material she produced always met the highest professional standards and did honor to the manuscript in which it appears. A great deal of assistance in the preparation of drafts of the book, getting figures ready for production, and so forth was provided by Irene Morrongiello. We thank her and wish to express our appreciation for a job well done.

Finally, to the crew at Prentice Hall we owe a debt of gratitude for the professional way the book was handled from inception to final production. Karen Gettman provided the incentive to the authors to actually produce the manuscript and make the book a reality. We thank Lisa Garboski, the production editor, for help in all phases of the book production cycle.

Lawrence R. Rabiner
Biing-Hwang Juang

Chapter 1

FUNDAMENTALS OF SPEECH RECOGNITION

1.1 INTRODUCTION

Automatic recognition of speech by machine has been a goal of research for more than four decades and has inspired such science fiction wonders as the computer HAL in Stanley Kubrick's famous movie *2001—A Space Odyssey* and the robot R2D2 in the George Lucas classic *Star Wars* series of movies. However, in spite of the glamour of designing an intelligent machine that can recognize the spoken word and comprehend its meaning, and in spite of the enormous research efforts spent in trying to create such a machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments. Thus, an important question in this book is, What do we mean by "speech recognition by machine." Another important question is, How can we build a series of bridges that will enable us to advance both our knowledge as well as the capabilities of modern speech-recognition systems so that the "holy grail" of conversational speech recognition and understanding by machine is attained?

Because we do not know how to solve the ultimate challenge of speech recognition, our goal in this book is to give a series of presentations on the fundamental principles of most modern, successful speech-recognition systems so as to provide a framework from which researchers can expand the frontier. We will attempt to avoid making absolute judgments on the relative merits of various approaches to particular speech-recognition problems. Instead we will provide the theoretical background and justification for each topic discussed so that the reader is able to understand why the techniques have proved valuable and how they can be used to advantage in practical situations.

One of the most difficult aspects of performing research in speech recognition by machine is its interdisciplinary nature, and the tendency of most researchers to apply a monolithic approach to individual problems. Consider the disciplines that have been applied to one or more speech-recognition problems:

1. **signal processing**—the process of extracting relevant information from the speech signal in an efficient, robust manner. Included in signal processing is the form of spectral analysis used to characterize the time-varying properties of the speech signal as well as various types of signal preprocessing (and postprocessing) to make the speech signal robust to the recording environment (signal enhancement).
2. **physics (acoustics)**—the science of understanding the relationship between the physical speech signal and the physiological mechanisms (the human vocal tract mechanism) that produced the speech and with which the speech is perceived (the human hearing mechanism).
3. **pattern recognition**—the set of algorithms used to cluster data to create one or more prototypical patterns of a data ensemble, and to match (compare) a pair of patterns on the basis of feature measurements of the patterns.
4. **communication and information theory**—the procedures for estimating parameters of statistical models; the methods for detecting the presence of particular speech patterns, the set of modern coding and decoding algorithms (including dynamic programming, stack algorithms, and Viterbi decoding) used to search a large but finite grid for a best path corresponding to a "best" recognized sequence of words.
5. **linguistics**—the relationships between sounds (phonology), words in a language (syntax), meaning of spoken words (semantics), and sense derived from meaning (pragmatics). Included within this discipline are the methodology of grammar and language parsing.
6. **physiology**—understanding of the higher-order mechanisms within the human central nervous system that account for speech production and perception in human beings. Many modern techniques try to embed this type of knowledge within the framework of artificial neural networks (which depend heavily on several of the above disciplines).
7. **computer science**—the study of efficient algorithms for implementing, in software or hardware, the various methods used in a practical speech-recognition system.
8. **psychology**—the science of understanding the factors that enable a technology to be used by human beings in practical tasks.

Successful speech-recognition systems require knowledge and expertise from a wide range of disciplines, a range far larger than any single person can possess. Therefore, it is especially important for a researcher to have a good understanding of the fundamentals of speech recognition (so that a range of techniques can be applied to a variety of problems), without necessarily having to be an expert in each aspect of the problem. It is the purpose of this book to provide this expertise by giving in-depth discussions of a number of

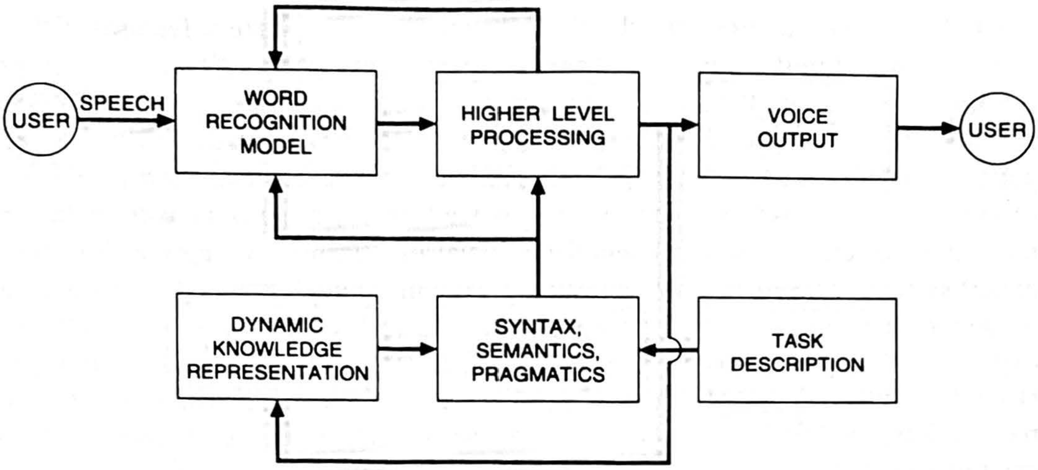


Figure 1.1 General block diagram of a task-oriented speech-recognition system.

fundamental topics in speech-recognition research.

1.2 THE PARADIGM FOR SPEECH RECOGNITION

A general model for speech recognition, shown in Figure 1.1, is used throughout this book. The model begins with a user creating a speech signal (speaking) to accomplish a given task. The spoken output is first recognized in that the speech signal is decoded into a series of words that are meaningful according to the syntax, semantics, and pragmatics of the recognition task. The meaning of the recognized words is obtained by a higher-level processor that uses a dynamic knowledge representation to modify the syntax, semantics, and pragmatics according to the context of what it has previously recognized. In this manner, things such as non sequitors are omitted from consideration at the risk of misunderstanding, but at the gain of minimizing errors for sequentially meaningful inputs. The feedback from the higher-level processing box reduces the complexity of the recognition model by limiting the search for valid (acceptable) input sentences (speech) from the user. The recognition system responds to the user in the form of a voice output, or equivalently, in the form of the requested action being performed, with the user being prompted for more input.

1.3 OUTLINE

The material in this book is organized into nine chapters. Chapters 2 through 9 each deals with a basic concept or a fundamental technique used in various speech-recognition systems. The material discussed in these chapters is as follows.

Chapter 2—The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization. In this chapter we review the speech production/perception process in human beings. We show how different speech sounds can be characterized by a set of spectral and temporal properties that depend on the acoustic-phonetic features of the sound and are manifest in the waveform, the sound spectrogram, or both. Included in the chapter is an overview of the three most common approaches to speech recognition, namely the acoustic-phonetic approach (which tries to directly exploit individual sound properties), the pattern recognition approach (which relies only on gross spectral and temporal properties of speech sounds and uses conventional as well as neural network pattern recognition technology to classify sounds), and the artificial intelligence (AI) approach in which an expert system or a self-organizing (learning) system, as implemented by neural networks, is used to classify sounds. We discuss the strengths and weaknesses of each approach and explain why the pattern-recognition approach is the one most heavily relied on in practical systems. We conclude the chapter with a discussion of the fundamental issues in speech recognition (i.e., those factors that most influence overall system performance), and with a brief overview of current applications.

Chapter 3—Signal Processing and Analysis Methods for Speech Recognition. In this chapter we present the two fundamental signal-processing approaches to speech spectral analysis: filter bank and linear predictive methods. We specialize the presentation of these two fundamental techniques to aspects related to speech analysis and compare and contrast the two methods in terms of robustness to speech sounds and required computation. For completeness we also discuss the popular source-coding technique referred to as vector quantization (VQ). Here, a codebook is created to represent the anticipated range of spectral vectors. This enables us to code an arbitrary continuous speech spectral vector into one of a fixed number of discrete codebook symbols at the cost of increased error in signal representation but with the benefit of significantly reduced computation in the recognition process. We conclude this chapter with a discussion of a spectral analysis model that attempts to mimic the processing in the human auditory system—the so-called ear model. Although our knowledge of the higher-order processing in the central nervous system is rudimentary, the importance of ear models is related to their robustness to noise, reverberation, and other environmental factors that often seriously degrade performance of current speech recognizers.

Chapter 4—Pattern-Comparison Techniques. In this chapter we discuss three fundamental aspects of comparing a pair of speech patterns. These are the basic concept of detecting speech (i.e., finding the speech signal in a background of noise or other acoustic interference), the idea of computing a measure of the local distance (or similarity) of a pair of spectral representations of a short-time piece of speech signal (a distance or distortion measure), and the concept of temporally aligning and globally comparing a pair of speech patterns corresponding to different speech utterances that may or may not be the same sequence of sounds or words (dynamic time warping algorithms). We show in this chapter how the basic pattern-comparison techniques can be combined in a uniform framework for speech-recognition applications.

