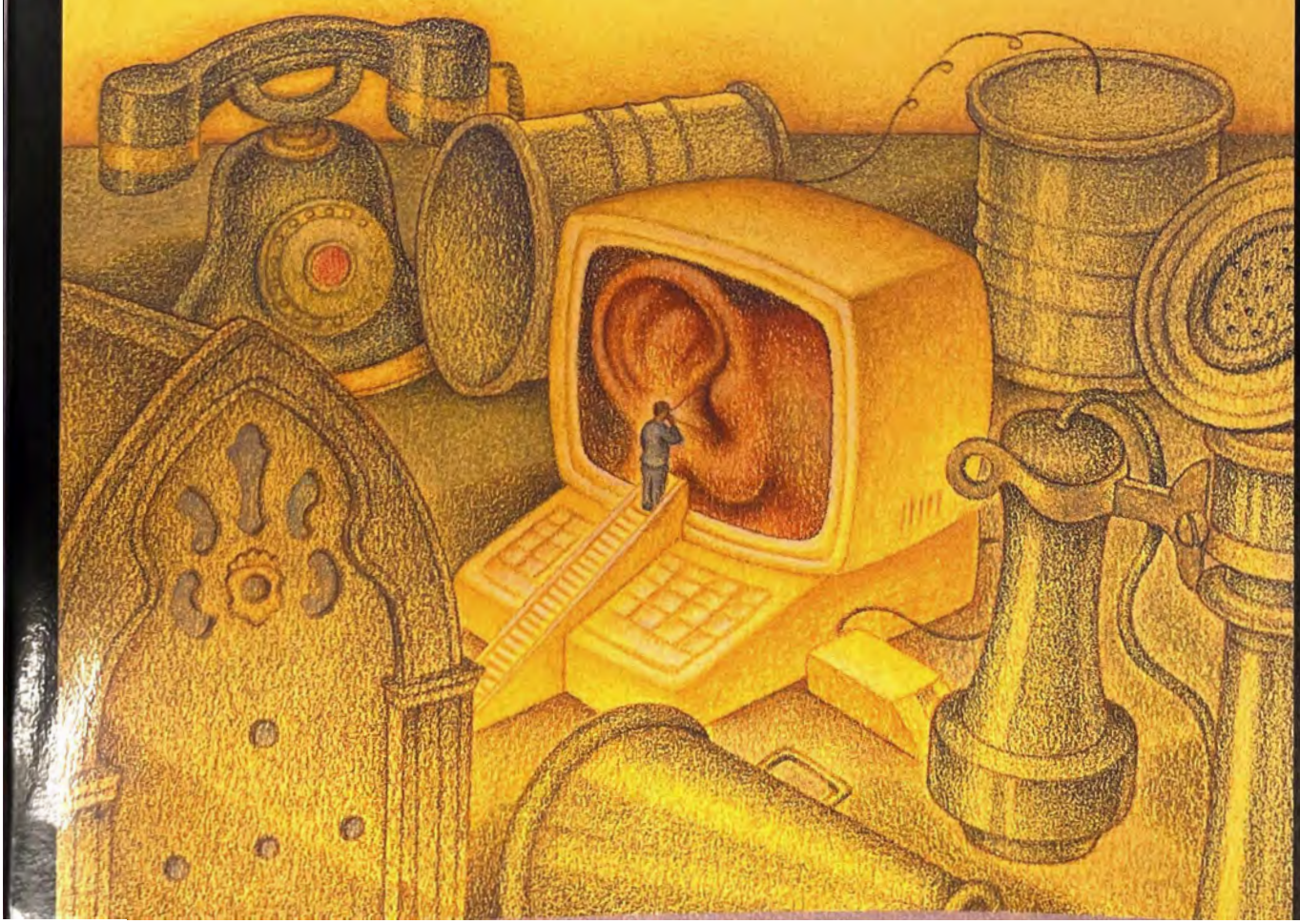
Christopher Schmandt
Foreword by Nicholas Negroponte
MIT Media Lab

Voice Communication with Computers

Conversational Systems



VNR Computer Library

IPR2023-00033
Apple EX1012 Page 1

Copyright © 1994 by Christopher Schmandt

Library of Congress Catalog Card Number 93-36404
ISBN 0-442-23935-1

All rights reserved. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without the written permission of the publisher.

I(T)P Van Nostrand Reinhold is an International Thomson Publishing company. ITP logo is a trademark under license.

Printed in

Van Nostrand Reinhold
115 Fifth Avenue
New York, NY 10003

International Thomson Publishing GmbH
Königswinterer Str. 418
53277 Bonn
Germany

International Thomson Publishing
Berkshire House, 168-173
High Holborn, London WC1V 7AA
England

International Thomson Publishing Asia
221 Henderson Building #05-10
Singapore 0315

Thomas Nelson Australia
102 Dodds Street
South Melbourne 3205
Victoria, Australia

International Thomson Publishing Japan
Kyowa Building, 3F
2-2-1 Hirakawacho
Chiyoda-Ku, Tokyo 102
Japan

Nelson Canada
1120 Birchmount Road
Scarborough, Ontario
M1K 5G4, Canada

16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data 93-36404

Schmandt, Chris.
Voice communication with computers / Chris Schmandt.
p. cm.
Includes bibliographical references and index.
ISBN 0-442-23935-1
1. Interactive computer systems. 2. Natural language processing (Computer science) I. Title.
QA76.9.I58S35 1993
006.4'54—dc20

93-36404
CIP

IPR2023-00033
Apple EX1012 Page 2

Contents

Speaking of Talk xvii
Preface xvii
Acknowledgments xxi
Introduction 1

Chapter 1. Speech as Communication 5

SPEECH AS CONVERSATION 6
HIERARCHICAL STRUCTURE OF CONVERSATION 8
REPRESENTATIONS OF SPEECH 12
Acoustic Representations 12
PHONEMES AND SYLLABLES 14
Phonemes 14
Syllables 17
Other Representations 17
SUMMARY 18

Chapter 2. Speech Production and Perception 19

VOCAL TRACT 19

IPR2023-00033
Apple EX1012 Page 3

THE SPEECH SOUNDS 24
Vowels 25
Consonants 26
Liquids and Glides 28
Acoustic Features of Phonemes 28
HEARING 28
Auditory System 29
Localization of Sounds 31
Psychoacoustics 33
SUMMARY 34
FURTHER READING 35

Chapter 3. Speech Coding 36

SAMPLING AND QUANTIZATION 37
SPEECH-CODING ALGORITHMS 44
Waveform Coders 44
Source Coders 51
CODER CONSIDERATIONS 53
Intelligibility 54
Editing 54
Silence Removal 55
Time Scaling 57
Robustness 58
SUMMARY 59
FURTHER READING 59

Chapter 4. Applications and Editing of Stored Voice 60

TAXONOMY OF VOICE OUTPUT APPLICATIONS 61
Playback-Only Applications 61
Interactive Record and Playback Applications 62
Dictation 63
Voice as a Document Type 64
VOICE IN INTERACTIVE DOCUMENTS 65

IPR2023-00033
Apple EX1012 Page 4

VOICE EDITING 69
Temporal Granularity 69
Manipulation of Audio Data 70
EXAMPLES OF VOICE EDITORS 74
Intelligent Ear, M.I.T. 74
Tioga Voice, Xerox PARC 75
PX Editor, Bell Northern Research 76
Sedit, Olivetti Research Center, and M.I.T. Media Laboratory 78
Pitchtool, M.I.T. Media Laboratory 79
SUMMARY 80

Chapter 5. Speech Synthesis 82

SYNTHESIZING SPEECH FROM TEXT 84
FROM TEXT TO PHONEMES 85
Additional Factors for Pronunciation 87
FROM PHONEMES TO SOUND 91
Parametric Synthesis 91
Concatenative Synthesis 93
QUALITY OF SYNTHETIC SPEECH 94
Measuring Intelligibility 95
Listener Satisfaction 96
Performance Factors 96
APPLICATIONS OF SYNTHETIC SPEECH 97
SUMMARY 99
FURTHER READING 99

Chapter 6. Interactive Voice Response 100

LIMITATIONS OF SPEECH OUTPUT 101
Speed 101
Temporal Nature 102
Serial Nature 102
Bulkiness 102
Privacy 103
ADVANTAGES OF VOICE 104

IPR2023-00033
Apple EX1012 Page 5

DESIGN CONSIDERATIONS 105
Application Appropriateness 106
Data Appropriateness 107
Responsiveness 108
Speech Rate 108
Interruption 109
Repetition 109
Exception Pronunciation 110
Multiple Voices 111
USER INPUT WITH TOUCHTONES 112
Menus 112
Data Entry 113
CASE STUDIES 117
Direction Assistance 117
Back Seat Driver 121
Voiced Mail 124
SUMMARY 130

Chapter 7. Speech Recognition 132

BASIC RECOGNIZER COMPONENTS 132
SIMPLE RECOGNIZER 133
Representation 134
Templates 134
Pattern Matching 135
CLASSES OF RECOGNIZERS 137
Who Can Use the Recognizer? 137
Speaking Style: Connected or Isolated Words? 139
Vocabulary Size 140
ADVANCED RECOGNITION TECHNIQUES 141
Dynamic Time Warping 142
Hidden Markov Models 144
Vector Quantization 147
Employing Constraints 149

IPR2023-00033
Apple EX1012 Page 6

ADVANCED RECOGNITION SYSTEMS 151
IBM's Tangora 151
CMU's Sphinx 151
MIT's SUMMIT 152
SUMMARY 152
FURTHER READING 153

Chapter 8. Using Speech Recognition 154

USES OF VOICE INPUT 154
Sole Input Channel 154
Auxiliary Input Channel 156
Keyboard Replacement 157
SPEECH RECOGNITION ERRORS 160
Classes of Recognition Errors 160
Factors Influencing the Error Rate 161
INTERACTION TECHNIQUES 163
Minimizing Errors 164
Confirmation Strategies 165
Error Correction 167
CASE STUDIES 169
Xspeak: Window Management by Voice 170
Put That There 175
SUMMARY 178

Chapter 9. Higher Levels of Linguistic Knowledge 179

SYNTAX 180
Syntactic Structure and Grammars 180
Parsers 185
SEMANTICS 186
PRAGMATICS 189
Knowledge Representation 190
Speech Acts 192
Conversational Implicature and Speech Acts 193

IPR2023-00033
Apple EX1012 Page 7

DISCOURSE 194
Regulation of Conversation 195
Discourse Focus 197
CASE STUDIES 199
Grunt 199
Conversational Desktop 204
SUMMARY 208
FURTHER READING 209

Chapter 10. Basics of Telephones 210

FUNCTIONAL OVERVIEW 211
ANALOG TELEPHONES 212
Signaling 213
Transmission 218
DIGITAL TELEPHONES 221
Signaling 222
Transmission 224
PBXS 226
SUMMARY 228
FURTHER READING 229

Chapter 11. Telephones and Computers 230

MOTIVATION 231
Access to Multiple Communication Channels 231
Improved User Interfaces 232
Enhanced Functionality 233
Voice and Computer Access 234
PROJECTS IN INTEGRATED TELEPHONY 234
Etherphone 234
MICE 237
BerBell 239
Personal eXchange 239
Phonetool 242

IPR2023-00033
Apple EX1012 Page 8

ARCHITECTURES 244
Distributed Architectures 244
Centralized Architectures 246
Comparison of Architectures 249
CASE STUDIES 251
Phone Slave 251
Xphone and Xrolo 256
Flexible Call Routing 260
SUMMARY 267

Chapter 12. Desktop Audio 268

EFFECTIVE DEPLOYMENT OF DESKTOP AUDIO 269
GRAPHICAL USER INTERFACES 271
AUDIO SERVER ARCHITECTURES 273
UBIQUITOUS AUDIO 278
CASE STUDIES 281
Evolution of a Visual Interface 282
Conversational Desktop 285
Phoneshell 287
Visual User Interfaces to Desktop Audio 292
SUMMARY 295

Chapter 13. Toward More Robust Communication 297

ROBUST COMMUNICATION 298
SPEECH RECOGNITION AND ROBUST PARSING 299
PROSODY 301
WHAT NEXT? 303

Bibliography 305
Index 315

IPR2023-00033
Apple EX1012 Page 9

Introduction

For most of us, speech has been an integral part of our daily lives since we were small children. Speech *is* communication; it is highly expressive and conveys subtle intentions clearly. Our conversations employ a range of interactive techniques to facilitate mutual understanding and ensure that we are understood.

But despite the effectiveness of speech communication, few of us use speech in our daily computing environments. In most workplaces voice is relegated to specialized industrial applications or aids to the disabled; voice is not a part of the computer interfaces based on displays, keyboards, and mice. Although current workstations have become capable of supporting much more sophisticated voice processing, the most successful speech application to date, voice mail, is tied most closely to the telephone.

As speech technologies and natural language understanding mature in the coming decades, many more potential applications will become reality. But much more than raw technology is required to bridge the gap between human conversation and computer interfaces; we must understand the assets and liabilities of voice communication if we are to gauge under which circumstances it will prove to be valuable to end users.

Conversational systems must speak and listen, but they also must understand, pose queries, take turns, and remember the topic of conversation. Understanding how people converse lets us develop better models for interaction with computers by voice. But speech is a very demanding medium to employ effectively, and unless user interaction techniques are chosen with great care, voice applications tend to be slow and awkward to use.

IPR2023-00033
Apple EX1012 Page 10

This book is about using speech in a variety of computing environments based on appreciating its role in human communication. Speech can be used as a method of *interacting* with a computer to place requests or receive warnings and notices. Voice can also be used as the underlying *data* itself, such as notes stored in a calendar, voice annotations of a text document, or telephone messages. Desktop workstations can already support both these speech functions. Speech excels as a method of interacting with the desktop computer over the telephone and has strong potential as the primary channel to access a computer small enough to fit in one's shirt pocket. The full utility of speech will be realized only when it is integrated across *all* these situations; when users find it effective to talk to their computers over the telephone, for example, they will suddenly have more utility for voice as data while in the office.

CONTENTS OF THIS BOOK

This book serves different needs for different readers. The author believes that a firm grounding in the theory of operation of speech technologies forms an important basis for appreciating the difficulties of building applications and interfaces to employ them. This understanding is necessary if we wish to be capable of making any predictions or even guesses of where this field will lead us over the next decade. Paired with descriptions of voice technologies are chapters devoted to applications and user interaction techniques for each, including case studies to illustrate potential applications in more detail. But many chapters stand more or less on their own, and individual readers may pick and choose among them. Readers interested primarily in user interface design issues will gain most benefit from Chapters 4, 6, 8, 9, and 12. Those most concerned about system architectures and support for voice in multimedia computing environments should focus on Chapters 3, 5, 7, and 12. A telecommunications perspective is the emphasis of Chapters 10, 11, and 6.

A conversation requires the ability to speak and to listen, and, if the parties are not in close proximity, some means of transporting their voices across a distance. Chapter 1 discusses the communicative role of speech and introduces some representations of speech and an analytic approach that frames the content of this book. Chapter 2 discusses the physiology of human speech and how we perceive it through our ears; although later chapters refer back to this information, it is not essential for understanding the remainder of the book.

Voice interface technologies are required for computers to participate in conversations. These technologies include digital recording, speech synthesis, and speech recognition; these are the topics of Chapters 3, 5, and 7. Knowledge of the operations of the speech technologies better prepares the reader to appreciate their limitations and understand the impact of improvements in the technologies in the near and distant future.

Although speech is intuitive and seemingly effortless for most of us, it is actually quite difficult to employ as a computer interface. This difficulty is partially due to limitations of current technology but also a result of characteristics inher-

IPR2023-00033
Apple EX1012 Page 11

ent in the speech medium itself. The heart of this book is both criteria for evaluating the suitability of voice to a range of applications and interaction techniques to make its use effective in the user interface. Although these topics are treated throughout this book, they receive particular emphasis in Chapters 4, 6, 8 and 12. These design guidelines are accentuated by case studies scattered throughout the book but especially in these chapters.

These middle chapters are presented in pairs. Each pair contains a chapter describing underlying technology matched with a chapter discussing how to apply the technology. Chapter 3 describes various speech coding methods in a descriptive form and differentiates coding schemes based on data rate, intelligibility, and flexibility. Chapter 4 then focuses on simple applications of stored voice in computer documents and the internal structure of audio editors used to produce those documents. Chapter 5 introduces text-to-speech algorithms. Chapter 6 then draws on both speech coding as well as speech synthesis to discuss *interactive* applications using speech output over the telephone.

Chapter 7 introduces an assortment of speech recognition techniques. After this, Chapter 8 returns to interactive systems, this time emphasizing voice input instead of touch tones. The vast majority of work to date on systems that speak and listen has involved short utterances and brief transactions. But both sentences and conversations exhibit a variety of structures that must be mastered if computers are to become *fluent.* Syntax and semantics constrain sentences in ways that facilitate interpretation; pragmatics relates a person's utterances to intentions and real-world objects; and discourse knowledge indicates how to respond and carry on the thread of a conversation across multiple exchanges. These aspects of speech communication, which are the focus of Chapters 9 and 13, must be incorporated into any system that can engage successfully in a conversation that in any way approaches the way we speak to each other.

Although a discussion of the workings of the telephone network may at first seem tangential to a book about voice in computing, the telephone plays a key role in any discussion of speech and computers. The ubiquity of the telephone assures it a central role in our voice communication tasks. Every aspect of telephone technology is rapidly changing from the underlying network to the devices we hold in our hands, and this is creating many opportunities for computers to get involved in our day-to-day communication tasks. Chapter 10 describes the telephone technologies, while Chapter 11 discusses the integration of telephone functionality into computer workstations. Much of Chapter 6 is about building telephone-based voice applications that can provide a means of accessing personal databases while not in the office.

When we work at our desks, we may employ a variety of speech processing technologies in isolation, but the full richness of voice at the desktop comes with the combination of multiple voice applications. Voice applications on the workstation also raise issues of interaction between both audio and window systems and operating system and run-time support for voice. This is the topic of Chapter 12. Speakers and microphones at every desk may allow us to capture many of the spontaneous conversations we hold every day, which are such an essential

IPR2023-00033
Apple EX1012 Page 12

aspect of our work lives. Desktop voice processing also enables remote telephone access to many of the personal information management utilities that we use in our offices.

ASSUMPTIONS

This book covers material derived from a number of specialized disciplines in a way that is accessible to a general audience. It is divided equally between background knowledge of speech technologies and practical application and interaction techniques. This broad view of voice communication taken in this book is by definition interdisciplinary. Speech communication is so vital and so rich that a number of specialized areas of research have risen around it, including speech science, digital signal processing and linguistics, aspects of artificial intelligence (computational linguistics), cognitive psychology, and human factors. This book touches on all these areas but makes no pretense of covering any of them in depth. This book attempts to open doors by revealing why each of these research areas is relevant to the design of conversational computer systems; the reader with further interest in any of these fields is encouraged to pursue the key overview references mentioned in each chapter.

Significant knowledge of higher mathematics as well as digital signal processing is assumed by many speech texts. These disciplines provide an important level of abstraction and on a practical level are tools required for any serious development of speech technology itself. But to be accessible to a wider audience, this book makes little use of mathematics beyond notation from basic algebra. This book provides an intuitive, rather than rigorous, treatment of speech signal processing to aid the reader in evaluation and selection of technologies and to appreciate their operation and design tradeoffs.

There is a wide gap between the goal of emulating conversational human behavior and what is commercially viable with today's speech technology. Despite the large amount of basic speech research around the world, there is little innovative work on how speech devices may be used in advanced systems, but it is difficult to discuss applications without examples. To this end, the author has taken the liberty to provide more detail with a series of voice projects from the Speech Research Group of M.I.T.'s Media Laboratory (including work from one of its predecessors, the Architecture Machine Group). Presented as case studies, these projects are intended both to illustrate applications of the ideas presented in each chapter and to present pertinent design issues. It is hoped that taken collectively these projects will offer a vision of the many ways in which computers can take part in communication.

IPR2023-00033
Apple EX1012 Page 13

1

Speech as Communication

Speech can be viewed in many ways. Although chapters of this book focus on specific aspects of speech and the computer technologies that utilize speech, the reader should begin with a broad perspective on the role of speech in our daily lives. It is essential to appreciate the range of capabilities that conversational systems must possess before attempting to build them. This chapter lays the groundwork for the entire book by presenting several perspectives on speech communication.

The first section of this chapter emphasizes the *interactive* and *expressive* role of voice communication. Except in formal circumstances such as lectures and dramatic performances, speech occurs in the context of a *conversation,* wherein participants take turns speaking, interrupt each other, nod in agreement, or try to change the topic. Computer systems that talk or listen may ultimately be judged by their ability to converse in like manner simply because conversation permeates human experience. The second section discusses the various components or *layers* of a conversation. Although the distinctions between these layers are somewhat contrived, they provide a means of analyzing the communication process; research disciplines have evolved for the study of each of these components. Finally, the last section introduces the *representations* of speech and conversation, corresponding in part to the layers identified in the second section. These representations provide abstractions that a computer program may employ to engage in a conversation with a human.

IPR2023-00033
Apple EX1012 Page 14

SPEECH AS CONVERSATION

Conversation is a process involving multiple participants, shared knowledge, and a protocol for taking turns and providing mutual feedback. Voice is our primary channel of interaction in conversation, and speech evolved in humans in response to the need among its members to communicate. It is hard to imagine many uses of speech that do not involve some interchange between multiple participants in a conversation; if we are discovered talking to ourselves, we usually feel embarrassed.

For people of normal physical and mental ability, speech is both rich in expressiveness and easy to use. We learn it without much apparent effort as children and employ it spontaneously on a daily basis.[1] People employ many layers of knowledge and sophisticated protocols while having a conversation; until we attempt to analyze dialogues, we are unaware of the complexity of this interplay between parties.

Although much is known about language, study of interactive speech communication has begun only recently. Considerable research has been done on natural language processing systems, but much of this is based on keyboard input. It is important to note the contrast between written and spoken language and between read or rehearsed speech and spontaneous utterances. Spoken language is less formal than written language, and errors in construction of spoken sentences are less objectionable. Spontaneous speech shows much evidence of the real-time processes associated with its production, including false starts, nonspeech noises such as mouth clicks and breath sounds, and pauses either silent or filled (". . . um . . .") [Zue *et al.* 1989b]. In addition, speech naturally conveys intonational and emotional information that fiction writers and playwrights must struggle to impart to written language.

Speech is rich in interactive techniques to guarantee that the listener understands what is being expressed, including facial expressions, physical and vocal gestures, "uh-huhs," and the like. At certain points in a conversation, it is appropriate for the listener to begin speaking; these points are often indicated by longer pauses and lengthened final syllables or marked decreases in pitch at the end of a sentence. Each round of speech by one person is called a **turn; interruption** occurs when a participant speaks before a break point offered by the talker. Instead of taking a turn, the listener may quickly indicate agreement with a word or two, a nonverbal sound ("uh-huh"), or a facial gesture. Such responses, called **back channels,** speed the exchange and result in more effective conversations [Kraut *et al.* 1982].[2]

Because of these interactive characteristics, speech is used for immediate communication needs, while writing often implies a distance, either in time or space,

[1]For a person with normal speech and hearing to spend a day without speaking is quite a novel experience.

[2]We will return to these topics in Chapter 9.

IPR2023-00033
Apple EX1012 Page 15

between the author and reader. Speech is used in transitory interactions or situations in which the process of the interaction may be as important as its result. For example, the agenda for a meeting is likely to be written, and a written summary or minutes may be issued "for the record," but the actual decisions are made during a conversation. Chapanis and his colleagues arranged a series of experiments to compare the effectiveness of several communication media, i.e., voice, video, handwriting, and typewriting, either alone or in combination, for problem-solving tasks [Ochsman and Chapanis 1974]. Their findings indicated an overwhelming contribution of voice for such interactions. Any experimental condition that included voice was superior to any excluding voice; the inclusion of other media with voice resulted in only a small additional effectiveness. Although these experiments were simplistic in their use of student subjects and invented tasks and more recent work by others [Minneman and Bly 1991] clarifies a role for video interaction, the dominance of voice seems unassailable.

But conversation is more than mere interaction; communication often serves a purpose of changing or influencing the parties speaking to each other. I tell you something I have learned with the intention that you share my knowledge and hence enhance your view of the world. Or I wish to obtain some information from you so I ask you a question, hoping to elicit a reply. Or perhaps I seek to convince you to perform some activity for me; this may be satisfied either by your physical performance of the requested action or by your spoken promise to perform the act at a later time. "Speech Act" theories (to be discussed in more detail in Chapter 9) attempt to explain language as action, e.g., to request, command, query, and promise, as well as to inform.

The intention behind an utterance may not be explicit. For example, "Can you pass the salt?" is not a query about one's ability; it is a request. Many actual conversations resist such purposeful classifications. Some utterances ("go ahead," "uh-huh," "just a moment") exist only to guide the flow of the conversation or comment on the state of the discourse, rather than to convey information. Directly purposeful requests are often phrased in a manner allowing flexibility of interpretation and response. This looseness is important to the process of people defining and maintaining their work roles with respect to each other and establishing socially comfortable relationships in a hierarchical organization. The richness of speech allows a wide range of "acceptance" and "agreement" from wholehearted to skeptical to incredulous.

Speech also serves a strong social function among individuals and is often used just to pass the time, tell jokes, or talk about the weather. Indeed, extended periods of silence among a group may be associated with interpersonal awkwardness or discomfort. Sometimes the actual occurrence of the conversation serves a more significant purpose than any of the topics under discussion. Speech may be used to call attention to oneself in a social setting or as an exclamation of surprise or dismay in which an utterance has little meaning with respect to any preceding conversation. [Goffman 1981]

The expressiveness of speech and robustness of conversation strongly support the use of speech in computer systems, both for stored voice as a data type as well as speech as a medium of interaction. Unfortunately, current computers are

IPR2023-00033
Apple EX1012 Page 16

capable of uttering only short sentences of marginal intelligibility and occasionally recognizing single words. Engaging a computer in a conversation can be like an interaction in a foreign country. One studies the phrase book, utters a request, and in return receives either a blank stare (wrong pronunciation, try again) or a torrent of fluent speech in which one cannot perceive even the word boundaries.

However, limitations in technology only reinforce the need to take advantage of conversational techniques to ensure that the user is understood. Users will judge the performance of computer systems employing speech on the basis of their expectations about conversation developed from years of experience speaking with fellow humans. Users may expect computers to be either deaf and dumb, or once they realize the system can talk and listen, expect it to speak fluently like you and me. Since the capabilities of current speech technology lie between these extremes, building effective conversational computer systems can be very frustrating.

HIERARCHICAL STRUCTURE OF CONVERSATION

A more analytic approach to speech communication reveals a number of different ways of describing what actually occurs when we speak. The hierarchical structure of such analysis suggests goals to be attained at various stages in computer-based speech communication.

Conversation requires apparatus both for listening and speaking. Effective communication invokes mental processes employing the mouth and ears to convey a message thoroughly and reliably. There are many layers at which we can analyze the communication process, from the lower layers where speech is considered primarily acoustically to higher layers that express meaning and intention. Each layer involves increased knowledge and potential for intelligence and interactivity.

From the point of view of the speaker, we may look at speech from at least eight layers of processing as shown in Figure 1.1.

Layers of Speech Processing

discourse The regulation of conversation for pragmatic ends. This includes taking turns talking, the history of referents in a conversation so pronouns can refer to words spoken earlier, and the process of introducing new topics.

pragmatics The intent or motivation for an utterance. This is the underlying reason the utterance was spoken.

semantics The meaning of the words individually and their meaning as combined in a particular sentence.

syntax The rules governing the combination of words in a sentence, their parts of speech, and their forms, such as case and number.

IPR2023-00033
Apple EX1012 Page 17

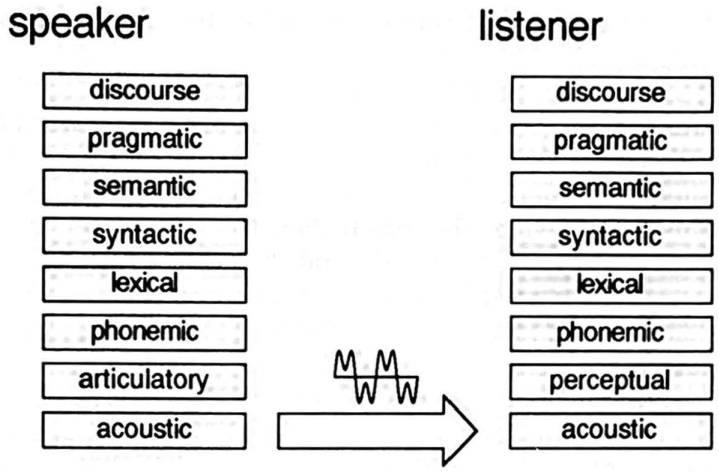


Figure 1.1. A layered view of speech communication.

lexical The set of words in a language, the rules for forming new words from affixes (prefixes and suffixes), and the stress ("accent") of syllables within the words.

phonetics The series of sounds that uniquely convey the series of words in the sentence.

articulation The motions or configurations of the vocal tract that produce the sounds, e.g., the tongue touching the lips or the vocal cords vibrating.

acoustics The realization of the string of phonemes in the sentence as vibrations of air molecules to produce pressure waves, i.e., sound.

Consider two hikers walking through the forest when one hiker's shoelace becomes untied. The other hiker sees this and says, "Hey, you're going to trip on your shoelace." The listener then ties the shoelace. We can consider this utterance at each layer of description.

Discourse analysis reveals that "Hey" serves to call attention to the urgency of the message and probably indicates the introduction of a new topic of conversation. It is probably spoken in a raised tone and an observer would reasonably expect the listener to acknowledge this utterance, either with a vocal response or by tying the shoe. Experience with discourse indicates that this is an appropriate interruption or initiation of a conversation at least under some circumstances. Discourse structure may help the listener understand that subsequent utterances refer to the shoelace instead of the difficulty of the terrain on which the conversants are traveling.

In terms of **pragmatics,** the speaker's intent is to warn the listener against tripping; presumably the speaker does not wish the listener to fall. But this utterance might also have been a ruse intended to get the listener to look down for the sake of playing a trick. We cannot differentiate these possibilities without know-

IPR2023-00033
Apple EX1012 Page 18

ing more about the context in which the sentence was spoken and the relationship between the conversants.

From a **semantics** standpoint, the sentence is about certain objects in the world: the listener, hiking, an article of clothing worn on the foot, and especially the string by which the boot is held on. The concern at the semantic layer is how the words refer to the world and what states of affairs they describe or predict. In this case, the meaning has to do with an animate entity ("you") performing some physical action ("tripping"), and the use of future tense indicates that the talker is making a prediction of something not currently taking place. Not all words refer to specific subjects; in the example, "Hey" serves to attract attention, but has no innate meaning.

Syntax is concerned with how the words fit together into the structure of the sentence. This includes the ordering of parts of speech (nouns, verbs, adjectives) and relations between words and the words that modify them. Syntax indicates that the correct word order is subject followed by verb, and syntax forces agreement of number, person, and case of the various words in the sentence. "You *is* going to . . ." is syntactically ill formed. Because the subject of the example is "you," the associated form of the verb "to be" is "are." The chosen verb form also indicates a future tense.

Lexical analysis tells us that "shoelace" comes from the root words "shoe" and "lace" and that the first syllable is stressed. Lexical analysis also identifies a set of definitions for each word taken in isolation. "Trip," for example, could be the act of falling or it could refer to a journey. Syntax reveals which definition is appropriate as each is associated with the word "trip" used as a different part of speech. In the example, "trip" is used as a verb and so refers to falling.

The **phonemic** layer is concerned with the string of phonemes of which the words are composed. Phonemes are the speech sounds that form the words of any language.[3] Phonemes include all the sounds associated with vowels and consonants. A grunt, growl, hiss, or gargling sound is not a phoneme in English, so it cannot be part of a word; such sounds are not referred to as speech. At the phoneme layer, while talking we are either continuously producing speech sounds or are silent. We are not silent at word boundaries; the phonemes all run together.

At the **articulatory** layer, the speaker makes a series of vocal gestures to produce the sounds that make up the phonemes. These sounds are created by a noise source at some location in the vocal tract, which is then modified by the configuration of the rest of the vocal tract. For example, to produce a "b" sound, the lips are first closed and air pressure from the lungs is built up behind them. A sudden release of air between the lips accompanied by vibration of the vocal cords produces the "b" sound. An "s" sound, by comparison, is produced by turbulence caused as a stream of air rushes through a constriction formed by the tongue and the roof of the mouth. The mouth can also be used to create nonspeech sounds, such as sighs and grunts.

[3]A more rigorous definition will be given in the next section.