Chapter 5—Speech-Recognition System Design and Implementation Issues. In this chapter we discuss the remaining pieces (after signal processing and pattern comparison) that enable us to build and study performance of a practical speech-recognition system.

In particular we discuss how speech recognizers are trained and how we can enhance the basic recognition procedure by adding features, by exploiting a preprocessor, by the use of methods of adaptation or by postprocessing the recognizer outputs using a set of pattern discriminators (as opposed to the pattern classifiers used in a conventional implementation). We conclude the chapter with a discussion of various ways of recognizing speech in adverse environments (e.g., noise, stress conditions, or mismatch between training and testing).

Chapter 6—Theory and Implementation of Hidden Markov Models. In this chapter we discuss aspects of the theory and implementation of the set of statistical modeling techniques collectively referred to as hidden Markov modeling. Included within these techniques are the algorithms for scoring a statistical (Markovian) model against a speech pattern, the techniques for aligning the model with the speech pattern so as to recover an estimate of the alignment path between different speech sounds and different model states, and the techniques for estimating parameters of the statistical models from a training set of utterances of the sounds being modeled. Also included is a discussion of the practical aspects of building hidden Markov models, including the issues of scaling of data, handling of multiple observation sequences, providing initial estimates of model parameters, and combating the problems of insufficient training data. We conclude the chapter with a practical example illustrating how a simple, isolated word recognizer would be implemented using hidden Markov models.

Chapter 7—Speech Recognition Based on Connected Word Models. In this chapter we show how the basic set of techniques developed for recognizing an isolated word or phrase can be readily extended to recognizing a sequence of words (e.g., a string of digits of a credit card number) spoken in a fluent or connected manner. We make the simplifying assumption that the connected word string is recognized by finding the optimal sequence of word models that best matches the spoken string. Hence we assume that the word is the basic recognition unit for these systems, and therefore the training problem is one of estimating the optimal parameters of word models on the basis of training data, which need not contain isolated versions of the words. We describe three "optimal" approaches to solving the recognition part of connected word-recognition problems: (1) the two-level dynamic programming method, (2) the level building method, and (3) the time synchronous level building (or the one-pass) method and discuss the properties, and the relative strengths and weaknesses of each method. We then show how we can optimally train connected word systems, even if isolated versions of the vocabulary words are not available. We conclude the chapter with a discussion of a connected digit recognizer implemented using the methods described in the chapter.

Chapter 8—Large Vocabulary Continuous Speech Recognition. In this chapter we discuss the issues in applying speech-recognition technology to the problem of recognizing fluently spoken speech with vocabulary sizes of 1000 or more words (with unlimited vocabularies as the ultimate goal). It is shown that a number of fundamental problems must be solved to implement such a system, including the choice of a basic subword speech unit (from which words, phrases, and sentences can be built up), an effective way of modeling the basic speech unit, a way of deriving models of the unit, a way of designing and implementing a word lexicon (which provides a mapping between words and subword units), a way of implementing task syntax (the system grammar), a way of implementing the overall recognition part of the system (via some type of network search), and a way of imposing task semantics onto the solution. We concentrate primarily on the issues involved in building large vocabulary recognition systems. For illustrative purposes we describe one reasonable way of building such a system and discuss the resulting performance on a standard database management task.

Chapter 9—Task Oriented Applications of Automatic Speech Recognition. The final chapter of the book provides a brief overview of how one might apply the ideas discussed in the book to building a real, task-oriented, speech recognition system. It includes discussions of how one would evaluate recognizer performance and how one might decide whether a proposed task is viable for speech recognition. We also discuss a set of broad classes of applications, which appear to be the most promising ones at this time, along with typical examples of how recognizers have been successfully employed within these broad classes. The chapter concludes with some broad performance projections through the year 2000.

1.4 A BRIEF HISTORY OF SPEECH-RECOGNITION RESEARCH

Research in automatic speech recognition by machine has been done for almost four decades. To gain an appreciation for the amount of progress achieved over this period, it is worthwhile to briefly review some research highlights. The reader is cautioned that such a review is cursory, at best, and must therefore suffer from errors of judgment as well as omission.

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [1]. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words [2]. The system again relied on spectral measurements (as provided by an analog filter bank) primarily during vowel regions. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [3]. They used a spectrum analyzer and a pattern matcher to make the recognition decision. A novel aspect of this research was the use of statistical information about allowable sequences of phonemes in English (a rudimentary form of language syntax) to improve overall phoneme accuracy for words consisting of two or more phonemes. Another effort of note in this period was the vowel recognizer of Forgie and Forgie, constructed at MIT Lincoln Laboratories in 1959, in which 10 vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker-independent manner [4]. Again a filter bank analyzer was used to provide spectral information, and a time varying estimate of the vocal tract resonances was made to decide which vowel was spoken.

In the 1960s several fundamental ideas in speech recognition surfaced and were published. However, the decade started with several Japanese laboratories entering the recognition arena and building special-purpose hardware as part of their systems. One early Japanese system, described by Suzuki and Nakata of the Radio Research Lab in Tokyo [5], was a hardware vowel recognizer. An elaborate filter bank spectrum analyzer was used along with logic that connected the outputs of each channel of the spectrum analyzer (in a weighted manner) to a vowel-decision circuit, and a majority decision logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita of Kyoto University in 1962, who built a hardware phoneme recognizer [6]. A hardware speech segmenter was used along with a zero-crossing analysis of different regions of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963 [7]. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program.

In the 1960s three key research projects were initiated that have had major implications on the research and development of speech recognition for the past 20 years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems associated with nonuniformity of time scales in speech events. Martin developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduced the variability of the recognition scores [8]. Martin ultimately developed the method and founded one of the first companies, Threshold Technology, which built, marketed, and sold speech-recognition products. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances [9]. Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected word recognition, were embodied in Vintsyuk's work, it was largely unknown in the West and did not come to light until the early 1980s; this was long after the more formal methods were proposed and implemented by others.

A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes [10]. Reddy's research eventually spawned a long and highly successful speech-recognition research program at Carnegie Mellon University (to which Reddy moved in the late 1960s) which, to this day, remains a world leader in continuous-speech-recognition systems.

In the 1970s speech-recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in Russia [11], Sakoe and Chiba in Japan [12], and Itakura in the United States [13]. The Russian studies helped advance the use of pattern-recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura's research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low-bit-rate speech coding, could be extended to speech-recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.

Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which researchers studied three distinct tasks over a period of almost two decades, namely the New Raleigh language [14] for simple database queries, the laser patent text language [15] for transcribing laser patents, and the office correspondence task, called Tangora [16], for dictation of simple memos.

Finally, at AT&T Bell Labs, researchers began a series of experiments aimed at making speech-recognition systems that were truly speaker independent [17]. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker-independent patterns are now well understood and widely used.

Just as isolated word recognition was a key focus of research in the 1970s, the problem of connected word recognition was a focus of research in the 1980s. Here the goal was to create a robust system capable of recognizing a fluently spoken string of words (e.g., digits) based on matching a concatenated pattern of individual words. A wide variety of connected word-recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) [18], the one-pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in England [19], the level building approach of Myers and Rabiner at Bell Labs [20], and the frame synchronous level building approach of Lee and Rabiner at Bell Labs [21]. Each of these "optimal" matching procedures had its own implementational advantages, which were exploited for a wide range of tasks.

Speech research in the 1980s was characterized by a shift in technology from template-based approaches to statistical modeling methods—especially the hidden Markov model approach [22, 23]. Although the methodology of hidden Markov modeling (HMM) was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analyses (IDA), and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs, in the mid-1980s, that the technique became widely applied in virtually every speech-recognition research laboratory in the world.

Another "new" technology that was reintroduced in the late 1980s was the idea of applying neural networks to problems in speech recognition. Neural networks were first introduced in the 1950s, but they did not prove useful initially because they had many practical problems. In the 1980s, however, a deeper understanding of the strengths and limitations of the technology was obtained, as well as the relationships of the technology to classical signal classification methods. Several new ways of implementing systems were also proposed [24, 25].

Finally, the 1980s was a decade in which a major impetus was given to large vocabulary, continuous-speech-recognition systems by the Defense Advanced Research Projects Agency (DARPA) community, which sponsored a large research program aimed at achieving high word accuracy for a 1000-word, continuous-speech-recognition, database management task. Major research contributions resulted from efforts at CMU (notably the well-known SPHINX system) [26], BBN with the BYBLOS system [27], Lincoln Labs [28], SRI [29], MIT [30], and AT&T Bell Labs [31]. The DARPA program has continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer, and the task shifting to retrieval of air travel information. At the same time, speech-recognition technology has been increasingly used within telephone networks to automate as well as enhance operator services.

REFERENCES

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits," *J. Acoust. Soc. Am.*, 24 (6): 637–642, 1952.

[2] H. F. Olson and H. Belar, "Phonetic Typewriter," *J. Acoust. Soc. Am.*, 28 (6): 1072–1081, 1956.

[3] D. B. Fry, "Theoretical Aspects of Mechanical Speech Recognition"; and P. Denes, "The Design and Operation of the Mechanical Speech Recognizer at University College London," *J. British Inst. Radio Engr.*, 19: 4, 211–229, 1959.

[4] J. W. Forgie and C. D. Forgie, "Results Obtained From a Vowel Recognition Computer Program," *J. Acoust. Soc. Am.*, 31 (11): 1480–1489, 1959.

[5] J. Suzuki and K. Nakata, "Recognition of Japanese Vowels—Preliminary to the Recognition of Speech," *J. Radio Res. Lab*, 37 (8): 193–212, 1961.

[6] T. Sakai and S. Doshita, "The Phonetic Typewriter, Information Processing 1962," *Proc. IFIP Congress*, Munich, 1962.

[7] K. Nagata, Y. Kato, and S. Chiba, "Spoken Digit Recognizer for Japanese Language," *NEC Res. Develop.*, No. 6, 1963.

[8] T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[9] T. K. Vintsyuk, "Speech Discrimination by Dynamic Programming," *Kibernetika*, 4 (2): 81–88, Jan.–Feb. 1968.

[10] D. R. Reddy, "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," Tech. Report No. C549, Computer Science Dept., Stanford Univ., September 1966.

[11] V. M. Velichko and N. G. Zagoruyko, "Automatic Recognition of 200 Words," *Int. J. Man-Machine Studies*, 2: 223, June 1970.

[12] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-26 (1): 43-49, February 1978.

[13] F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23 (1): 67-72, February 1975.

[14] C. C. Tappert, N. R. Dixon, A. S. Rabinowitz, and W. D. Chapman, "Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery," Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146, 1971.

[15] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a Linguistic Statistical Decoder for the

Recognition of Continuous Speech," *IEEE Trans. Information Theory*, IT-21: 250–256, 1975.

[16] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proc. IEEE*, 73 (11): 1616–1624, 1985.

[17] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27: 336–349, August 1979.

[18] H. Sakoe, "Two Level DP Matching—A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27: 588–595, December 1979.

[19] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Inst. Acoust. Autumn Conf.*, 25–28, November 1979.

[20] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-29: 284–297, April 1981.

[21] C. H. Lee and L. R. Rabiner, "A Frame Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37 (11): 1649–1658, November 1989.

[22] J. Ferguson, Ed., *Hidden Markov Models for Speech*, IDA, Princeton, NJ, 1980.

[23] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, 77 (2): 257–286, February 1989.

[24] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Mag.*, 4 (2): 4–22, April 1987.

[25] A. Weibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37: 393–404, 1989.

[26] K. F. Lee, H. W. Hon, and D. R. Reddy, "An Overview of the SPHINX Speech Recognition System," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 38: 600–610, 1990.

[27] Y. L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, S. Roucos, and R. M. Schwartz, "BBYLOS: The BBN Continuous Speech Recognition System," *Proc. ICASSP 87*, 89–92, April 1987.

[28] D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proc. ICASSP 89*, Glasgow, Scotland, 449–452, May 1989.

[29] M. Weintraub et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. ICASSP 89*, Glasgow, Scotland, 699–702, May 1989.

[30] V. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT Summit Speech Recognition System: A Progress Report," *Proc. DARPA Speech and Natural Language Workshop*, 179–189, February 1989.

[31] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, 4: 127–165, 1990.

Chapter 2

THE SPEECH SIGNAL: PRODUCTION, PERCEPTION, AND ACOUSTIC-PHONETIC CHARACTERIZATION

2.1 INTRODUCTION

In this chapter we discuss the mechanics of producing and perceiving speech in human beings, and we show how an understanding of these processes leads naturally to several different approaches to speech recognition by machine. We begin by showing how the different classes of speech sounds, or phonetics, can each be characterized in terms of broad acoustic features whose properties are relatively invariant across words and speakers. The ideas of acoustic-phonetic characterization of sounds lead naturally to straightforward implementation of a speech-recognition algorithm based on sequential detection of sounds and sound classes. The strengths and weaknesses of such an approach are discussed. An alternative approach to speech recognition is to use standard pattern-recognition techniques in a framework in which all speech knowledge is "learned" via a training phase. We show that such a "blind" approach has some natural advantages for a wide range of speech-recognition systems. Finally we show how aspects of both the acoustic-phonetic approach and the pattern-recognition approach can be integrated into a hybrid method that includes techniques from artificial intelligence as well as neural network methods.

2.1.1 The Process of Speech Production and Perception in Human Beings

Figure 2.1 shows a schematic diagram of the speech-production/speech-perception process in human beings. The production (speech-generation) process begins when the talker

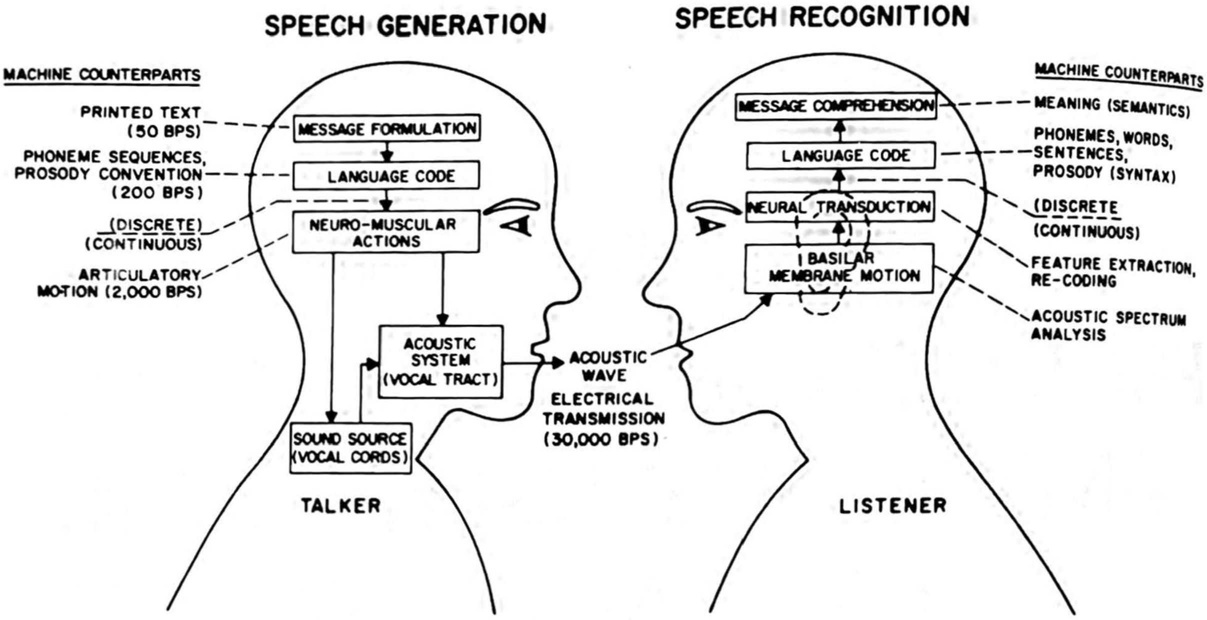


Figure 2.1 Schematic diagram of speech-production/speech-perception process (after Flanagan [unpublished]).

formulates a message (in his mind) that he wants to transmit to the listener via speech. The machine counterpart to the process of message formulation is the creation of printed text expressing the words of the message. The next step in the process is the conversion of the message into a language code. This roughly corresponds to converting the printed text of the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody markers denoting duration of sounds, loudness of sounds, and pitch accent associated with the sounds. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output. The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the lips, jaw, tongue, and velum (a "trapdoor" controlling the acoustic flow to the nasal mechanism).

Once the speech signal is generated and propagated to the listener, the speech-perception (or speech-recognition) process begins. First the listener processes the acoustic signal along the basilar membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process. In a manner that is not well understood, the neural activity along the auditory nerve is converted into a language code at the higher centers of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

A slightly different view of the speech-production/speech-perception process is shown in Figure 2.2. Here we see the steps in the process laid out along a line corresponding to the basic information rate of the signal (or control) at various stages of the