IPR2023-00033
Apple EX1012 Page 19

Finally, the **acoustics** of the utterance is its nature as sound. Sound is transmitted as variations in air pressure over time; sound can be converted to an electrical signal by a microphone and represented as an electrical waveform. We can also analyze sound by converting the waveform to a spectrogram, which displays the various frequency components present in the sound. At the acoustic layer, speech is just another sound like the wind in the trees or a jet plane flying overhead.

From the perspective of the listener, the articulatory layer is replaced by a **perceptual** layer, which comprises the processes whereby sound (variations in air pressure over time) is converted to neural signals in the ear and ultimately interpreted as speech sounds in the brain. It is important to keep in mind that the hearer can directly sense only the acoustic layer of speech. If we send an electric signal representing the speech waveform over a telephone line and convert this signal to sound at the other end, the listening party can understand the speech. Therefore, the acoustic layer alone must contain all the information necessary to understand the speaker's intent, but it can be represented at the various layers as part of the process of understanding.

This layered approach is actually more descriptive than analytic in terms of human cognitive processes. The distinctions between the layers are fuzzy, and there is little evidence that humans actually organize discourse production into such layers. Intonation is interpreted in parallel at all these layers and thus illustrates the lack of sharp boundaries or sequential processing among them. At the pragmatic layer, intonation differentiates the simple question from exaggerated disbelief; the same words spoken with different intonation can have totally different meaning. At the syntactic layer, intonation is a cue to phrase boundaries. Intonation can differentiate the noun and verb forms of some words (e.g., conduct, convict) at the syntactic layer by conveying lexical stress. Intonation is not phonemic in English, but in some other languages a change in pitch does indicate a different word for the otherwise identical articulation. And intonation is articulated and realized acoustically in part as the fundamental frequency at which the vocal cords vibrate.

Dissecting the communication process into layers offers several benefits, both in terms of understanding as well as for practical implementations. Understanding this layering helps us appreciate the complexity and richness of speech. Research disciplines have evolved around each layer. A layered approach to representing conversation is essential for modular software development; a clean architecture isolates each module from the specialized knowledge of the others with information passed over well-defined interfaces. Because each layer consists of a different perspective on speech communication, each is likely to employ its own representation of speech for analysis and generation.

As a cautionary note, it needs to be recognized from the start that there is little evidence that humans actually function by invoking each of these layers during conversation. The model is descriptive without attempting to explain or identify components of our cognitive processes. The model is incomplete in that there are some aspects of speech communication that do not fit it, but it can serve as a framework for much of our discussion of conversational computer systems.

IPR2023-00033
Apple EX1012 Page 20

REPRESENTATIONS OF SPEECH

We need a means of describing and manipulating speech at each of these layers. Representations for the lower layers, such as acoustic waveforms or phonemes, are simpler and more complete and also more closely correspond to directly observable phenomena. Higher-layer representations, such as semantic or discourse structure, are subject to a great deal more argument and interpretation and are usually abstractions convenient for a computer program or a linguistic comparison of several languages. Any single representation is capable of conveying particular aspects of speech; different representations are suitable for discussion of the different layers of the communication process.

The representation chosen for any layer should contain all the information required for analysis at that layer. Since higher layers possess a greater degree of abstraction than lower layers, higher-layer representations extract features from lower-layer representations and hence lose the ability to recreate the original information completely. For example, numerous cues at the acoustic layer may indicate that the talker is female, but if we represent the utterance as a string of phones or of words we have lost those cues. In terms of computer software, the representation is the data type by which speech is described. One must match the representation and the particular knowledge about speech that it conveys to the algorithms employing it at a particular layer of speech understanding.

Acoustic Representations

Sounds consist of variations in air pressure over time at frequencies that we can hear. Speech consists of a subset of the sounds generated by the human vocal tract. If we wish to analyze a sound or save it to hear again later, we need to capture the variations in air pressure. We can convert air pressure to electric voltage with a microphone and then convert the voltage to magnetic flux on an audiocassette tape using a recording head, for example.

We can plot the speech signal in any of these media (air pressure, voltage, or magnetic flux) over time as a **waveform** as illustrated in Figure 1.2. This representation exhibits positive and negative values over time because the speech radiating from our mouths causes air pressure to be temporarily greater or less than that of the ambient air.

A waveform describing sound pressure in air is continuous, while the waveforms employed by computers are **digital,** or **sampled,** and have discrete values for each sample; these concepts are described in detail in Chapter 3. Tape recorders store analog waveforms; a compact audio disc holds a digital waveform. A digitized waveform can be made to very closely represent the original sound, and it can be captured easily with inexpensive equipment. A digitized sound stored in computer memory allows for fast random access. Once digitized, the sound may be further processed or compressed using digital signal processing techniques. The analog audiotape supports only sequential access (it must be rewound or fast-forwarded to jump to a different part of the tape) and is prone to mechanical breakdown.

IPR2023-00033
Apple EX1012 Page 21

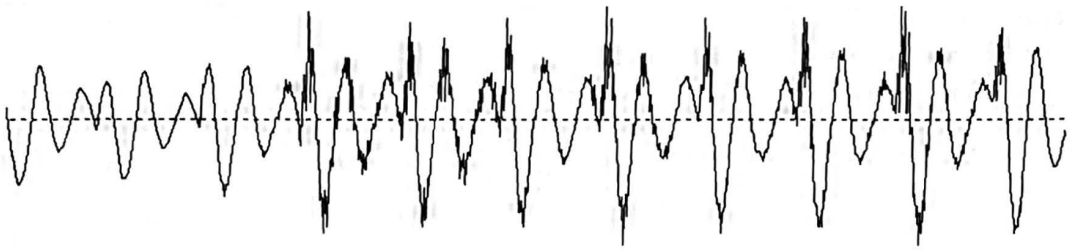


Figure 1.2. A waveform, showing 100 milliseconds of the word "me." The vertical axis depicts amplitude, and the horizontal axis represents time. The display depicts the transition from "m" to "e."

A waveform can effectively represent the original signal visually when plotted on a piece of paper or a computer screen. But to make observations about what a waveform sounds like, we must analyze it across a span of time not just at a single point. For example, in Figure 1.2 we can determine the amplitude ("volume") of the signal by looking at the differences between its highest and lowest points. We can also see that it is periodic: The signal repeats a pattern over and over. Since the horizontal axis represents time, we can determine the frequency of the signal by counting the number of periods in one second. A periodic sound with a higher frequency has a higher pitch than a periodic sound with a lower frequency.

One disadvantage of working directly with waveforms is that they require considerable storage space, making them bulky; Figure 1.2 shows only 100 milliseconds of speech. A variety of schemes for compressing speech to minimize storage are discussed in Chapter 3. A more crucial limitation is that a waveform simply shows the signal as a function of time. A waveform is in no way speech specific and can represent any acoustical phenomenon equally well. As a general-purpose representation, it contains all the acoustic information but does not explicitly describe its content in terms of properties of speech signals.

A speech-specific representation more succinctly conveys those features salient to speech and phonemes, such as syllable boundaries, fundamental frequency, and the higher-energy frequencies in the sound. A **spectrogram** is a transformation of the waveform into the frequency domain. As seen in Figure 1.3, the spectrogram reveals the distribution of various frequency components of the signal as a function of time indicating the energy at each frequency. The horizontal axis represents time, the vertical axis represents frequency, and the intensity or blackness at a point indicates the acoustic energy at that frequency and time.

A spectrogram still consists of a large amount of data but usually requires much less storage than the original waveform and reveals acoustic features specific to speech. Because of this, spectral analysis[4] is often employed to process

[4]To be precise, spectral or Fourier analysis uses mathematical techniques to derive the values of energy at particular frequencies. We can plot these as described above; this visual representation is the spectrogram.

IPR2023-00033
Apple EX1012 Page 22

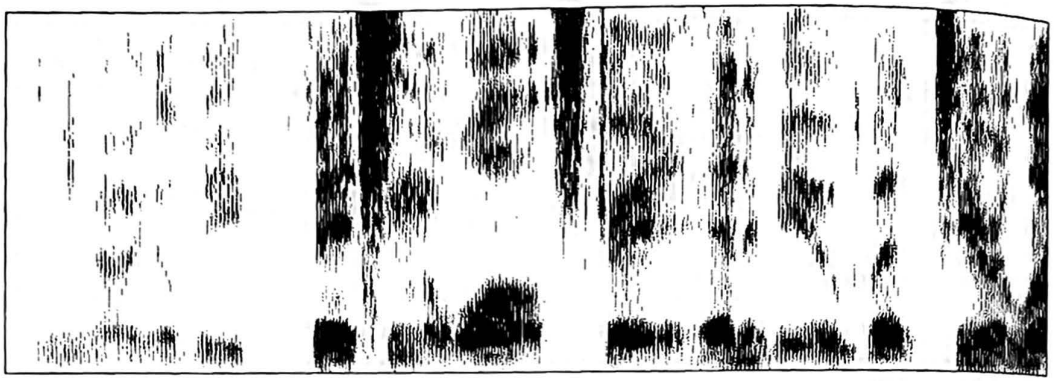


Figure 1.3. A spectrogram of 2.5 seconds of speech. The vertical axis is frequency, the horizontal axis is time, and energy maps to darkness.

speech for analysis by a human or a computer. People have been trained to read spectrograms and determine the words that were spoken. Although the spectrogram conveys salient features of a sound, the original acoustic signal cannot be reconstructed from it without some difficulty. As a result, the spectrogram is more useful for analysis of the speech signal than as a means of storing it for later playback.

Other acoustic representations in the frequency domain are even more succinct, though they are more difficult for a human to process visually than a spectrogram. **Linear Prediction Coefficients** and **Cepstral Analysis,** for example, are two such techniques that rely heavily on digital signal processing.[5] Both of these techniques reveal the resonances of the vocal tract and separate information about how the sound was produced at the noise source and how it was modified by various parts of the vocal tract. Because these two techniques extract salient information about how the sound was articulated they are frequently used as representations for computer analysis of speech.

PHONEMES AND SYLLABLES

Two representations of speech which are more closely related to its lexical structure are phonemes and syllables. Phonemes are important small units, several of which make up most syllables.

Phonemes

A **phoneme** is a unit of speech, the set of which defines all the sounds from which words can be constructed in a particular language. There is at least one pair of

[5]Linear prediction will be explained in Chapter 3. Cepstral analysis is beyond the scope of this book.

IPR2023-00033
Apple EX1012 Page 23

words in a language for which replacing one phoneme with another will change what is spoken into a different word. Of course, not every combination of phonemes results in a word; many combinations are nonsense.

For example, in English, the words "bit" and "bid" have different meanings, indicating that the "t" and "d" are different phonemes. Two words that differ in only a single phoneme are called a **minimal pair.** "Bit" and "bif" also vary by one sound, but this example does not prove that "t" and "f" are distinct phonemes as "bif" is not a word. But "tan" and "fan" are different words; this proves the phonemic difference between "t" and "f".

Vowels are also phonemes; the words "heed," "had," "hid," "hide," "howed," and "hood" each differ only by one sound, showing us that English has at least six vowel phonemes. It is simple to construct a minimal pair for any two vowels in English, while it may not be as simple to find a pair for two consonants.

An **allophone** is one of a number of different ways of pronouncing the same phoneme. Replacing one allophone of a phoneme with another does not change the meaning of a sentence, although the speaker will sound unnatural, stilted, or like a non-native. For example, consider the "t" sound in "sit" and "sitter." The "t" in "sit" is somewhat aspirated; a puff of air is released with the consonant. You can feel this if you put your hand in front of your mouth as you say the word. But in "sitter" the same phoneme is not aspirated; we say the aspiration is not phonemic for "t" and conclude that we have identified two allophones. If you aspirate the "t" in "sitter," it sounds somewhat forced but does not change the meaning of the word.

In contrast, aspiration of stop consonants is phonemic in Nepali. For an example of English phonemes that are allophones in another language, consider the difficulties Japanese speakers have distinguishing our "l" and "r." The reason is simply that while these are two phonemes in English, they are allophonic variants on the same phoneme in Japanese. Each language has its own set of phonemes and associated allophones. Sounds that are allophonic in one language may be phonemic in another and may not even exist in a third. When you learn a language, you learn its phonemes and how to employ the permissible allophonic variations on them. But learning phonemes is much more difficult as an adult than as a child.

Because phonemes are language specific, we can not rely on judgments based solely on our native languages to classify speech sounds. An individual speech sound is a **phone,** or **segment.** For any particular language, a given phoneme will have a set of allophones, each of which is a segment. Segments are properties of the human vocal mechanism, and phonemes are properties of languages. For most practical purposes, phone and phoneme may be considered to be synonyms.

Linguists use a notation for phones called the **International Phonetic Alphabet,** or **IPA.** IPA has a symbol for almost every possible phone; some of these symbols are shown in Figure 1.4. Since there are far more than 26 such phones, it is not possible to represent them all with the letters of the English alphabet. IPA borrows symbols from the Greek alphabet and elsewhere. For example, the "th" sound in "thin" is represented as "θ" in IPA, and the sound

IPR2023-00033
Apple EX1012 Page 24

Phoneme	Example Word	Phoneme	Example Word	Phoneme	Example Word
i	beet	p	put	č	chin
I	bit	t	tap	ǰ	judge
ɛ	bet	k	cat	m	map
e	bait	b	bit	n	nap
æ	bat	d	dill	ŋ	sing
ɑ	cot	g	get	r	ring
ɔ	caught	f	fain	l	lip
ʌ	but	θ	thin	w	will
o	boat	s	sit	y	yell
U	foot	ʃ	shoe	h	head
u	boot	v	veal		
ɜ	bird	ð	then		
ɑj (ɑI)	bite	z	zeal		
ɔj (ɔI)	boy	ž	azure		
ɑw (ɑU)	bout				
ə	about				

Figure 1.4. The English phonemes in IPA, the International Phonetic Alphabet.

in "then" is "δ."[6] American linguists who use computers have developed the **Arpabet,** which uses ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters. Arpabet[7] was developed for the convenience of computer manipulation and representation of speech using ASCII-printable characters.

To avoid the necessity of the reader learning either IPA or Arpabet, this book indicates phones by example, such as, "the 't' in bottle." Although slightly awkward, such a notation suffices for the limited examples described. The reader will find it necessary to learn a notation (IPA is more common in textbooks) to make any serious study of phonetics or linguistics.

A phonemic transcription, although compact, has lost much of the original signal content, such as pitch, speed, and amplitude of speech. Phonemes are abstractions from the original signal that highlight the speech-specific aspects of that

[6]Are these sounds phonemic in English?

[7]The word comes from the acronym ARPA, the Advanced Research Projects Agency (sometimes called DARPA), a research branch of the U.S. Defense Department that funded much early speech research in this country and continues to be the most significant source of government support for such research.

IPR2023-00033
Apple EX1012 Page 25

signal; this makes a phonemic transcription a concise representation for lexical analysis as it is much more abstract than the original waveform.

Syllables

Another natural way to divide speech sounds is by the **syllable.** Almost any native speaker of English can break a word down into syllables, although some words can be more difficult, e.g., "chocolate" or "factory." A syllable consists of one or more consonants, a vowel (or diphthong[8]), followed by one or more consonants; consonants are optional, but the vowel is not. Two or more adjacent consonants are called a consonant **cluster;** examples are the initial sounds of "screw" and "sling." Acoustically, a syllable consists of a relatively high energy core (the vowel) optionally preceded or followed by periods of lower energy (consonants). Consonants have lower energy because they impose constrictions on the air flow from the lungs. Many natural languages, such as those written with the Arabic script and the northern Indian languages, are written with a syllabic system in which one symbol represents both a consonant and its associated vowel.

Other Representations

There are many other representations of speech appropriate to higher layer aspects of conversation.[9] Lexical analysis reveals an utterance as a series of **words.** A dictionary, or **lexicon** lists all the words of a language and their meanings. The **phrase,** sometimes called a "breath group" when describing intonation, is relevant both to the study of prosody (pitch, rhythm, and meter) as well as syntax, which deals with structures such as the noun phrase and verb phrase. A **parse tree,** as shown in Figure 1.5, is another useful representation of the syntactic relationships among words in a sentence.

Representations for higher layers of analysis are varied and complex. Semantics associates meaning with words, and meaning implies a relationship to other words or concepts. A **semantic network** indicates the logical relationships between words and meaning. For example, a door is a physical object, but it has specific meaning only in terms of other objects, such as walls and buildings, as it covers entrance holes in these objects.