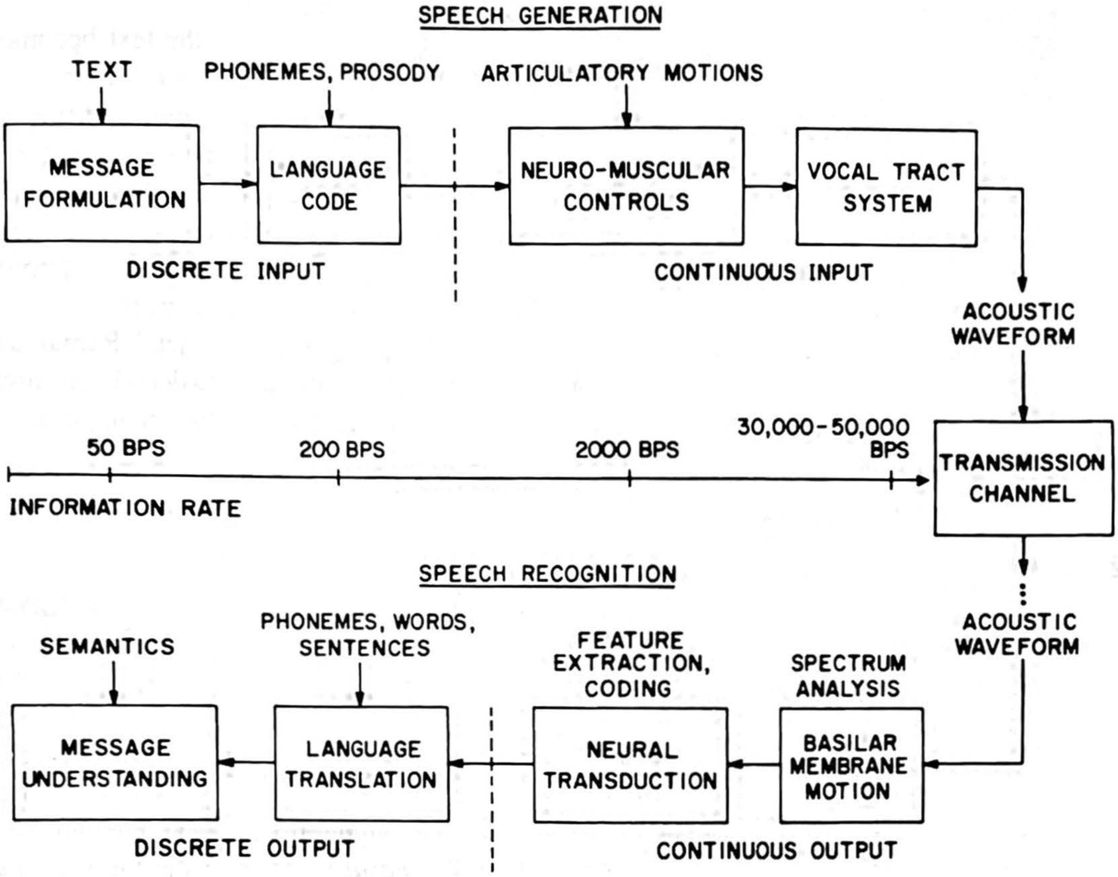


Figure 2.2 Alternative view of speech-production/speech-perception process (after Rabiner and Levinson [1]).

process. The discrete symbol information rate in the raw message text is rather low (about 50 bps [bits per second] corresponding to about 8 sounds per second, where each sound is one of about 50 distinct symbols). After the language code conversion, with the inclusion of prosody information, the information rate rises to about 200 bps. Somewhere in the next stage the representation of the information in the signal (or the control) becomes continuous with an equivalent rate of about 2000 bps at the neuromuscular control level, and about 30,000–50,000 bps at the acoustic signal level.

A transmission channel is shown in Figure 2.2 [1], indicating that any of several well-known coding techniques could be used to transmit the acoustic waveform from the talker to the listener. The steps in the speech-perception mechanism can also be interpreted in terms of information rate in the signal or its control and follows the inverse pattern of the production process. Thus the continuous information rate at the basilar membrane is in the range of 30,000–50,000 bps, while at the neural transduction stage it is about 2000 bps. The higher-level processing within the brain converts the neural signals to a discrete representation, which ultimately is decoded into a low-bit-rate message.

To illustrate, in a trivial way, how the speech-production/speech-perception process works, consider that the speaker has a goal of finding out whether his office mate has eaten his lunch yet. To express this thought, the speaker formulates the message "Did you eat yet?" In the process of converting the message to a language code, the text becomes a phonetic sequence of sounds of the form /dId yu it yɛt?/, in which each word is expressed as a sequence of phonemes constituting the ideal pronunciation of the sounds of the word (as spoken in isolation) within the spoken language. However, because the words are not spoken in isolation, and a physical mechanism is used to produce the sounds (the human vocal tract system), and because physical systems obey continuity and smoothness constraints, by the time the message is spoken the sounds become more like the phonetic string /dI jə it jɛt?/. The final d in dId is dropped, the word *you* becomes converted to a word that sounds a lot like "juh," and finally the word *yet* is pronounced as "jet." Remarkably, through the speech-perception process, human beings are usually able to decode this highly stylized version of the text into the correct string; sadly, however, this remains a most difficult task for almost all speech-recognition machines.

2.2 THE SPEECH-PRODUCTION PROCESS

Figure 2.3 shows a mid-sagittal plane (longitudinal cross-section) X-ray of the human vocal apparatus [2]. The *vocal tract*, outlined by the dotted lines in Figure 2.3, begins at the opening of the vocal cords, or *glottis*, and ends at the lips. The vocal tract consists of the *pharynx* (the connection from the esophagus to the mouth) and the mouth, or *oral cavity*. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract, determined by the positions of the tongue, lips, jaw, and velum, varies from zero (complete closure) to about 20 cm^2. The *nasal tract* begins at the velum and ends at the nostrils. When the *velum*, (a trapdoor-like mechanism at the back of the mouth cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

A schematic diagram of the human vocal mechanism is shown in Figure 2.4 [3]. Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the *trachea* (or windpipe), the tensed vocal cords within the *larynx* are caused to vibrate (in the mode of a relaxation oscillator) by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the *pharynx* (the throat cavity), the mouth cavity, and possibly the nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, mouth), different sounds are produced.

Figure 2.5 shows plots of the glottal air flow (volume velocity waveform) and the resulting sound pressure at the mouth for a typical vowel sound [4]. The glottal waveform shows a gradual build-up to a quasi-periodic pulse train of air, taking about 15 msec to reach steady state. This build-up is also reflected in the acoustic waveform shown at the bottom of the figure.

A simplified representation of the complete physiological mechanism for creating speech is shown in Figure 2.6 [3]. The lungs and the associated muscles act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the bronchi and trachea. When the vocal cords are tensed, the air flow causes them to vibrate, producing

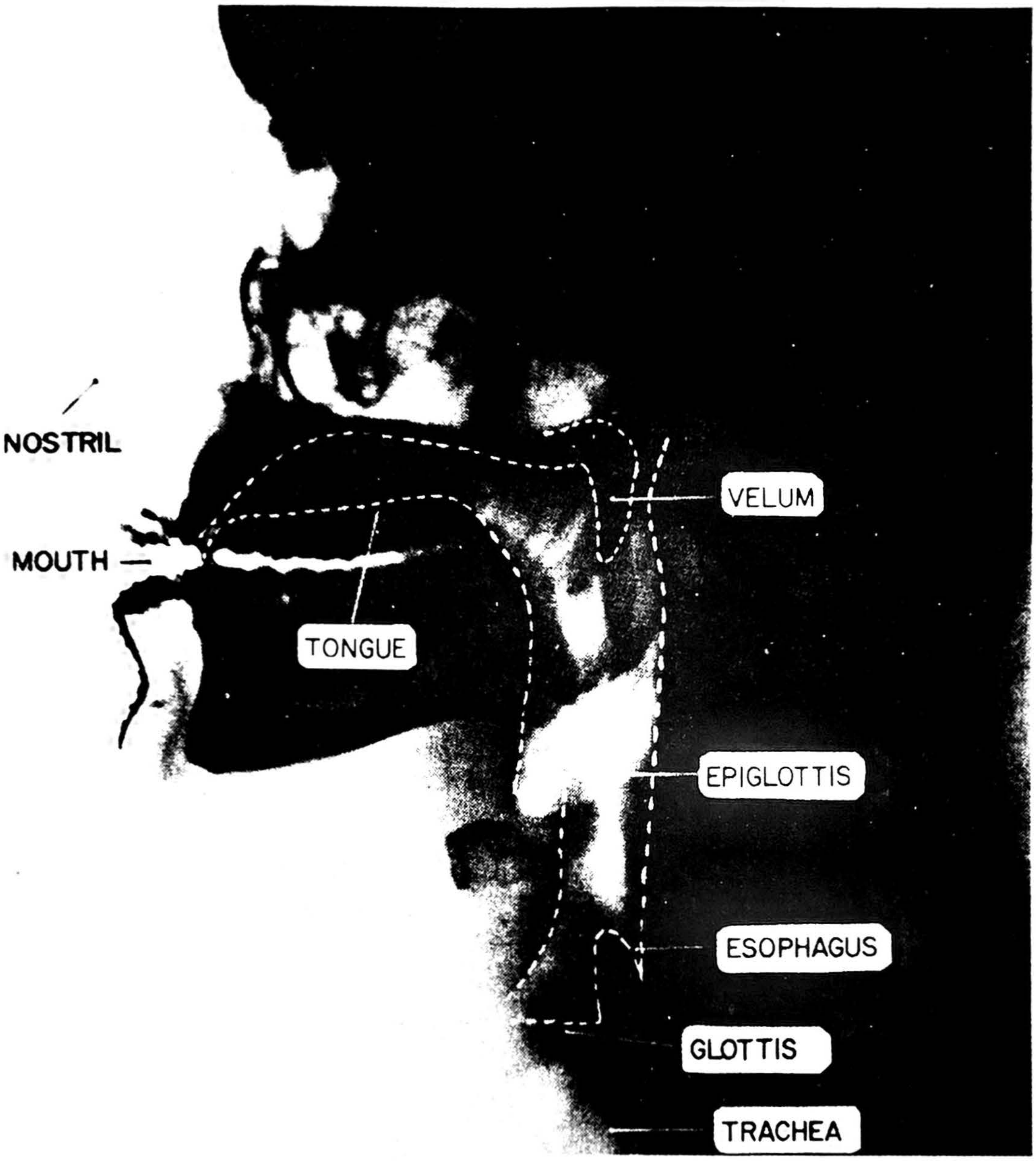


Figure 2.3 Mid-sagittal plane X-ray of the human vocal apparatus (after Flanagan et al. [2]).

so-called voiced speech sounds. When the vocal cords are relaxed, in order to produce a sound, the air flow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sounds, or it can build up pressure behind a point of total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly released, causing a brief transient sound.