Discourse analysis has produced a variety of models of the focus of a conversation. For example, one of these uses a **stack** to store potential topics of current focus. New topics are pushed onto the stack, and a former topic again becomes the focus when all topics above it are popped off the stack. Once removed from the stack, a topic cannot become the focus without being reintroduced.

---

[8]A diphthong consists of two vowel sounds spoken in sequence and is considered a single phoneme. The two vowel sounds in a diphthong cannot be separated into different syllables. The vowels in "hi" and "bay" are examples of diphthongs.

[9]Most of the speech representations mentioned in this section will be detailed in Chapter 9.

IPR2023-00033
Apple EX1012 Page 26

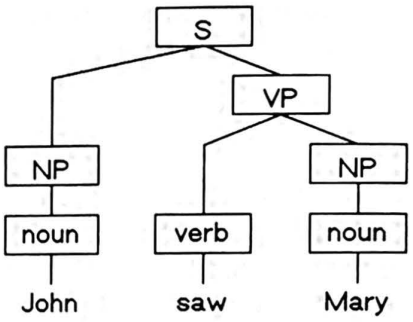


Figure 1.5. A parse tree.

SUMMARY

The perspective of speech as a conversational process constitutes the foundation and point of view of this book. Conversations employ spontaneous speech and a variety of interaction techniques to coordinate the exchange of utterances between participants. Speech is our primary medium of communication, although writing may be favored for longer-lasting or more formal messages. Computer systems would do well to exploit the richness and robustness of human conversation.

But "conversation" is simply too rich and varied a term to be amenable to analysis without being analyzed in components. This chapter identified a number of such components of a conversation: acoustic, articulatory, phonetic, lexical, syntactic, semantic, pragmatic, and discourse. Disciplines of research have been established for each of these areas, and the rest of this book will borrow heavily from them. Each discipline has its own set of representations of speech, which allow utterances to be described and analyzed.

Representations such as waveforms, spectrograms, and phonetic transcriptions provide suitable abstractions that can be embedded in the computer programs that attempt to implement various layers of speech communication. Each representation highlights particular features or characteristics of speech and may be far removed from the original speech sounds.

The rationale for this book is that the study of speech in conversations is interdisciplinary and that the designers of conversational computer systems need to understand each of these individual components in order to fully appreciate the whole. The rest of this book is organized in part as a bottom-up analysis of the layers of speech communication. Each layer interacts with the other layers, and the underlying goal for conversational communication is unification of the layers.

IPR2023-00033
Apple EX1012 Page 27

2

Speech Production and Perception

A basic knowledge of the physiology of speech production and its perception is necessary to understand both speech synthesis and speech recognition as well as compression techniques. This chapter provides an overview of the production and perception of speech. It briefly treats the organs of speech, i.e., the vocal tract and the auditory system. An articulatory model of speech is presented; this explains the different speech sounds of a language in terms of how they are produced. Finally, it introduces some basic findings of psychoacoustics that relate to the manner in which sound is perceived by the listener.

To distinguish the various types of sounds in English and the mechanisms whereby they are produced some signal processing terms are introduced. In this text, very simple and intuitive descriptions of such terms are given; more formal definitions may be found in texts on signal processing.

VOCAL TRACT

The vocal tract is the set of organs that produce speech sounds; these organs are also used for eating, drinking, and breathing. As seen in Figure 2.1, the vocal tract includes portions of the throat, mouth, and nasal cavities. These organs, the **articulators,** are moved to various configurations to produce the different sounds that constitute speech. The primary topic of this section is the production of speech sounds in general; the next section classifies the sounds specific to English.

IPR2023-00033
Apple EX1012 Page 28

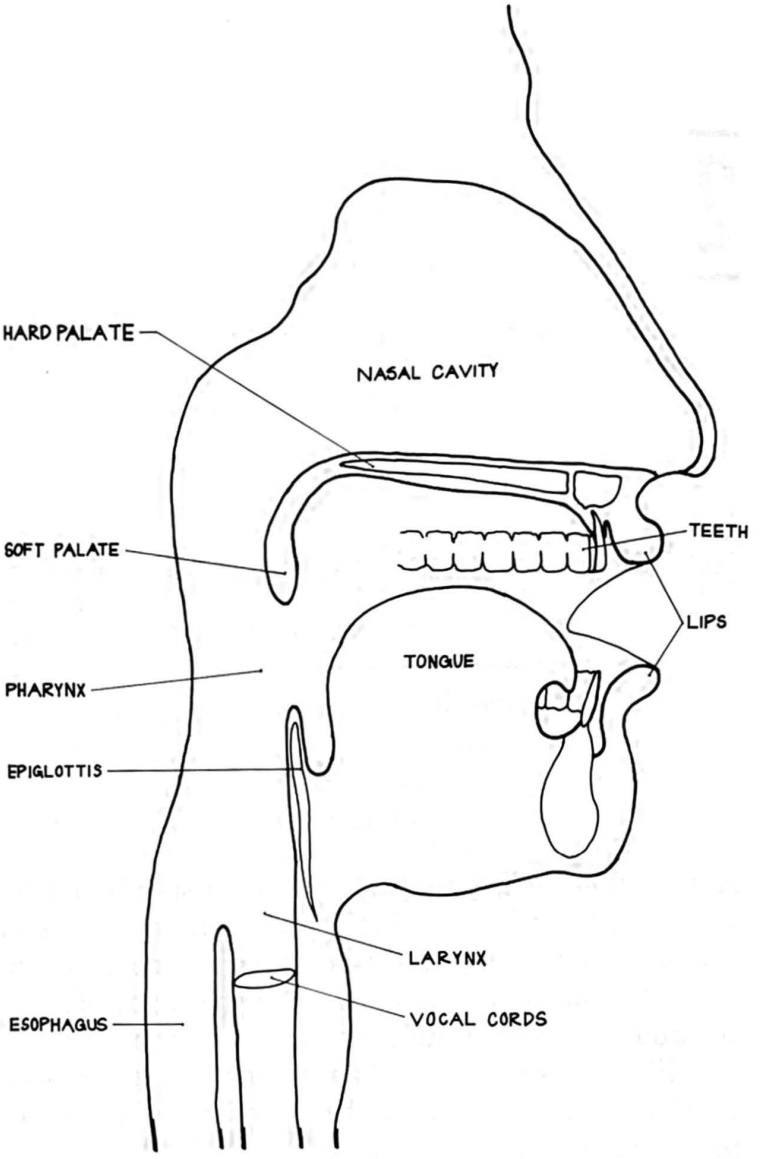


Figure 2.1. The major components of the vocal tract.

Sound is vibration in a medium, usually air, at frequencies that we can hear. In air, sound is carried as variations in pressure over time transmitted by the motion of air molecules. The air molecules create pressure variations by moving back and forth while not traveling a substantial distance. This variation is similar to the wave pattern generated when one end of a Slinky toy is moved back and forth. Some force must be created to cause this motion of molecules. In the vocal tract, this force is provided by air being expelled from the lungs by muscles in the diaphragm.

Without further modification, the air coming out of the lungs does not create a significant sound. If you open your mouth wide and exhale, you hear a little noise created as the air rushes out of the mouth, but this is usually not audible across a room. To become speech this air flow must be focused or concentrated in some

IPR2023-00033
Apple EX1012 Page 29

way. Focusing is accomplished by one of two mechanisms, either vibration of the vocal cords to produce a periodic sound or turbulent air flow through a constriction to produce an aperiodic, or "noisy," sound.[1]

If the vocal cords are vibrating, the speech produced is termed **voiced.** The vocal cords are folds of tissue capable of opening and closing in a regular fashion as controlled by several sets of muscles and driven by the pressure of air being expelled from the lungs. In operation the vocal cords remain closed until the pressure behind them is strong enough to force them open. As the air rushes by and pressure is released, a partial vacuum (Bernoulli effect) helps pull the cords back together again to shut off the flow of air.

This opening and closing pattern occurs at a regular rate and results in a series of pulses at that frequency. This frequency, called the **fundamental frequency** of voicing, or **F0,** corresponds to the pitch that we perceive in speech. F0 is typically higher for female and young speakers and may be varied within a limited range by muscle tension. The waveform of this glottal pulse is roughly triangular (see Figure 2.2), which results in a spectrum rich in harmonics, i.e., higher frequency components. The energy of speech drops at about 12 dB[2] per octave from the fundamental. Figure 2.3 illustrates the spectrum of the sound produced by the vocal chords. The **spectrum** is a measure of the magnitude of all the frequency components that constitute a signal. Since the source is of a single frequency F0, energy is found at that frequency and at integral multiples of that frequency. For example, the author's typical fundamental frequency is at 110 Hz,[3] which means that energy will be found at 220 Hz, 330 Hz, 440 Hz, etc.

You can experience voicing directly for yourself. Place a finger at the front bottom of your throat, slightly above the level of the shoulders. Speak out loud, alternating making "s" and "z" sounds. During the voiced "z" you can feel the vocal cords vibrate with your finger.

Another source of sound in the vocal tract is turbulence as air rushes through a constriction. Turbulence is not periodic; it is caused by the continuous stream of air molecules bouncing around randomly. It is similar in nature to the sound made by a babbling brook or the wind blowing through trees or surf moving onto shore. Such sounds contain a relatively large amount of high frequency energy, or "hiss," because the acoustic waveform varies a great deal from moment to moment. Turbulence is characteristic of both continuous **frication** sounds, such as the phoneme "s," and sudden **plosive** ones associated with a quick opening of a closed vocal tract, as in the phoneme "p." This turbulence can take place at a number of locations in the vocal tract. In the fricative "s," it is at a constriction formed between the tongue and the roof of the mouth. With the plosive "p," it is

[1]A periodic signal repeats a pattern at a regular interval or frequency. Aperiodic signals do not exhibit such patterns. Wind instruments produce periodic sounds, while percussion instruments are generally aperiodic.

[2]1 dB, short for decibel, corresponds to a just noticeable difference in sound energy.

[3]Hz, or Hertz, means cycles (or periods) per second. 1 kHz, or one kiloHertz, is one thousand cycles per second. 1 MHz, or MegaHertz, is one million cycles per second.

IPR2023-00033
Apple EX1012 Page 30

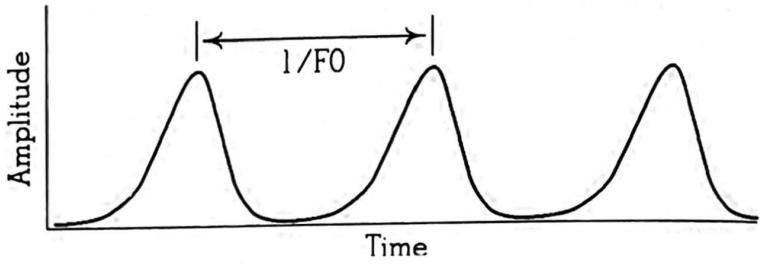


Figure 2.2. The glottal pulse is a roughly triangular-shaped waveform.

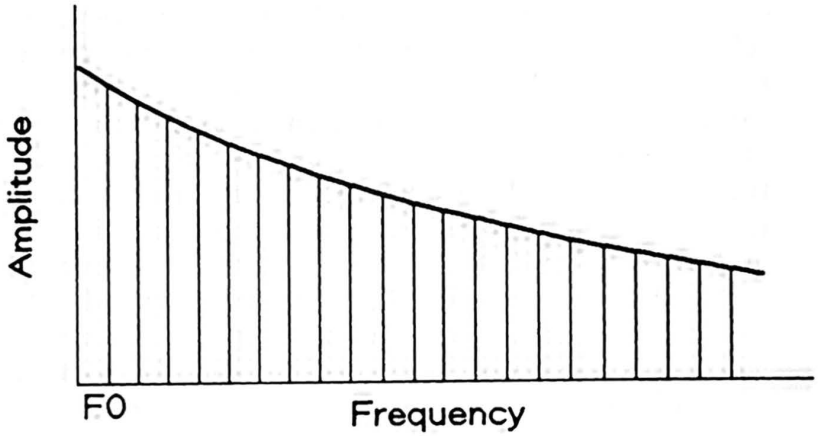


Figure 2.3. The spectrum of the sound produced by the vocal chords shows gradually decreasing magnitude with increasing frequency. This is properly a **line spectrum,** as energy is found only at multiples of F0, shown by the vertical line. The curve connecting the tops of these lines shows the **envelope,** or overall shape of the spectrum.

formed at the lips. The frication in "f" occurs around the lower lip making contact with the upper teeth.

There is more to classifying speech than identifying the source of the noise. The sound created by the noise source must then pass through the rest of the vocal tract that modifies it. These modifications are responsible for many of the distinguishing features of the different phonemes, or speech sounds. Even if the sound source is at the lips such that it does not have to pass through any further portions of the vocal tract, the size and shape of the oral cavity behind the lips affects the sound quality.

The vocal tract is usually analyzed as a series of **resonances.** A resonator is a physical system that enhances certain frequencies or range of frequencies. Resonators that amplify a very narrow range of frequencies are referred to as "high

IPR2023-00033
Apple EX1012 Page 31

Q",[4] and those which have a very broad band are "low Q" (see Figure 2.4). Because the vocal tract is composed of soft tissue, its resonators are rather low Q. In contrast, a very "clear" sound, such as a flute note, is produced by a high Q resonator.

The vocal tract is about 17 cm long and can be approximated as a series of joined tubes of varying diameters and lengths that add up to 17. One tube extends from the vocal cords up to the top of the mouth, the other, at approximately a right angle to the first, extends from there to the lips. The relative lengths of these tubes can be varied by the positioning of the tongue. If the tongue is low and flat (as in "ah"), then the front tube formed by the mouth is pronounced and the back tube is shorter. When the tongue is high in the mouth (as when saying the letter "e"), the front tube in front of the tongue is quite short, while the back tube is elongated.

If a column of air in a tube is excited by a source, it resonates at a characteristic frequency dependent upon the length of the tube. This is the principle upon which the flute is designed; opening the finger holes allows the air to escape at various points along the tube, thereby changing the length of the air column and thus its frequency. In the same way, the sound produced from the vocal cord vibration is enhanced at the frequencies corresponding to the length of the tubes in the vocal tract (see Figure 2.5). The frequencies at which these resonances occur are called **formants.** The lowest-frequency formant is labeled **F1** and called the first formant; the second formant is labeled **F2**, and so on. Most analysis stops at five or six formants. In the multitube model, each formant corresponds to a tube section.

Whichever the noise source, either voiced or turbulent, its spectrum is enhanced around the formant frequencies, which depend on the physical configuration of the vocal tract. Another part of the vocal tract that influences the noise source is the **nasal cavity,** a fairly soft region above the roof of the mouth and behind the nose. The passage from the back of the mouth to the nasal cavity can be closed by the **velum,** a small flap of tissue toward the top back of the mouth. When the velum is opened, the nasal cavity is physically connected and acoustically coupled with the vocal tract. The nasal cavity absorbs a significant amount of sound energy in the lower portion of the energy spectrum.

The radiation of the speech signal from the vocal tract into the surrounding air must also be considered. In addition to the mouth, sound is radiated from the cheeks, throat, and bony structure of the head and jaw, resulting in a signal that is much less directional than it would be otherwise with radiational patterns that are heavily frequency dependent. Flanagan [Flanagan 1960] measured sound energies of various frequencies at multiple locations about the head and found that at approximately 45 degrees (horizontally) away from the axis of the mouth

[4]Q, or Quality factor, is a measure of the difference between the higher and lower bounds of the frequencies amplified by the resonator, as measured at the point of half-power relative to the peak power at the center frequency. This number is divided by the center frequency; i.e., it is a factor of frequencies rather than a difference.

IPR2023-00033
Apple EX1012 Page 32

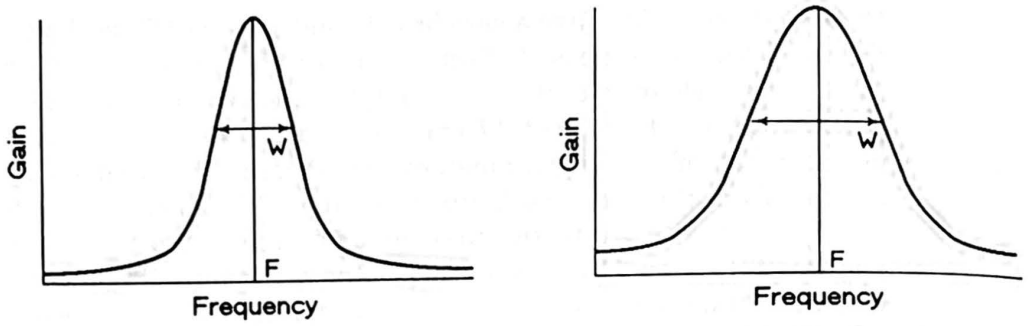


Figure 2.4. A high Q filter, left, and a low Q filter, right. F is the center frequency, and W is its bandwidth.