Speech is produced as a sequence of sounds. Hence the state of the vocal cords, as well as the positions, shapes, and sizes of the various articulators, changes over time to reflect the sound being produced. The manner in which different sounds are created will be described later in this chapter. First we divert to a brief discussion of the speech waveform

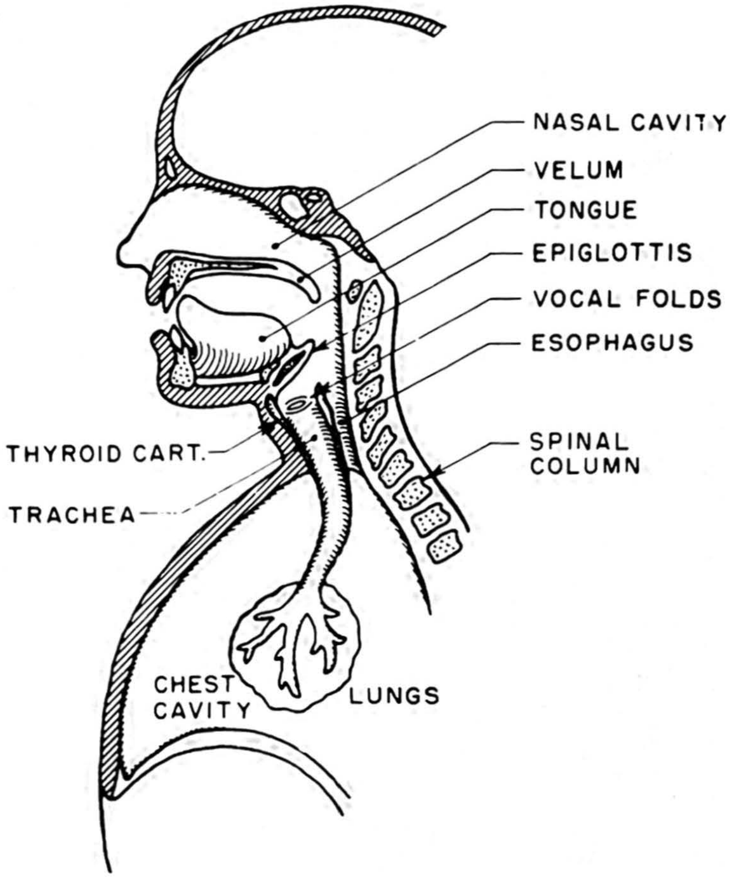


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).

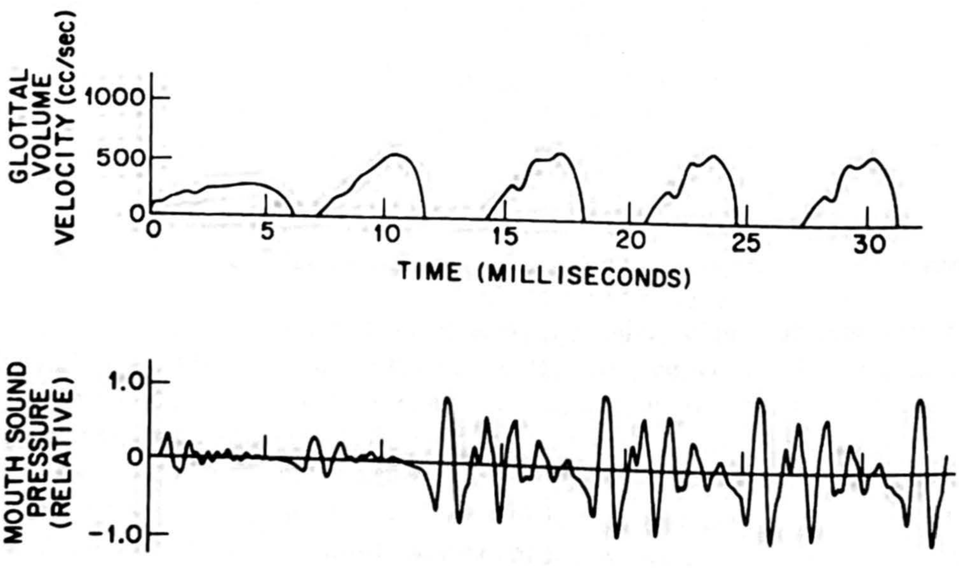


Figure 2.5 Glottal volume velocity and resulting sound pressure at the start of a voiced sound (after Ishizaka and Flanagan [4]).

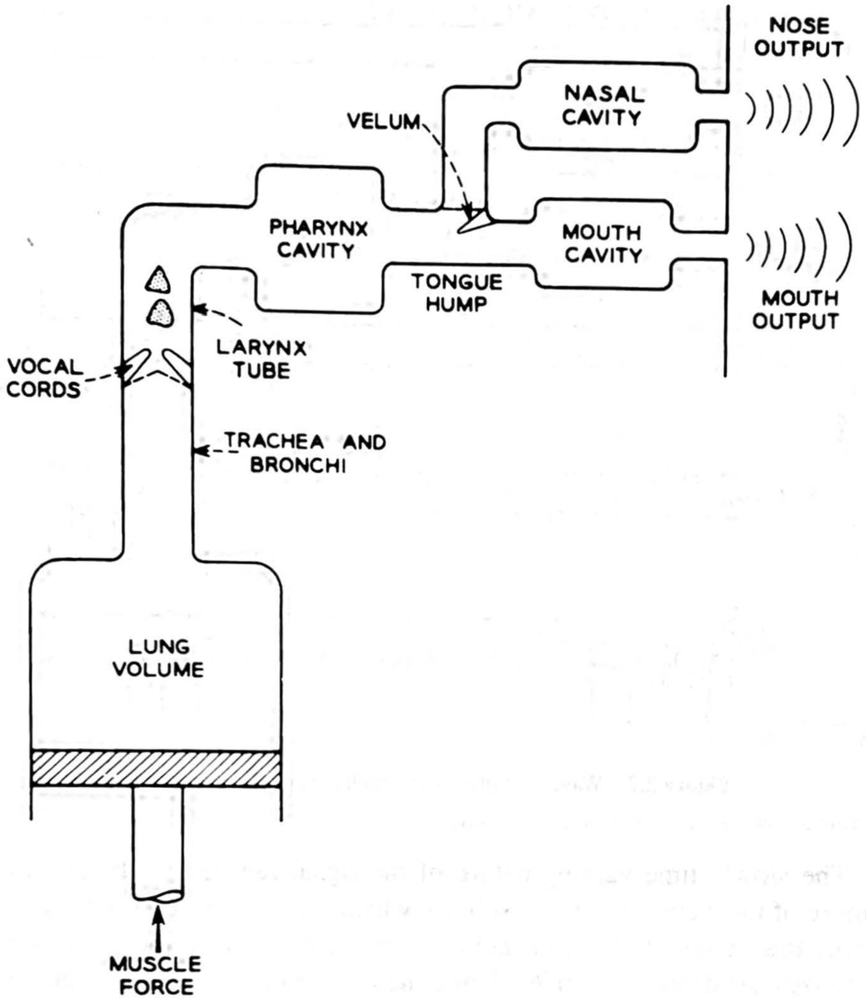


Figure 2.6 Schematic representation of the complete physiological mechanism of speech production (after Flanagan [3]).

and its spectral representation.

2.3 REPRESENTING SPEECH IN THE TIME AND FREQUENCY DOMAINS

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. An illustration of this effect is given in Figure 2.7, which shows the time waveform corresponding to the initial sounds in the phrase, "It's time . . ." as spoken by a male speaker. Each line of the waveform corresponds to 100 msec (1/10 second) of signal; hence the entire plot encompasses about 0.5 sec.

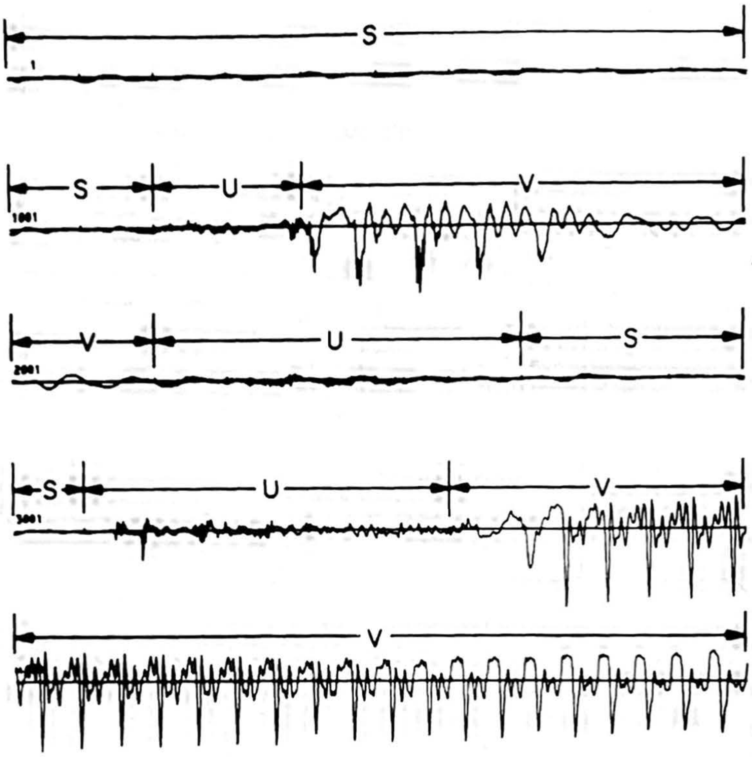


Figure 2.7 Waveform plot of the beginning of the utterance "It's time."

The slowly time varying nature of the signal can be seen by contrasting the first 100 msec of the waveform (the first line), which corresponds to background silence and is therefore low in amplitude, to the next 100 msec of the waveform (the second line), which first shows a small increase in level, and then a sharp increase in level and a gross change in waveform shape and regularity (it becomes almost periodic).

There are several ways of classifying (labeling) events in speech. Perhaps the simplest and most straightforward is via the state of the speech-production source—the vocal cords. It is accepted convention to use a three-state representation in which the states are (1) silence (S), where no speech is produced; (2) unvoiced (U), in which the vocal cords are not vibrating, so the resulting speech waveform is aperiodic or random in nature; and (3) voiced (V), in which the vocal cords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting speech waveform is quasi-periodic. The result of applying this type of classification to the waveform of Figure 2.7 is shown in the figure. Initially, before speaking begins, the waveform is classified as silence (S). A brief period of unvoiced (U) sound (whisper or aspiration) is seen prior to the voicing (V) corresponding to the initial vowel in the word *It's*. Following the voicing region, there is a brief, unvoiced aspiration (devoicing of the vowel), followed by a silence region (prior to the /t/ in *It's*), and then a relatively long, unvoiced (U) region corresponding to the /t/ release, followed by the /s/, followed by the /t/ in *time*. Finally there is a long voicing (V) region corresponding to the diphthong $/a^y/$ in *time*.

It should be clear that the segmentation of the waveform into well-defined regions of

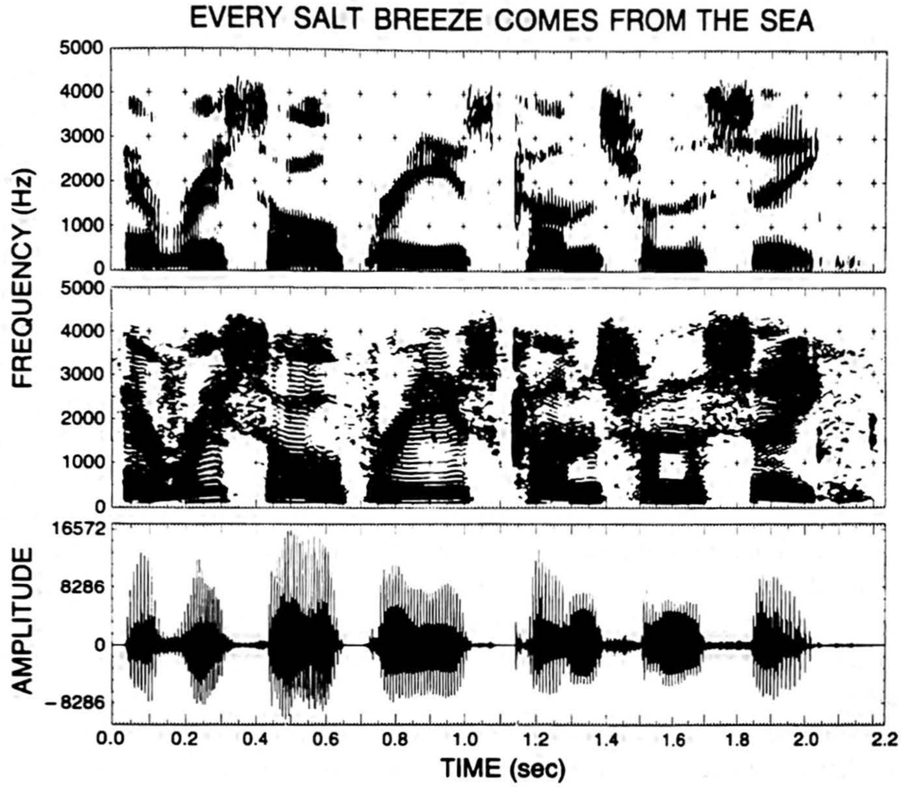


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea."

silence, unvoiced, and voiced signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or a weak, voiced sound (like /v/ or /m/) from unvoiced sounds or even silence. However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications.

An alternative way of characterizing the speech signal and representing the information associated with the sounds is via a spectral representation. Perhaps the most popular representation of this type is the sound spectrogram in which a three-dimensional representation of the speech intensity, in different frequency bands, over time is portrayed. An example of this type of speech representation is given in Figure 2.8, which shows a *wideband spectrogram* in the first panel, a *narrowband spectrogram* in the second panel, and a waveform amplitude plot in the third panel, of a spoken version of the utterance "Every salt breeze comes from the sea" by a male speaker. The wideband spectrogram corresponds to performing a spectral analysis on 15-msec sections of waveform using a broad analysis filter (125 Hz bandwidth) with the analysis advancing in intervals of 1 msec. The spectral intensity at each point in time is indicated by the intensity (darkness) of the plot at a particular analysis frequency. Because of the relatively broad bandwidth of the analysis filters, hence the relatively short duration of the analysis window, the spectral envelope of individual periods of the speech waveform during voiced sections are resolved and are seen as vertical striations in the spectrogram.

The narrowband spectrogram (shown in the second panel of Figure 2.8) corresponds to performing a spectral analysis on 50-msec sections of waveform using a narrow analysis filter (40 Hz bandwidth), with the analysis again advancing in intervals of 1 msec. Because of the relatively narrow bandwidth of the analysis filters, individual spectral harmonics corresponding to the pitch of the speech waveform, during voiced regions, are resolved and are seen as almost-horizontal lines in the spectrogram. During periods of unvoiced speech, we see primarily high-frequency energy in the spectrograms; during periods of silence we essentially see no spectral activity (because of the reduced signal level).

A third way of representing the time-varying signal characteristics of speech is via a parameterization of the spectral activity based on the model of speech production. Because the human vocal tract is essentially a tube, or concatenation of tubes, of varying cross-sectional area that is excited either at one end (by the vocal cord puffs of air) or at a point along the tube (corresponding to turbulent air at a constriction), acoustic theory tells us that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube. Such resonances are called *formants* for speech, and they represent the frequencies that pass the most acoustic energy from the source to the output. Typically there are about three resonances of significance, for a human vocal tract, below about 3500 Hz. Figure 2.9 [5] shows a wideband spectrogram, along with the computed formant frequency estimates, for the utterance "Why do I owe you a letter," spoken by a male speaker. There is a good correspondence between the estimated formant frequencies and the points of high spectral energy in the spectrogram. The formant frequency representation is a highly efficient, compact representation of the time-varying characteristics of speech. The major problem, however, is the difficulty of reliably estimating the formant frequencies for low-level voiced sounds, and the difficulty of defining the formants for unvoiced or silence regions. As such, this representation is more of theoretical than of practical interest.

Figures 2.10 and 2.11 show spectral and temporal representations of the phrase "Should we chase," spoken by a male speaker, along with a detailed segmentation of the waveform into individual sounds. The ultimate goal of speech recognition is to uniquely and automatically provide such a segmentation and labeling of speech into constituent sounds or sound groups such as words, then sentences. To understand the limitations on this approach, we will next discuss, in detail, the general sounds of English and the relevant acoustic and phonetic features of the sounds.