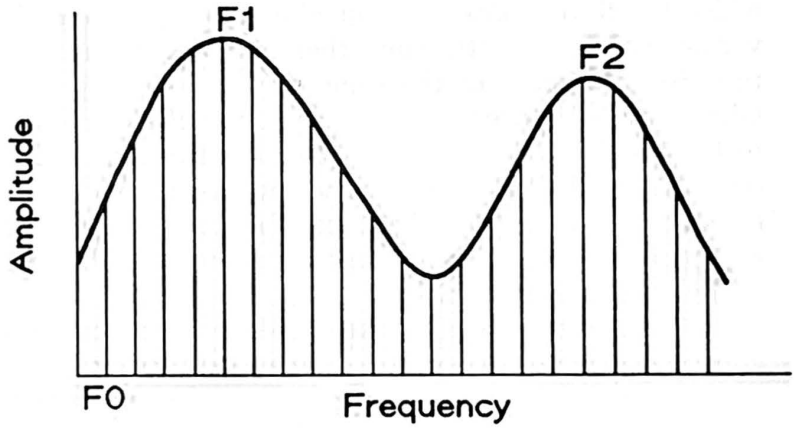


Figure 2.5. Resonances or formants of the vocal tract. The formants, labelled F1 and F2, are peaks in the vocal tract's filter function.

the signal was one half as strong, while directly behind the head the signal decreased by one order of magnitude less than in front of the head.

The noise source is referred to as the **source,** or **excitation,** signal. The manner in which the vocal tract enhances or diminishes the various frequencies of the signal as a function of its physical configuration is called its **transfer function.** A well-defined mathematical relationship allows computation of the signal that results from a source and a transfer function, but a detailed analysis of this relationship is beyond the scope of this book.

THE SPEECH SOUNDS

Each language employs a characteristic set of phonemes or allophones to convey words. These speech sounds are generated by manipulating the components of the vocal tract into specific physical configurations. The set of sounds comprising

IPR2023-00033
Apple EX1012 Page 33

the phonemes of a language must be distinctive because the listener perceives only the sound, not the vocal tract configuration that produces it. The classification of phonemes by the manner in which they are produced is referred to as an **articulatory model** of speech.

The three fundamental articulatory classifications are whether or not the phoneme is **voiced,** the **place** where the sound of the phoneme is made, and the **manner** in which it is made. The remainder of this section describes the various sounds of English according to their articulation.

Vowels

Vowels are sounds produced by vocal cord vibration (voicing) and a relatively open vocal tract, i.e., the lips, tongue, or teeth do not close the passageway. In the English alphabet there are only five letters we call vowels, but in spoken English there are closer to 17 vowel phonemes. Vowels are steady state; after we begin speaking a vowel, we can continue it until we run out of breath. Because vowels are voiced, they are periodic as can be seen in Figure 2.6.

The vowels can be differentiated by the acoustical effect of the position of the tongue and lips during their pronunciation. We can distinguish three degrees of constriction of the vocal tract that indicate how close to the roof of the mouth the hump of the tongue gets (high, medium, and low). We can also differentiate three positions laterally where the hump of the tongue makes the greatest constriction (front, central, and back). In addition, vowels can be *nasalized* by moving the velum to open the air passage leading from the throat to the nose. English does not employ nasalization as a phonemic distinction; however, certain other languages such as French use nasalization as a distinguishing feature to contrast phonemes.

In English the vowels are distinguished solely by the position of the tongue, which changes the lengths of the two primary vocal tract tube sections. In other words, the first two formants form a set of distinctive features for the vowels. Although the typical positions of the formants will vary with gender and among speakers, they can generally be clustered together into regions in a space defined



Figure 2.6. A waveform of the vowel in "but." The display shows amplitude mapped against time, over a 100 millisecond interval. Nine pitch periods can be seen; the frequency of voicing is about 90 Hz.

IPR2023-00033
Apple EX1012 Page 34

by F1 and F2 as illustrated in Figure 2.7. Such a diagram is sometimes called the "vowel triangle" because all of the vowels fit in a space that approximates a triangle in the chosen coordinate system of F1 versus F2.

Some vowels are **diphthongs.** Diphthongs are actually combinations of two vowel sounds with the vocal tract moving from one position to another during their production. The vowel sounds in "bay" and "boy" are examples of diphthongs.

Consonants

Consonants are characterized by constrictions in the vocal tract and can be differentiated by place of closure, manner (type or degree) of closure, and whether they are voiced.

Place refers to the location in the vocal tract of the closure associated with a consonant. **Labial** closure uses both lips as in "p" and "b" sounds. **Labial-dental** closure involves the lower lip and upper teeth as in "f" and "v" sounds. **Alveolar** closure involves the tongue and the gum ridge behind the front teeth as in "n," "d," "s," and "z" sounds. **Palatal** closure uses the tongue on the soft palate or roof of the mouth slightly farther back as in "sh" and "zh" sounds.

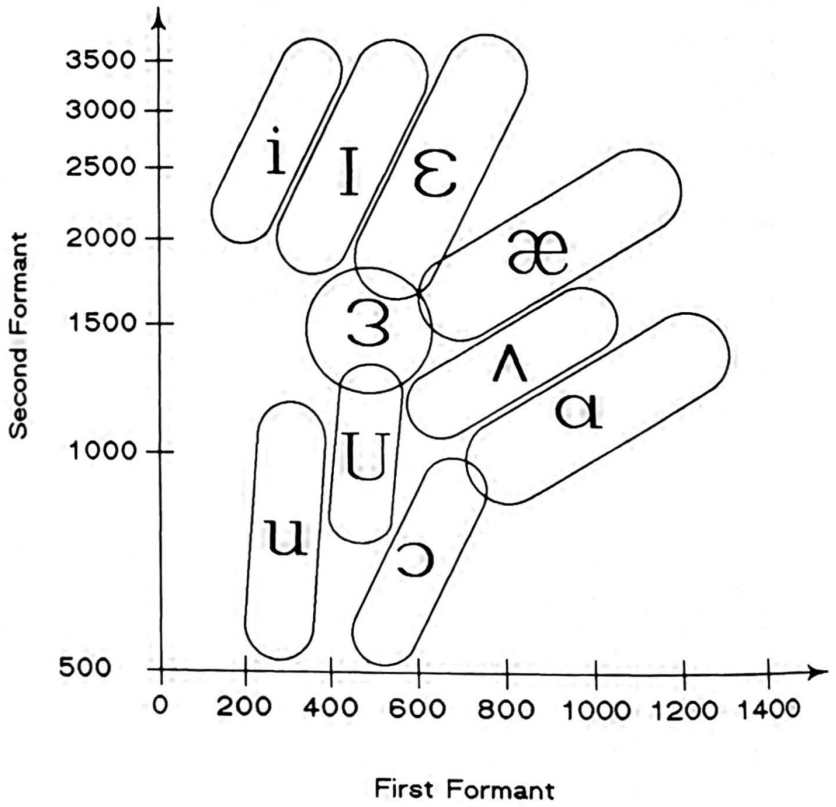


Figure 2.7. The vowel triangle displays English vowels mapped onto a Cartesian space defined by the the first and second formants. The IPA notation from Figure 1.4 is used for labeling.

IPR2023-00033
Apple EX1012 Page 35

Velar closure occurs at the back of the mouth, or hard palate, as in "g" and "k" sounds. Each of these places of closure results in a distinct sound as the shape of the vocal tract changes and consequently the frequencies of resonance vary. The closure divides the vocal tract into cavities in front of and behind the closure; these have different resonant frequencies depending on their size. For example, the position of the third formant is often a strong cue to the identity of a stop consonant.

Closure of the vocal tract is achieved in several different manners, each resulting in a different sound. **Stops** ("t," "b," "p," "d") involve a sudden and total cessation of air flow. Stop consonants are very dynamic. During the closure, no air flows through the vocal tract, resulting in silence. Noise resumes suddenly as the closure is released, and the air pressure from behind the closure may result in **aspiration** noise, such as in the "t" in "top." Stops are also called **plosives,** which focuses on the release rather than the closure of the consonant.

Fricatives ("s," "z," "sh," "zh") involve constriction to the point of producing turbulence and hence noise but not total closure. The partial closure for "s" is alveolar and that for "sh" is palatal. Because closure is incomplete the fricatives result in a continuous sound as air flows by. Sound created by air rushing through a small opening is aperiodic (see Figure 2.8) and dominated by high-frequency components. Combining these two traits, the characteristic acoustic feature of the fricatives is a moderate duration of high-frequency energy.

Nasals ("m," "n," "ng" as in "sing") are produced by closing the oral cavity but opening the velum to the nasal cavity. The nasal cavity absorbs a significant amount of low-frequency energy giving a distinctive cue to nasalization. The three English nasals are all voiced. The nasals are differentiated from each other by the place at which the vocal tract is closed, which is either labial ("m"), palatal ("n"), or velar ("ng").

Most of the consonant types mentioned so far have come in pairs, such as ("d," "t") and ("v," "f"). Both members of each of these pairs is produced with the articulators in the same place but are distinguished by the presence or absence of

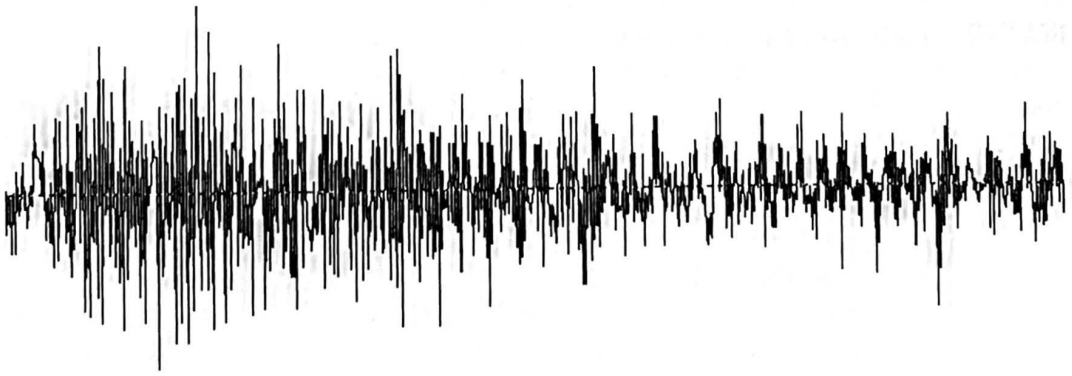


Figure 2.8. The aperiodic waveform of the consonant in "see." This figure also shows 100 milliseconds of speech. No periodicity is evident.

IPR2023-00033
Apple EX1012 Page 36

voicing. Voiced consonants thus have two noise sources: one at the place of articulation and the other at the periodic glottal pulses.

Whether or not a consonant is voiced depends on the type of consonant. In a stop there is no air flow so a stop is considered voiced if voicing resumes soon after the release; otherwise it is unvoiced. Air flow continues during a fricative so voicing is defined by whether the vocal cords are vibrating during the consonant itself.

Liquids and Glides

A small class of special-case phonemes behave much like vowels. The **liquids** ("l," "r") invariably precede a vowel and are dynamic, involving a transition into the vowel. It is sometimes hard to identify "r" as a distinct consonant but rather to observe the effect, or **coloration,** of the vowel caused by the motion of the lips and the rise of the tongue in the middle of the mouth. The **glides** ("w," "y" as in "you") are similar to vowels except that the vocal tract is more constricted than for the vowels. According to another point of view, the glides are simply vowels that do not occur during the energy peak of the syllable.

Acoustic Features of Phonemes

The methods used to articulate the various phonemes of a language must produce sounds with adequate acoustical cues to allow the listener to distinguish between possible phonemes in speech. Some examples of acoustical cues have just been mentioned: the distinctive structure of the first two formants in vowels, the energy at high frequencies that distinguish fricatives, the presence or absence of periodicity due to voicing, and the absorbing effect of the nasal cavity.

The listener can select from a multitude of cues occurring simultaneously to differentiate phonemes. In fluent speech the articulators often move only partially in the direction of target vocal tract configurations described above for each phoneme. The acoustical effect is still sufficiently pronounced to allow the listener to detect at least some of the cues and identify each phoneme.

HEARING

Sound arrives at our ears as variations in air pressure. We can hear vibrations in the range of approximately 20 Hz up to 15 or 20 kHz; this figure varies among individuals, and the range decreases as we age. This sensory stimulus triggers neural activity through a complex but versatile mechanical transformation in the ear. There is a chain of processes whereby the physical sound from its source causes the auditory event of "hearing." In turn, these neural firings stimulate perceptual processing at higher levels of the brain. Although little is known about such processing from direct neural evidence, the domain of **psychoacoustics** studies our perception of sound by observing subjects' responses to acoustic stimuli.

IPR2023-00033
Apple EX1012 Page 37

Auditory System

The complement to the vocal tract is the auditory system. Although primarily designed to turn air pressure variations into corresponding neural signals, the ear also contains the **vestibular organs** that are used to maintain physical balance; we are not concerned with them here. The ear can be divided into three distinct sections. The outer ear directs sound toward the eardrum, the middle ear converts the pressure variations of sound into mechanical motion, and the inner ear converts this motion to electrical signals in the auditory neurons. The ear is shown in Figure 2.9.

The **outer ear** includes the **pinna** and the **ear canal,** which leads to the eardrum. The pinna consists of the fleshy protrusion on the side of the head and is what we usually refer to when we use the term "ear" (as in "Van Gogh cut off his ear"). The pinna with its various folds of cartilage around the ear opening serves primarily as a protective mechanism. The pinna provides some amplification of the sound by focusing it into the ear canal in much the same way that we may cup our hands behind our ears to better hear quiet sounds. It is directional at high frequencies and is used as a localization aid to find a sound source because it makes the ear more sensitive to sounds coming from in front of rather than behind the listener. In certain animals such as owls, the outer ear occupies a much larger surface area with respect to the head and is more fundamental to localization.

The ear canal is a tube about 1 cm wide by 2.5 cm long leading to the middle ear. It has a resonance of about 3000 Hz and therefore amplifies sound at this frequency. The length of the ear canal shields the middle ear from physical injury if the side of the head is struck.

The **middle ear** provides the linkage between the outer and inner ear; its function is to effectively convert variations in air pressure to mechanical motion of the liquid inside the inner ear. The **eardrum,** or **tympanic membrane,** covers the interior end of the ear canal. Vibrations in the air cause the eardrum to vibrate; thus it produces physical motion from the sound in the air. The middle ear is filled with air. The **eustachean tube** runs between the middle ear and the throat. When we yawn, the eustachean tube opens, allowing the air pressure to equalize across the eardrum. As experienced when descending rapidly in an airplane, this pressure balance is essential for effective transfer of sound across the eardrum.

The inner ear is filled with water. There is a large impedance mismatch between air- and water-like fluids. Water is much denser than air so air does not cause significant displacement in the liquid that it impinges; even a strong wind on a pond does not penetrate much deeper than to form surface ripples because of this mismatch. A series of three **ossicular bones** (the **malleus, incus,** and **stapes**) provide mechanical advantage (leverage) from the ear drum to the **oval window,** a membrane on the surface of the inner ear. As the eardrum is 17 times as large as the oval window, this difference in area provides further amplification. The size difference plus the mechanical advantage of the ossicular bones combine to provide a 22:1 amplification and thereby an impedance match for efficient transfer of acoustical energy to the inner ear.

IPR2023-00033
Apple EX1012 Page 38

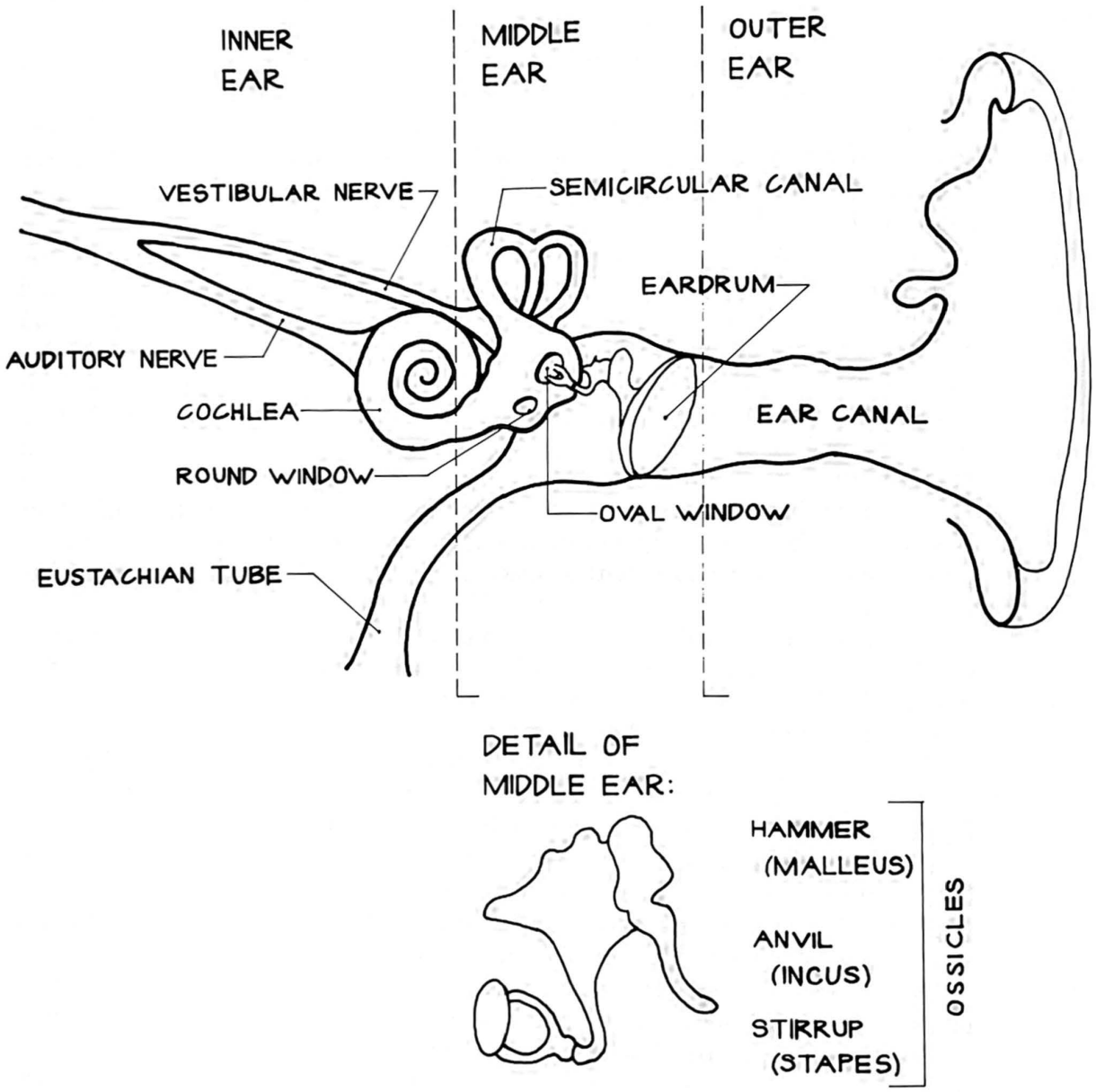


Figure 2.9. The ear.

The middle ear has several different modes of vibration that provide improved dynamic range without distortion (the goal of all good audio amplifiers). For loud sounds, the vibration mode is changed from pumping to a rotational action. In addition, for very loud sounds a reflex occurs in the muscles of the middle ear to damp the vibration and thereby protect the sensitive inner ear.

IPR2023-00033
Apple EX1012 Page 39

The **inner ear** is the transducer from mechanical to electrical energy. At its center is the **cochlea,** a spiral-shaped chamber looking somewhat like a snail shell and making 2⅝ turns. Sound-induced waves in the liquid medium of the cochlea cause vibration on the **basilar membrane.** Different portions of this tapering membrane vibrate in response to sound energy at specific frequencies. At the basal end it is thin and stiff, while at its apex it is flexible and massive.

The frequency dependent vibration of the basilar membrane induces motion in the microscopic hairs that penetrate the membrane. These hairs are linked to neurons to the brain that produce neural firings in response to the stimulus of the hairs' bending. Each neuron is connected to a small number of hairs in a particular location on the basilar membrane, and the basilar membrane responds to different frequencies along its length so the firing of a neuron corresponds to the presence of sound of the appropriate frequency. Because of this relationship between neural firings and basilar membrane vibration, an individual neuron is not very responsive to frequencies much higher or lower than its preferred one. Measurements of neural responses indicate that each acts as a bandpass filter and that all the neurons and associated hair cells exhibit a fairly constant Q.

Neural activity is highest at the onset of a sound and decays rapidly at first and then more slowly. This is known as **adaptation;** we are more sensitive to changing sounds than continual tones. Neurons fire more rapidly for louder sounds, reaching peak rates of up to 1000 firings per second for short periods. At frequencies below about 1 kHz, neurons tend to fire in phase with the vibration of the basilar membrane, i.e., every cycle of membrane motion induces one neural spike.

The ear's response to a sound is a series of neural spikes from many neurons simultaneously. The firing pattern of any particular neuron is a function of amplitude of the acoustic energy (at the basilar membrane) within the frequency range to which that neuron is sensitive. The pattern adapts over a short period of time and may be at least partially in phase with the acoustic signal. The neural firings are transmitted to a number of stages of processing in the central auditory system. At several points along the way are sites that are particularly sensitive to time or amplitude differences between signals arriving from each ear. Such differences are central to our ability to locate a sound spatially.

Localization of Sounds

When we hear a sound, we perceive it as coming from some location in space outside of our head; this is known as **localization.** Much of our ability to localize sound depends on the differences in the sound that arrives at each of our two ears from a single source. These differences are due to the positions of the ears and hence their different distances from the sound source as well as the tendency of the head to partially block sound coming to an ear from the opposite side of the head.

We can localize a sound source within about 4 degrees in front of the head, 6 degrees behind the head, and only within 10 degrees at the sides. Localization is very frequency dependent and we are most sensitive to sounds at around 800 Hz.

IPR2023-00033
Apple EX1012 Page 40

Our hearing is about 100 times less sensitive to position than our visual system, but we can hear sounds behind our heads. We have a strong reflex to turn and look at a sound that occurs behind our heads. This moves the source of the sound into our field of view, which could aid survival, and places the source where we can better localize it. The motion of the head also provides cues to enhance localization as we seem to be especially sensitive to differences in location.

If a sound source is off to either side of our head, it takes the sound longer to reach the ear on the other side as it is further away (see Figure 2.10). In other words, when a sound begins one ear hears it before the other. But more importantly, while the sound continues the air vibrations will arrive at each ear out of phase with the other ear; this is defined as the **interaural phase difference.** The human head has a thickness of 21 or 22 centimeters, which means that it can take sound up to about 600 microseconds longer to reach the further ear. This phase difference becomes confused when it is greater than the wavelength of the sound so phase is most effective for localizing sound below 1500 Hz.

When a sound is off to one side, the mass of the head also blocks the direct path from the sound source to the opposite ear. This results in an **interaural intensity difference;** the sound is louder at the ear that has an unobscured path. This effect is most pronounced at frequencies above 10 kHz, as the head more effectively blocks small wavelengths.

The differences in phase and intensity of the sound arriving at each ear can help localize it to the left or right side but leave confusion as to whether the sound is in front of or behind the head. A sound in front of and to the left of the listener creates the same interaural phase and intensity differences as a sound the same distance behind and also to the left. The shape of the pinna interacts with incom-

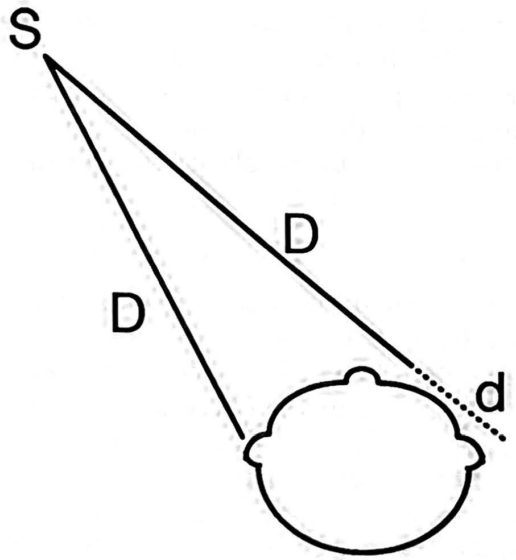


Figure 2.10. A sound on the left side of the head arrives at the right ear later than the left ear because it must travel further.

IPR2023-00033
Apple EX1012 Page 41

ing sound and provides front-back cues. Its amplification effect provides directionality through increased gain for sounds from the front. In addition, the folds of the pinna act as a complex filter enhancing some frequencies; this filtering is different for sounds from the front than from the back providing further front/back cues.

Psychoacoustics

So far we have described the process whereby sound energy is converted to neural signals; we have not yet described how the listener perceives the sound characterized by such signals. Although more is known about higher-level processing of sound in the brain than we have mentioned in the brief overview above, there is still not sufficient knowledge to describe a model of such higher auditory processing. However, researchers can build an understanding of how we perceive sounds by presenting acoustical stimuli to subjects and observing their responses. The field of psychoacoustics attempts to quantify perception of pitch, loudness, and position of a sound source.

The ear responds to a wide but limited range of sound loudness. Below a certain level, sound cannot be perceived; above a much higher level, sound causes pain and eventually damage to the ear. Loudness is measured as **sound pressure level** (SPL) in units of **decibels** (dB). The decibel is a logarithmic scale; for power or energy, a signal that is X times greater than another is represented as $10 \log_{10} X$ dB. A sound twice as loud is 3 dB higher, a sound 10 times as loud is 10 dB higher, and a sound 100 times as loud is 20 dB higher than the other. The reference level for SPL is 0.0002 dyne/cm^2; this corresponds to 0 dB.

The ear is sensitive to frequencies of less than 20 Hz to approximately 18 kHz with variations in individuals and decreased sensitivity with age. We are most sensitive to frequencies in the range of 1 kHz to 5 kHz. At these frequencies, the threshold of hearing is 0 dB and the threshold of physical feeling in the ear is about 120 dB with a comfortable hearing range of about 100 dB. In other words, the loudest sounds that we hear may have ten billion times the energy as the quietest sounds. Most speech that we listen to is in the range of 30 to 80 dB SPL.

Perceived loudness is rather nonlinear with respect to frequency. A relationship known as the Fletcher-Munson curves [Fletcher and Munson 1933] maps equal input energy against frequency in equal loudness contours. These curves are rather flat around 1 kHz (which is usually used as a reference frequency); in this range perceived loudness is independent of frequency. Our sensitivity drops quickly below 500 Hz and above about 3000 Hz; thus it is natural for most speech information contributing to intelligibility to be within this frequency range.

The temporal resolution of our hearing is crucial to understanding speech perception. Brief sounds must be separated by several milliseconds in order to be distinguished, but in order to discern their order about 17 milliseconds difference is necessary. Due to the firing patterns of neurons at initial excitation, detecting details of multiple acoustic events requires about 20 milliseconds of averaging. To note the order in a series of short sounds, each sound must be between 100 and 200 milliseconds long, although this number decreases when the sounds have

IPR2023-00033
Apple EX1012 Page 42

gradual onsets and offsets.[5] The prevalence of such transitions in speech may explain how we can perceive 10 to 15 phonemes per second in ordinary speech.

Pitch is the perceived frequency of a sound, which usually corresponds closely to the fundamental frequency, or F0, of a sound with a simple periodic excitation. We are sensitive to *relative* differences of pitch rather than absolute differences. A 5 Hz difference is much more significant with respect to a 100 Hz signal, than to a 1 kHz signal. An **octave** is a doubling of frequency. We are equally sensitive to octave differences in frequencies regardless of the absolute frequency; this means that the perceptual difference between 100 Hz and 200 Hz is the same as between 200 Hz and 400 Hz or between 400 Hz and 800 Hz.

The occurrence of a second sound interfering with an initial sound is known as **masking.** When listening to a steady sound at a particular frequency, the listener is not able to perceive the addition of a lower energy second sound at nearly the same frequency. These frequencies need not be identical; a sound of a given frequency interferes with sounds of similar frequencies over a range called a **critical band.** Measuring critical bands reveals how our sensitivity to pitch varies over the range of perceptible frequencies. A pitch scale of **barks** measures frequencies in units of critical bands. A similar scale, the **mel** scale, is approximately linear below 1 kHz and logarithmic above this.

Sounds need not be presented simultaneously to exhibit masking behavior. **Temporal masking** occurs when a sound masks the sound succeeding it (**forward** masking) or the sound preceding it (**backward** masking). Masking influences the perception of phonemes in a sequence as a large amount of energy in a particular frequency band of one phoneme will mask perception of energy in that band for the neighboring phonemes.

SUMMARY

This chapter has introduced the basic mechanism for production and perception of speech sounds. Speech is produced by sound originating at various locations in the vocal tract and is modified by the configuration of the remainder of the vocal tract through which the sound is transmitted. The articulators, the organs of speech, move so as to make the range of sounds reflected by the various classes of phonemes.

Vowels are steady-state sounds produced by an unobstructed vocal tract and vibrating vocal cords. They can be distinguished by their first two resonances, or formants, the frequencies of which are determined primarily by the position of the tongue. Consonants are usually more dynamic; they can be characterized according to the place and manner of articulation. The positions of the articula-

[5]The onset is the time during which the sound is just starting up from silence to its steady state form. The offset is the converse phenomenon as the sound trails off to silence. Many sounds, especially those of speech, start up and later decay over at least several pitch periods.

IPR2023-00033
Apple EX1012 Page 43

tors during consonant production create spectra characteristic of each phoneme; each phoneme has multiple cues as to its identity.

Sound consists of vibrations in a medium (usually air) at a frequency we can hear. The variations in sound pressure are focused by the outer ear and cause mechanical motion in the middle ear. This mechanical motion results in waves in the liquid in the inner ear, which causes neurons to fire in response. Which neurons fire and the pattern of their firings is dependent on the spectral and temporal characteristics of the sound. We localize a sound based on interaural phase and intensity differences. Each ear receives a slightly different sound signal due to the position of the ears on either side of the head. The pinna enables us to distinguish between sounds in front and behind.

This chapter also discussed several topics in psychoacoustics and the frequency and temporal responses of the auditory system; what we perceive with our sense of hearing is as much due to the characteristics of our auditory system as to the qualities of the sounds themselves. These factors influence our perception of speech in particular as well as all sounds.

The next chapter explores methods to digitally capture, encode, and compress the speech signal for computer storage, analysis, or transmission. Data compression techniques for speech take advantage of temporal and frequency characteristics of the speech signal as well as the sensitivity of the auditory system to reduce the data rate without loss of intelligibility or perceived quality of the reconstructed signal.

FURTHER READING

A concise overview of the human speech system can be found in Denes and Pinson and a more analytic and rigorous one in Flanagan (1972). Ladefoged offers an excellent treatment of phonetics. O'Shaughnessy covers many issues in speech production and hearing as a prelude to describing speech technologies. Handel offers an excellent overview of the perception of speech and music, while Yost is a more introductory text with superb illustrations.

IPR2023-00033
Apple EX1012 Page 44

7

Speech Recognition

This chapter is about the technologies used to allow computers to recognize the words in human speech. It describes the basic components of all speech recognition systems and illustrates these with an example of a simple recognizer typical of several inexpensive commercial products. After this discussion of basic recognition, the chapter details a larger range of features that can be used to differentiate various recognizers according to the styles of speech interactions they support. Several more advanced recognition techniques are then introduced followed by brief descriptions of selected research projects in large vocabulary and speaker independent word recognition.

BASIC RECOGNIZER COMPONENTS

There are three basic components of any speech recognizer.

1. A speech **representation** that is computationally efficient for pattern matching. The representation is the form into which the recognizer converts the speech signal before it begins analysis to identify words. Typical representations include the output of a bank of filters (similar to a spectrogram), Linear Predictive Coding (LPC) coefficients,[1] and zero crossings of the speech waveform. Recognizers of

[1]See Chapter 3.

IPR2023-00033
Apple EX1012 Page 45

increased sophistication incorporate more abstract representations of speech such as phonemes or distinctive spectral features. Hidden Markov Models, described later in this chapter, are a statistical representation based on the various ways words or phonemes may be pronounced.