2.4 SPEECH SOUNDS AND FEATURES

The number of linguistically distinct speech sounds (phonemes) in a language is often a matter of judgment and is not invariant to different linguists. Table 2.1 shows a condensed list of phonetic symbols of American English, their ARPABET representation [6], and an example word in which the sound occurs. Shown in this table are 48 sounds, including 18 vowels or vowel combinations (called diphthongs), 4 vowel-like consonants, 21 standard consonants, 4 syllabic sounds, and a phoneme referred to as a glottal stop (literally a symbol

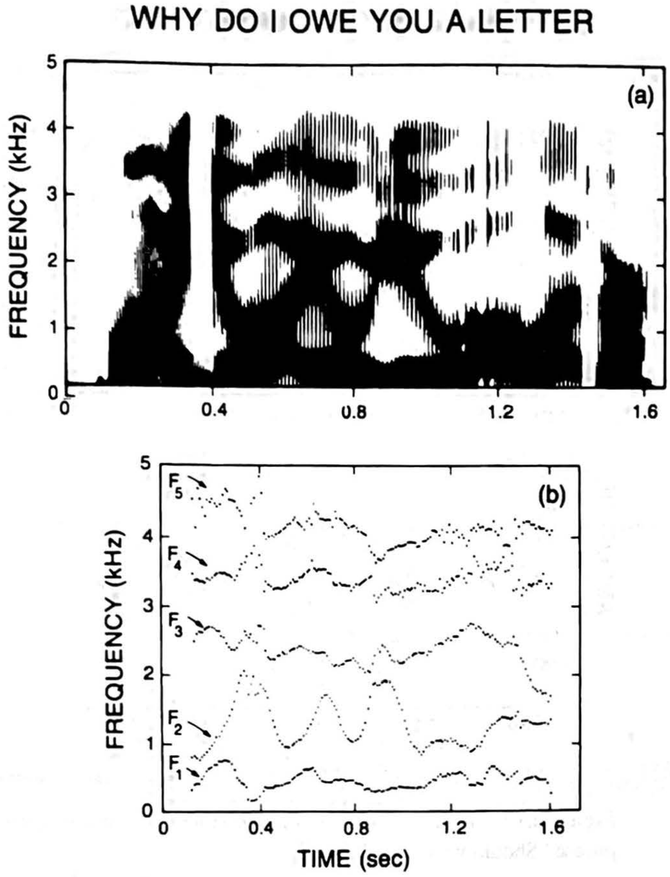


Figure 2.9 Wideband spectrogram and formant frequency representation of the utterance "Why do I owe you a letter" (after Atal and Hanauer [5]).

for a sound corresponding to a break in voicing within a sound).

Many of the sounds or phonemes shown in Table 2.1 are not considered standard; they represent specialized cases such as the so-called barred I (/ɨ/) in the word *roses*. As such, a more standard representation of the basic sounds and sound classes of American English is shown in Figure 2.12. Here we see the conventional set of 11 vowels, classified as front, mid, or back, corresponding to the position of the tongue hump in producing the vowel; 4 vowel combinations or diphthongs; the 4 semivowels broken down into 2 liquids and 2 glides; the nasal consonants, the voiced and unvoiced stop consonants; the voiced and unvoiced fricatives; whisper; and the affricates. There are a total of 39 of the 48 sounds of Table 2.1 represented in Figure 2.12.

2.4.1 The Vowels

The vowel sounds are perhaps the most interesting class of sounds in English. Their importance to the classification and representation of written text is very low; however, most practical speech-recognition systems rely heavily on vowel recognition to achieve

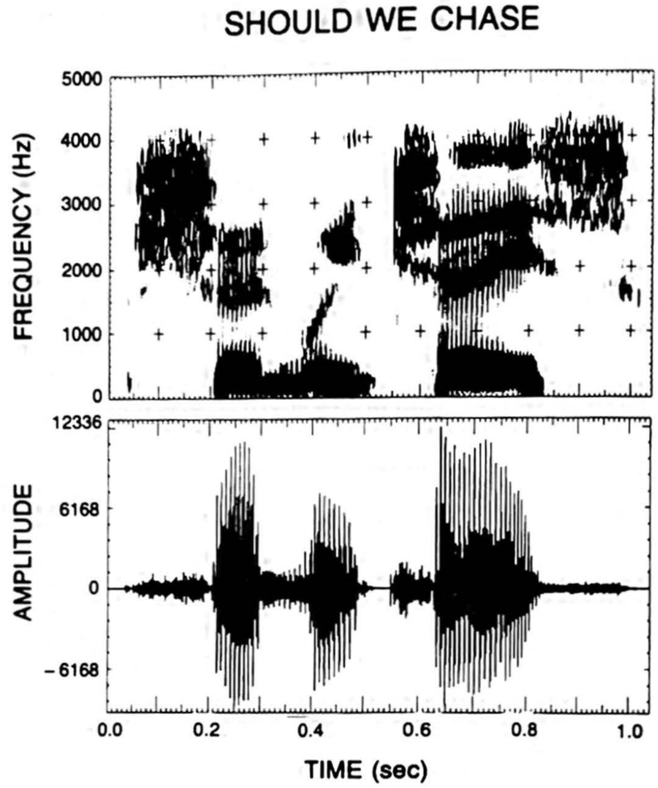


Figure 2.10 Wideband spectrogram and intensity contour of the phrase "Should we chase."

high performance. To partially illustrate this point, consider the following sections of text:

Section I

> Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nts i_ th_ c_mp_ny's _m_g_, s_p_rv_s__n, th__r w_rk_ng c_nd_t__ns, b_n_f_ts _nd _pp_rt_n_t__s f_r gr_wth.

Section II

> A__i_u_e_ _o_a___a___a_e_ e__e__ia___ __e _a_e, _i__ __e __o_e_ o_ o__u_a_io_a_ e___o_ee_ __i_____ _e__ea_i__.

In Section I we have omitted the conventional vowel letters (a,e,i,o,u); however, with a little effort the average reader can "fill in" the missing vowels and decode the section so that it reads

> They noted significant improvements in the company's image, supervision, their working conditions, benefits and opportunities for growth.

In Section II we have omitted the conventional consonant letters; the resulting text is essentially not decodable. The actual text is

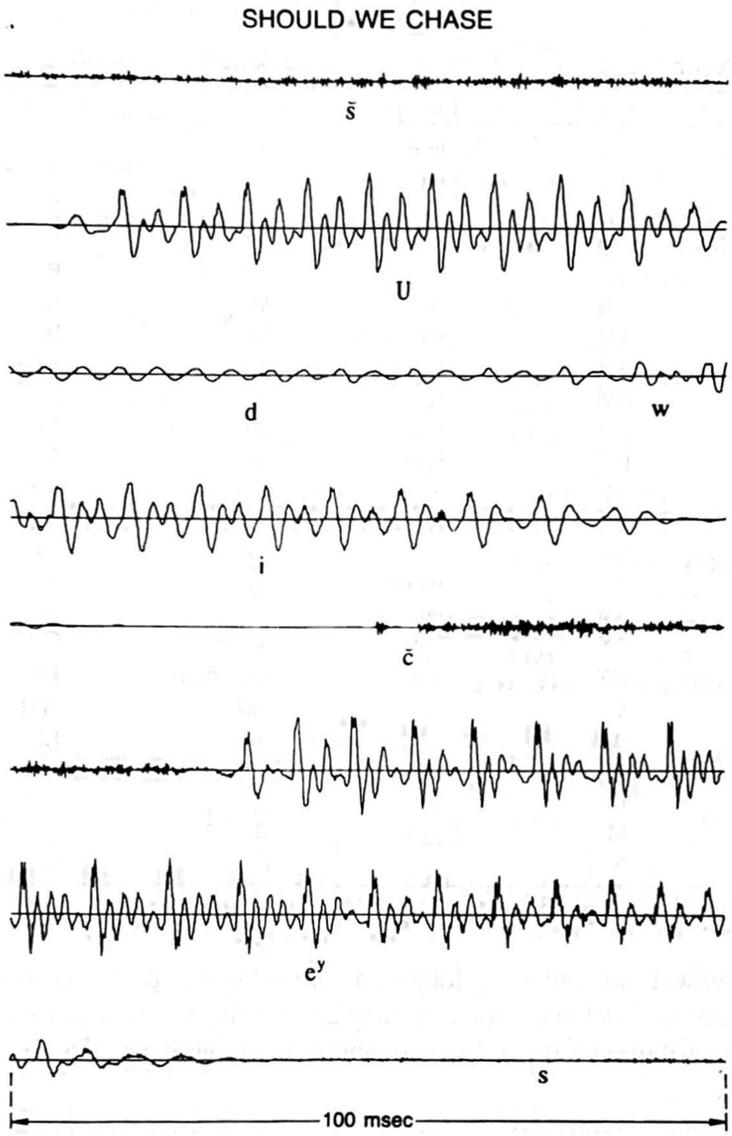


Figure 2.11 The speech waveform and a segmentation and labeling of the constituent sounds of the phrase "Should we chase."

Attitudes toward pay stayed essentially the same, with the scores of occupational employees slightly decreasing.

In speaking, vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. The way in which the cross-sectional area varies along the vocal tract determines the resonance frequencies of the tract (the formants) and thereby the sound that is produced. The vowel sound produced is determined primarily by the position of the tongue, but the positions of the jaw, lips, and to a small extent, the velum, also influence the resulting sound.

TABLE 2.1. A condensed list of phonetic symbols for American English.

Phoneme	ARPABET	Example	Phoneme	ARPABET	Example
/i/	IY	beat	/ŋ/	NX	sing
/I/	IH	bit	/p/	P	pet
/e/ (eʸ)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/a/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	H	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (oʷ)	OW	boat	/f/	F	fat
/U/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/š/ (sh)	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/aʷ/	AW	down	/ž/ (zh)	ZH	azure
/aʸ/	AY	buy	/č/ (tsh)	CH	church
/ɔʸ/	OY	boy	/ǰ/ (dzh, j)	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/l̩/	EL	battle
/r/	R	rent	/m̩/	EM	bottom
/l/	L	let	/n̩/	EN	button
/m/	M	met	/ɾ/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

The vowels are generally long in duration (as compared to consonant sounds) and are spectrally well defined. As such they are usually easily and reliably recognized and therefore contribute significantly to our ability to recognize speech, both by human beings and by machine.