2. A set of **templates** or **models,** which are descriptions of each word to be recognized, in the representation of speech used by the recognizer. The templates describe the words in the recognizer's vocabulary, i.e., those words that the recognizer can identify. They are reference models against which an input can be compared to determine what was spoken.
3. A **pattern matching** algorithm to determine which template is most similar to a specimen of speech input. This element of the speech recognizer must determine word boundaries, locate the most similar template, and decide whether the difference between the input and the selected template is minor enough to accept the word. In very large vocabulary recognizers, the pattern matching technique usually includes access to more advanced knowledge of language such as syntax and semantics, which constrain how words can be combined into sentences.

To identify a word the recognizer must capture incoming speech and convert it to the chosen internal representation (see Figure 7.1). The pattern matcher then selects the template that most closely matches the input or rejects the utterance if no template is a close enough match.

SIMPLE RECOGNIZER

This section describes a simple recognizer typical of many inexpensive, commercially available devices. This recognizer is designed to identify a small number of

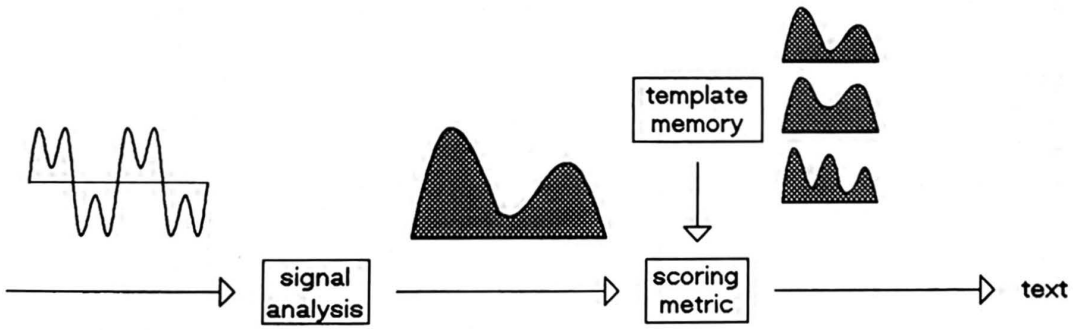


Figure 7.1. The functional elements of a speech recognizer. The user's speech is digitized and converted to the recognizer's internal representation. The captured speech is then compared with the words stored in the recognizer's template memory. The pattern matching algorithm then determines which template, if any, is the closest match.

IPR2023-00033
Apple EX1012 Page 46

words spoken in isolation by a specific individual. The purpose of the description that follows is not to provide details of a particular product but rather to offer a sample implementation of the basic components described previously.

Representation

A simple recognizer digitizes incoming audio by means of a codec[2] and then uses a digital signal processing algorithm to extract frames of Linear Predictive Coding (LPC) coefficients every 20 milliseconds. The LPC coefficients are a concise representation of the articulatory characteristics of speech as they capture both the resonances of the vocal tract as well as its voicing characteristics. The 20 millisecond sampling interval results in 50 LPC frames per second, providing significant data reduction to simplify later pattern-matching.

Templates

Templates are gathered by prompting the user from a word list and then saving a set of LPC frames for each word generated as just described. To build a template, the recognizer must determine when the user begins and finishes speaking to know which LPC frames to include in the template. Since each LPC frame or set of coefficients in one 20 millisecond sampling period includes an energy value, this task can be accomplished easily. As shown in Figure 7.2, once the audio exceeds the threshold of background noise, LPC frames are saved until the audio drops below the threshold. It is necessary to wait until the signal has dropped below the threshold for a short period of time as silence or low energy can occur within a word, such as at stop consonants.

In the case of this simple recognizer, templates are trained by a user saying each word once. More sophisticated recognizers may take multiple specimens of a

[2]See Chapter 3.

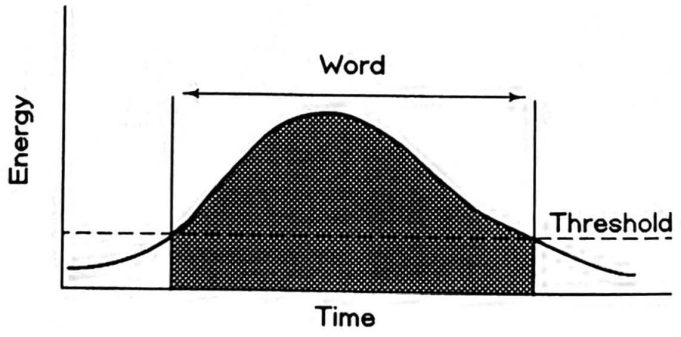


Figure 7.2. Energy thresholds can be used to find word boundaries if the words are spoken in isolation.

IPR2023-00033
Apple EX1012 Page 47

word to build a more robust template. Once templates are trained successfully, the user can save them as a disk file on the host computer and reload them before each session, to avoid template training each time the recognizer is used.

This recognizer also allows retraining of a single template, in which case the new LPC frames replace the old. Retraining is useful if the vocabulary changes, if the original training was poor, or if the user decides that a different pronunciation of the word is more comfortable. Retraining may also be necessary when the user has a cold, due to an obstructed nasal cavity.

Pattern Matching

Word recognition requires the detection and classification of speech input. Just as with template creation, a word is captured by detecting that the audio input exceeds the background noise threshold and then converting the input to LPC frames until the input again drops to the background level. The input word is then compared with each template; each frame or set of LPC parameters of the input is compared to the corresponding frame of the template. The frame-by-frame error is the sum of the differences between each of the dozen or so LPC parameters; the error for the word is the sum of the errors for each frame. The template with the smallest word error most closely matches the audio input. If the error exceeds a rejection threshold, a failed recognition attempt is reported. Otherwise the recognizer reports the word corresponding to the closest template.

Both the templates and the captured audio input are multidimensional vectors, with one degree-of-freedom per LPC parameter and one for time. But to illustrate pattern matching let us assume that the templates consist of only a pair of values that display on two axes. One parameter is assigned to the X-axis and the other parameter maps to the Y-axis so that each template occupies a point whose coordinates are the two parameters (see Figure 7.3). Speech input is converted into this representation, i.e., a pair of parameters specifying a point in the same parameter space as the templates. The closest template is the nearest neighbor template to the point representing the input. The four templates divide the space into four regions; any pair of parameters in each region most closely matches the template in that region.

Even the simplest of speech recognizers can improve its accuracy by employing a better pattern matching algorithm. Two additional refinements, illustrated in Figure 7.4, contribute to this recognition improvement. The first refinement is that the input cannot exceed a threshold distance (**r** in the figure) from the nearest template. When the audio input is further from the nearest template than this distance, a rejection error is reported. This is necessary to avoid inadvertently accepting noise, breath sounds, or a mistaken utterance as one of the acceptable words in the recognizer's vocabulary.

The second refinement is a requirement that the input word maps significantly closer to the best match template than any other. If a point is very nearly the same distance from two or more templates, the recognizer is confident that one of these was spoken but uncertain which particular word was spoken. This is depicted in Figure 7.4 as the shaded regions around the lines partitioning the

IPR2023-00033
Apple EX1012 Page 48

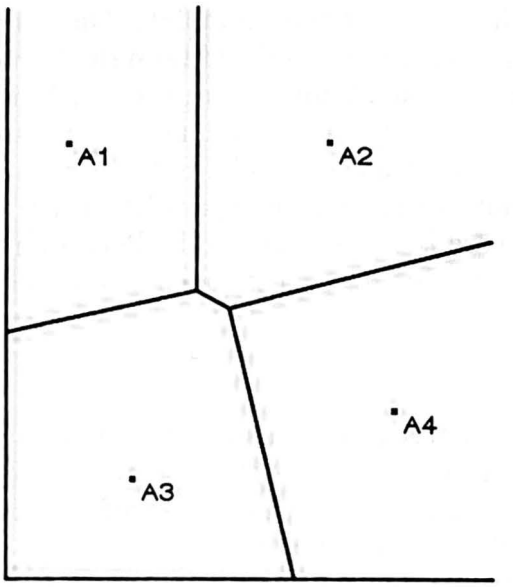


Figure 7.3. To illustrate pattern matching decisions, assume that each template represents two degrees of freedom defining a Cartesian coordinate space. Four templates occupy four points in this space.

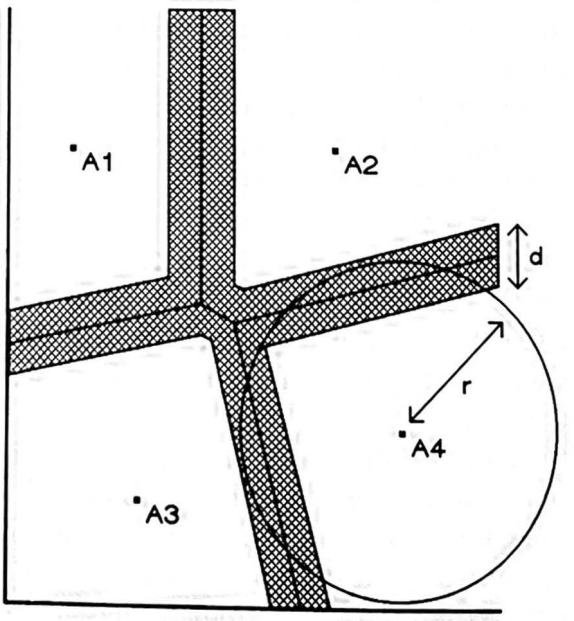


Figure 7.4. A more refined classifier has upper bounds **r** on the distance between an input specimen and a chosen template. It will also reject values which are closer to one template but also close to another, i.e., within distance **d.**

IPR2023-00033
Apple EX1012 Page 49

space between neighbors; the distance **d** defines the minimal difference required to differentiate candidate choices. Points within the shaded region are nearly the same distance from each of the two closest templates and will also cause a rejection error.

This section described an extremely simple recognizer that identifies only isolated words. The speech recognizer's simplicity is due in part to its unsophisticated processing of acoustical information: It does not extract significant features but merely considers LPC frames in isolation. Because each template may contain many frames, the pattern matcher has to make a large number of frame-by-frame difference calculations, making it computationally difficult to recognize a large number of words. More sophisticated recognizers detect and classify key acoustic features of an utterance and use this for the internal representation.

Additionally, the pattern matcher just described makes no provision for templates of different lengths and, more importantly, assumes that the input utterance will be of the same length as the correct template. If acoustic boundary detection operates slightly differently from word to word or if the words are spoken at different speeds, this pattern matcher will likely fail. A technique known as **dynamic time warping** or **dynamic programming** (discussed later in this chapter) is often used to compensate for time differences.

CLASSES OF RECOGNIZERS

Recognizers vary widely in their functionality. In addition to performance, other distinctions between recognition techniques are important in selecting a recognizer for a particular application. These distinctions are summarized in Figure 7.5, which displays a three-dimensional space representing the range of possible recognizer configurations. Some of these recognizers are commercially available and have been in use for some time, whereas others are the subject of ongoing research and are just beginning to emerge from research laboratories.

Who Can Use the Recognizer?

A **speaker-independent** recognizer is designed to recognize anyone, while a **speaker-dependent** recognizer can be expected to understand only a single particular speaker. A **speaker-adaptive** recognizer functions to an extent as a combination of the two; it accommodates a new user without requiring that the user train every word in the vocabulary.

Speaker-independent recognition is more difficult than speaker-dependent recognition because we all speak the same words slightly differently. Although we usually understand each other, this variability in pronunciation and voice quality among talkers plays havoc with the pattern matching algorithms of simple recognizers. Speaker-independent recognition requires more elaborate template generation and a clustering technique to identify various ways a particular word may be pronounced.

IPR2023-00033
Apple EX1012 Page 50

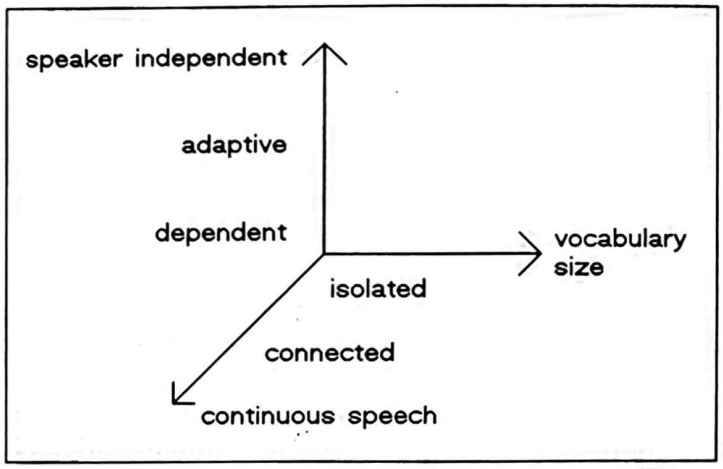


Figure 7.5. A three-dimensional space defined by the different functionalities provided by a recognizer.

This process of classifying the acceptable variations in word pronunciations is the equivalent of the training method just described for the simple illustrative recognizer. Since speaker-independent recognition is such a difficult task, most speaker-independent recognizers are limited to small vocabularies (10 to 20 words) and recognize only single words spoken in isolation. Several commercially available speaker-independent recognizers accept the digits "0" through "9" plus "yes" and "no." Speaker-independent recognition for a larger vocabulary or with words spoken together must rely on additional constraints such as a heavily restricted grammar to achieve adequate recognition performance.

An individual pronounces a single word in a much more consistent fashion than that of a number of different people; this is why speaker-dependent recognition is less difficult. Most speaker-dependent recognizers use a training method similar to the one described for the simple recognizer in the previous section. Multiple templates may be trained and merged together to give an average pronunciation for each word in the vocabulary.

A speaker-adaptive recognizer learns to adapt to a particular speaker either by calibrating itself to a known sample of that person's speech or by incorporating user feedback when it recognizes a word correctly or incorrectly. Such a recognizer has a basic acoustic model for each word that it refines to fit a particular speaker. One common technique is to have the user read a passage of training data; the recognizer can then determine how this particular person speaks.

If a large vocabulary size is needed, a recognizer must be either speaker-adaptive or independent because training many templates is tedious and time consuming. Some adaptive recognizers require many hours of computer time to derive the user model from the training data, but once the data has been acquired the user is freed from the task. The resulting user model can be saved in a host computer and downloaded later, similar to the set of word templates for the speaker-dependent recognizer.

IPR2023-00033
Apple EX1012 Page 51

Speaking Style: Connected or Isolated Words?

A **discrete speech** recognizer requires pauses between each word to identify word boundaries. A **connected speech** recognizer is designed to recognize a short series of words spoken together as a phrase. A **continuous speech** recognizer is capable of identifying words in a long string of ordinary speech without the talker pausing between groups of words. A **keyword spotting** recognizer can locate a few words in the midst of any amount of speech.

"Connected speech" for recognition purposes is not natural speech. Users of a connected speech recognizer can speak a few words or perhaps a whole sentence at a time but then must pause to let the recognizer catch up. Users must also speak distinctly. Pausing after each sentence provides reliable boundary points for the first and last word, which facilitates recognition of the entire phrase. But we do not pause between sentences in fluent speech. Human listeners can keep up, and this is the goal of continuous speech recognition, which may become a reality as computer speeds increase.

Keyword spotting searches for a small number of words in a stream of continuous speech. For example, a keyword recognizer might be designed to recognize the digits from an utterance such as "The number is three five seven . . . um . . . four one, please." Successful keyword spotting is comparatively new [Wilpon *et al.* 1990] and more difficult than simply separating speech from non-speech background noise.

Most currently-available commercial products recognize isolated or connected speech. Connected speech is much faster to use because it eliminates the need for unnatural pauses and requires less attention to speaking style on the part of the user. But because connected recognition is more difficult, it may manifest higher error rates, which could outweigh the speed advantage. It can also be argued that speaking discretely is more effective because it requires the user to keep in mind the need to speak clearly from a limited vocabulary while using speech recognition [Biermann *et al.* 1985].

Connected word recognition is much more difficult than isolated word recognition for several reasons.

- Coarticulation changes the pronunciation of a word as a function of its neighbors. Initial and final syllables are particularly subject to modification. Words spoken in isolation do not suffer from coarticulation effects between words.
- It is difficult to find word boundaries reliably from within fluent speech. There is no pause between words, nor is there a significant decrease in speech energy at word boundaries; low energy syllables containing stop consonants are often more discernible than word boundaries.
- The probability of error increases with the number of words in an utterance. If the first word is incorrectly matched against a template which is either too long or too short, then the data to be analyzed for the second word will be incorrect, making the error propagate to subsequent words.

IPR2023-00033
Apple EX1012 Page 52

Because of these factors many current applications of speech recognition are based on isolated word recognizers.