There are several ways to characterize and classify vowels, including the typical articulatory configuration required to produce the sounds, typical waveform plots, and typical spectrogram plots. Figures 2.13–2.15 show typical articulatory configurations of the vowels (2.13), examples of vowel waveforms (2.14), and examples of vowel spectrograms (2.15). A convenient and simplified way of classifying vowel articulatory configurations is in terms of the tongue hump position (i.e., front, mid, back), and tongue hump height (high, mid, low), where the tongue hump is the mass of the tongue at its narrowest constriction within the vocal tract. According to this classification the vowels /i/, /I/, /æ/, and /ɛ/ are front vowels, (with different tongue heights) /a/, /ʌ/, and /ɔ/ are mid vowels, and /U/, /u/, and /o/ are back vowels (see also Figure 2.12).

As shown in the acoustic waveform plots of the vowels, in Figure 2.14, the front vowels show a pronounced, high-frequency resonance, the mid vowels show a balance of energy over a broad frequency range, and the back vowels show a predominance of low-frequency spectral information. This behavior is evidenced in the vowel spectrogram

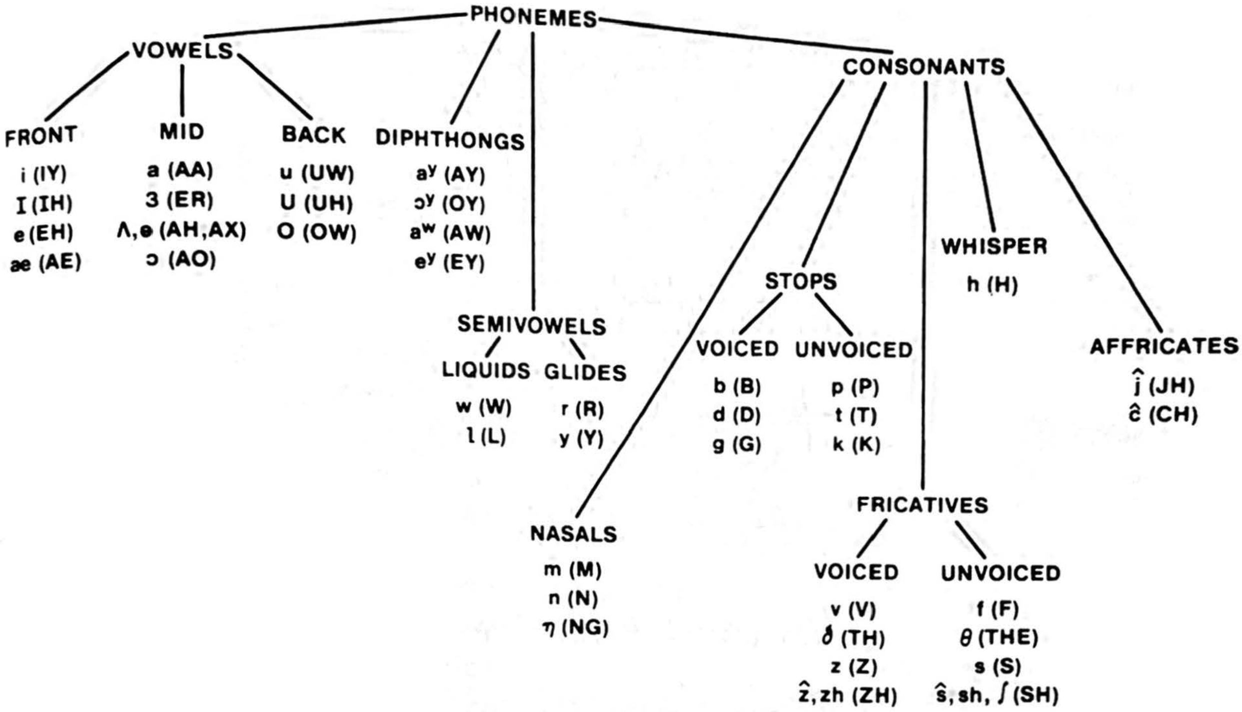


Figure 2.12 Chart of the classification of the standard phonemes of American English into broad sound classes.

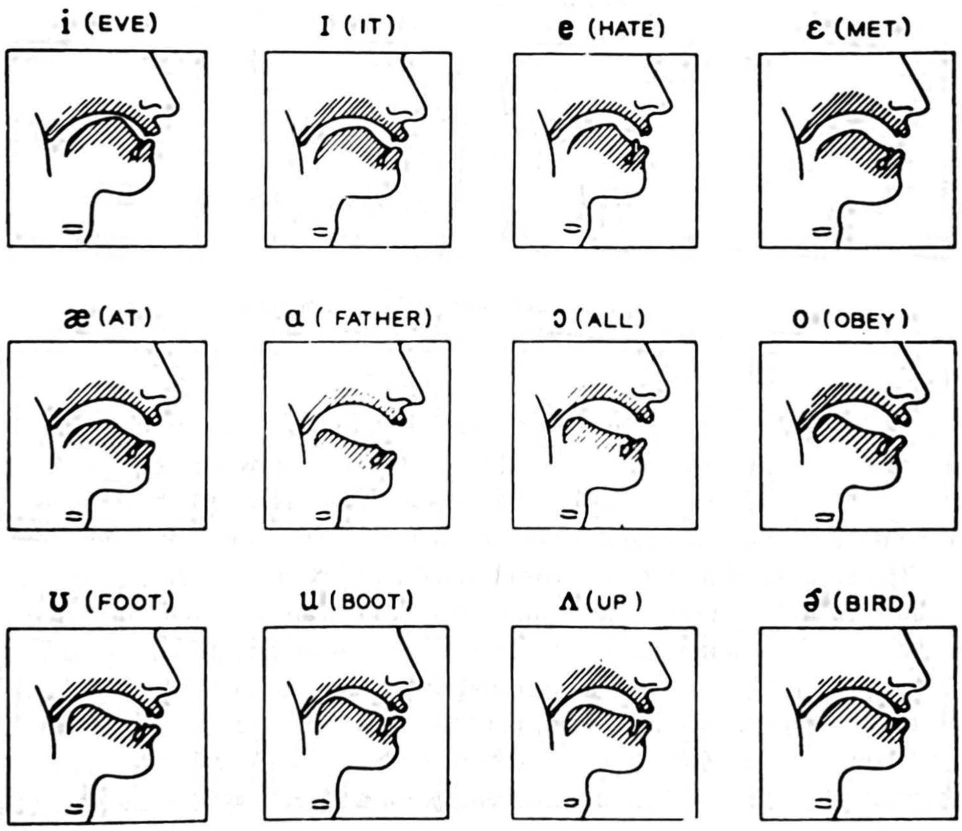


Figure 2.13 Articulatory configurations for typical vowel sounds (after Flanagan [3]).

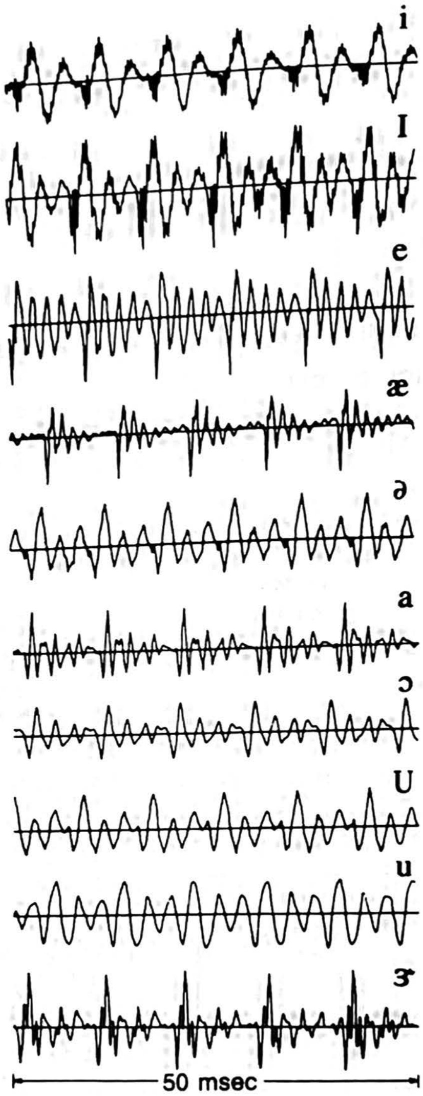


Figure 2.14 Acoustic waveform plots of typical vowel sounds.

plots of Figure 2.15, in which the front vowels show a relatively high second and third formant frequency (resonance), whereas the mid vowels show well-separated and balanced locations of the formants, and the back vowels (especially /u/) show almost no energy beyond the low-frequency region with low first and second formant frequencies.

The concept of a "typical" vowel sound is, of course, unreasonable in light of the variability of vowel pronunciation among men, women and children with different regional accents and other variable characteristics. To illustrate this point, Figure 2.16 shows a classic plot, made by Gordon Peterson and Harold Barney, of measured values of the first and second formant for 10 vowels spoken by a wide range of male and female talkers who attended the 1939 World's Fair in New York City [7].A wide range of variability can be seen in the measured formant frequencies for a given vowel sound, and also there is