Vocabulary Size

Another criterion by which to differentiate recognizers is vocabulary size, which can be grossly categorized as small, medium, or large. Small vocabulary recognizers with less than 200 words have been available for some time. Medium size recognizers (200 to 5000 words) are being developed, usually based on the same algorithms used for smaller vocabulary systems but running on faster hardware. Large vocabulary recognizers aim for the 5000 to 10,000 word level. Much ordinary office language could be handled with a vocabulary of this breadth, which marks the range being aimed for in "listening typewriters" designed to accommodate dictation of business correspondence.

Several issues conspire to complicate the recognition of large vocabularies. One factor is the computational power required for the pattern matching algorithm. The input must be compared with each template, so classification time is a function of the number of templates, i.e., vocabulary size. The requirement of an acceptable response time therefore puts an upper limit on vocabulary size. As microprocessor speeds increase, this computational limit will become less of an issue. Some search-pruning techniques can also be employed to quickly remove candidate templates from the search if they are obviously poor choices, such as if they are much too long or short. If the recognizer employs a grammar, it can use this knowledge of how words can be combined to eliminate syntactically incorrect candidates at each point in the search.

A more serious limitation to vocabulary size is simply that as the number of words increases, it is more likely that the recognizer will find some of them similar. To return to the two parameter template model discussed earlier, compare Figure 7.3 with Figure 7.6. As the number of templates increases, the average distance between them decreases, allowing for a smaller margin of difference in pronunciation between the input sample and its associated template.

Decreased distance between templates is confounded by the fact that as templates or words are added to the recognizer's vocabulary, they are not uniformly distributed throughout the recognizer's representation space. In fact, some words in the vocabulary are likely to sound so similar that to distinguish among them on acoustic evidence alone is extremely difficult. For example, if a vocabulary consisted of the words, "sun," "moon," and "stars," we might expect that distinguishing which word was spoken to be easy. But if we add "Venus," "Earth," and "Mars" to the vocabulary, we might anticipate confusion between "stars" and "Mars."[3]

[3]This is a simplistic example. The strong fricative and stop at the beginning of "stars" should be easily differentiated from the nasal of "Mars" if suitable acoustic features are used for the recognizer's representation. It is hard to predict which words will sound similar without intimate knowledge of the recognizer's internal speech representation.

IPR2023-00033
Apple EX1012 Page 53

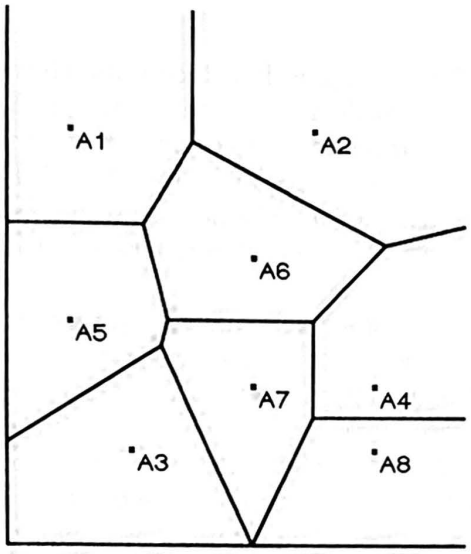


Figure 7.6. As the number of words increases, the mean distance between templates decreases.

ADVANCED RECOGNITION TECHNIQUES

This section explores several techniques to enhance speech recognition. The simplest recognizers, supporting a very small speaker-dependent vocabulary of isolated words, occupy the region near the origin of the three-dimensional model depicted earlier in Figure 7.5. In moving away from the origin in any direction, recognition errors are more likely as the process becomes more complicated. But a large vocabulary, continuous speech, and speaker independence are precisely those attributes that make recognition more widely useful.

Unless users can be convinced to change their habits to speak more clearly and consistently, recognizers must be improved. More advanced recognizers generally employ one of two techniques to make pattern matching more powerful for dealing with variations in speech patterns. These techniques, **Dynamic Time Warping** and **Hidden Markov Models,** are the topics of the next two sections. The descriptions that follow provide an overview of the two techniques; the curious reader is encouraged to consult the references for a more rigorous description of the algorithms. A third approach to managing speech pattern variation uses **neural networks** for speech recognition; this approach is far less developed than the previous two and still speculative.

Data reduction can facilitate pattern matching by minimizing its computational requirements. **Vector Quantization** is a technique employed by many recognizers to capture important variations in speech parameters without overloading the classifier—so it will be described in this section as well. Finally, the last part of this section considers how nonspeech evidence could be used to improve large vocabulary speech recognition.

IPR2023-00033
Apple EX1012 Page 54

Dynamic Time Warping

Dynamic Time Warping is a technique that compensates for variability in the rate at which words are spoken. Dynamic Time Warping (DTW) was developed primarily as a mechanism to compensate for variable word duration in connected speech recognition [Sakoe and Chiba 1978]. This method can also help determine word boundaries when an unknown number of words are spoken together. DTW is based on a more general computational technique known as **dynamic programming.**

The duration of spoken words is quite variable, especially if we compare connected speech with the isolated words that may have been used for training a recognizer. Coarticulation may shorten words by combining their boundary syllables with the preceding or subsequent words. Words spoken in isolation (citation form) are longer and often more clearly enunciated. The stress pattern of the sentence lengthens some syllables, as stressed syllables are longer than unstressed syllables. Such changes in length are not linear; every phoneme is not lengthened by the same scale factor. DTW compensates for such nonlinear phenomena.

Consider the simple recognizer discussed earlier in this chapter. It computes an error between the input speech and a template by computing the frame-by-frame difference between the two. But if the talker speaks at a different rate, successive frames of input may not align with the same phoneme in the template, giving a deceptively large error. For example, if the word "fast" is trained but the talker lengthens the vowel ("faaast") during recognition, some frames of the lengthened vowel would be scored against the frames corresponding to the unvoiced "s" in the template. DTW detects that successive "a" frames of the input match the fewer "a" frames of the template better than the "s" frames that follow, and it computes an error based on the selected matchup.

DTW operates by selecting which frames of the reference template best match each frame of the input such that the resulting error between them is minimized. By allowing multiple frames of one to be matched against a single repeated frame of the other, DTW can compress or expand relative time. Because this decision is made on a frame-by-frame basis, the time-scaling is local; DTW may compress one portion of the input and expand another if necessary.

DTW provides a mapping between the sample and a reference template such that the error when the two are compared is minimized. This mapping defines a path between sample and reference frames; for each frame of the sample it specifies the template frame best matched to the next sample frame. In Figure 7.7, the *nth* sample frame is compared to the *mth* reference frame. If the sample is spoken more quickly, then multiple reference frames correspond to a single template frame; this case is indicated by the vertical path, which allows the *nth* sample frame to also match the *m + 1st* reference frame. If the reference and sample are proceeding at the same rate, the *m + 1st* reference frame will match the *n + 1st* sample frame forming a path at a 45-degree angle (slope = 1). Finally, if the sample is spoken more slowly than the reference, multiple sample frames must be compared to a single reference frame as indicated by the horizontal line.

Figure 7.8 shows how these frame-by-frame decisions combine to produce a path for comparing the sample to the reference. The DTW algorithm computes

IPR2023-00033
Apple EX1012 Page 55

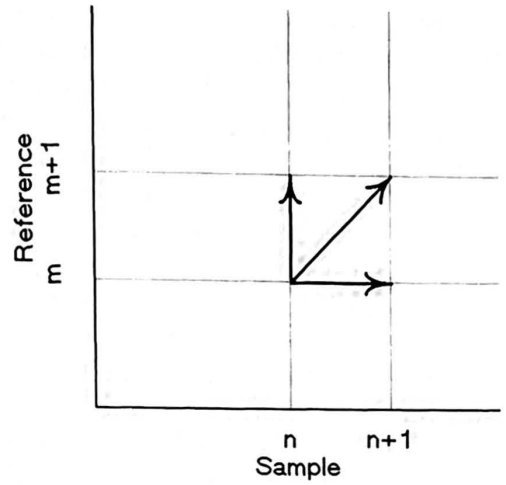


Figure 7.7. Dynamic Time Warping defines a path between frames of a sample utterance and a template such that the frame-by-frame error between the two is minimized. If sample point **m** matches reference point **m,** then reference point **m + 1** may match either sample point **n** or **n + 1.**

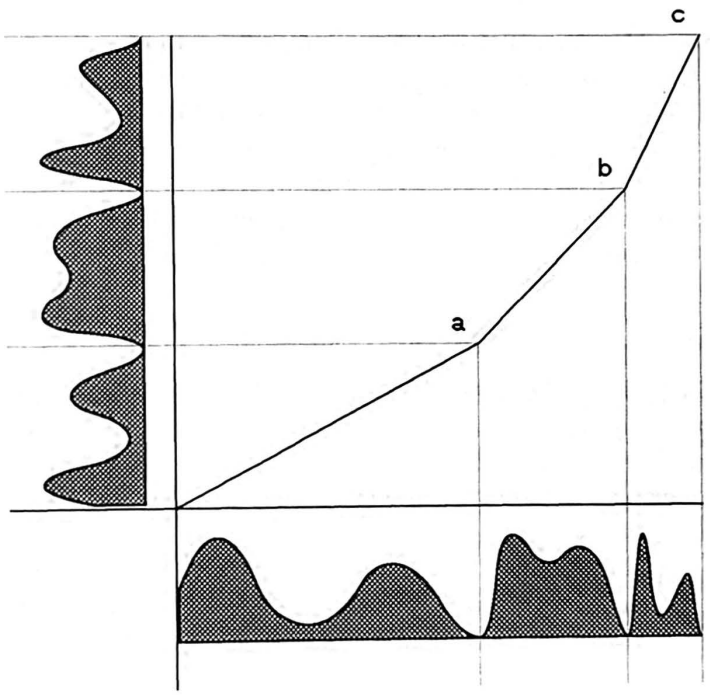


Figure 7.8. Dynamic Time Warping provides nonlinear time-scaling between a sample (horizontal-axis) and reference (vertical-axis) utterances. In region **a** the sample is spoken more slowly; in region **b** they are spoken at the same rate; and in region **c** the template is spoken more slowly.

IPR2023-00033
Apple EX1012 Page 56

the minimal error path between the input and template frames; relative speaking rate can be inferred from the slope of this path. A slope of 1 (45-degree line) is optimal when both are spoken at the identical rate.

It is necessary to place limits on the degree of time-scaling allowed by DTW. It would be senseless to compare 50 milliseconds of the sample to 1200 milliseconds of reference; normal speech does not exhibit such wide variability. Constraints on the maximum time-warping ratio are imposed by placing upper and lower bounds on the slope of the DTW path. The typical range of slopes is one-half to two, which accommodates a sample being spoken either half or twice as fast as the reference.

"Two-level" dynamic programming can be used to find word boundaries in a connected utterance containing an unknown number of words. At the lower level, DTW allows time-scaling of each word by computing the minimal error path matching the input to each template. At the higher level, these word-by-word error values are used to compute the best path through the set of all possible word combinations that could comprise a sentence. The two levels operate in concert to apportion the duration of the utterance among the set of selected templates. Figure 7.9 shows how a number of words of different lengths could consist of similar phones, or speech sounds.[4] It may be that the first three phones most closely match the short word "her," which would suggest that the utterance contained three words. But a better *overall* score might be obtained by matching these three phones against the beginning of "heartfelt" and selecting the utterance with only two words.

Dynamic programming is most beneficial for connected speech recognition. Word durations show greater temporal variation in connected speech than in isolated speech. Additionally, word boundaries must be found by considering how the continuous stream of phones can best be segmented into words in the absence of interword pauses. Dynamic programming is used in a number of commercial connected speech recognition systems, but its popularity for research purposes has faded in favor of Hidden Markov Models.

Hidden Markov Models

A **Hidden Markov Model (HMM)** is a two-stage probabilistic process that can be used as a powerful representation for speech. A Hidden Markov Model is a well-behaved mathematical construct, and a number of detailed algorithms exist for solving problems associated with HMMs; introductions to these can be found in [Rabiner and Juang 1986a]. This section first explains HMMs in the abstract and then demonstrates how they can be applied to speech recognition.

An HMM consists of a number of internal states; the model passes from an initial state to the final state as a step-by-step process generating an observable output at each step (state transition). For example, the states may correspond to phonemes contained in a word with the observable output corresponding to the

[4]Such a representation is called a **word lattice.**

IPR2023-00033
Apple EX1012 Page 57

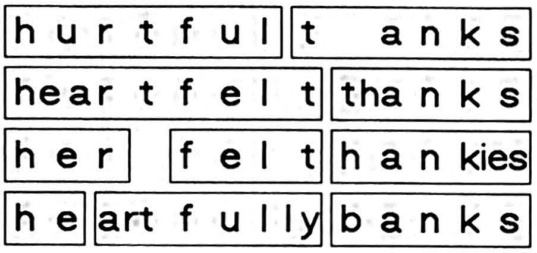


Figure 7.9. Different sequences of words could match an utterance; each sequence implies different word boundaries in the input utterance.

presence or absence of a number of acoustic features. At each step, the model can either move to a new state or stay in the current one. The model is "hidden" in that we cannot observe the state directly but only its output. From the series of observable outputs, we can attempt to guess when the model was in each state. Alternatively, we can say whether some series of outputs was likely to have been generated by a particular HMM.

For example, consider the arrangement shown in Figure 7.10, in which a box and a bowl are each full of black balls and white balls. The box has many more black balls than white while white balls dominate the bowl. Starting with the box, we remove one ball from it at random and pass it to an observer in another room who cannot see what is being done. Then a normal six-sided die is tossed. If the result is less than four, then the source of the next ball is the bowl. If the die is greater than three, then the next ball will be selected from the box. The cycle is then repeated.[5] Whenever we select from the box, a die throw of less than four shifts attention to the bowl. Once we start selecting from the bowl, we continue selecting from it unless the die shows a one, at which point we are finished. This model is likely to spend a majority of its cycles selecting from the bowl since a state transition from the box has a probability ½ of shifting to the bowl, but a bowl transition has a probability of only ⅙ of terminating and ⅚ of continuing with the bowl.[6] We should expect this arrangement to initially present more black balls than white to the observer (while it is in the box state) and then to produce more white balls. We should also expect more white balls overall. But keep in mind that this is a probabilistic process: Not all black balls come from the box and not all selections come from the bowl.

The box-and-bowl setup is a Hidden Markov Model; it is characterized by a number of *states* and associated probabilities for each *transition* from any state. The box and bowl are each a state; the rules about throwing the die define the

[5]This pedagogic example was inspired by Rabiner and Juang [Rabiner and Juang 1986b].

[6]Probability is expressed as a number between zero and one. If an event occurs with probability ½, we would expect it to happen once in two trials. A probability of ⅙ implies that we can expect the event once out of six trials. A smaller probability indicates that an event is less likely to occur.

IPR2023-00033
Apple EX1012 Page 58

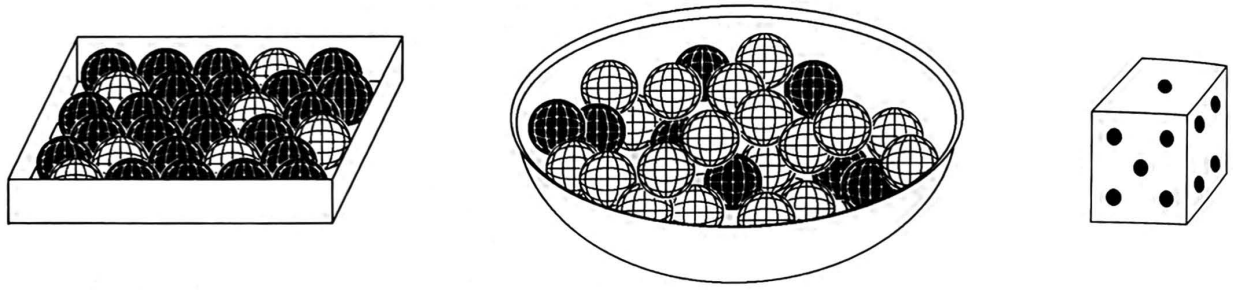


Figure 7.10. A box and a bowl full of colored balls illustrate a Hidden Markov Model.

transition properties. Figure 7.11 shows a more general HMM; in each state **S(n)** there is some probability **P(n,n)** of remaining in that state and some probability **P(n, n + 1)** of transitioning to the next state. In some models, it may be possible to bypass a state or a series of states as indicated by the dashed arc labeled P_{13} in the figure. The sum of the probabilities of all the arcs leaving a state is one; the model accounts for every possible path through its states.

How does this apply to speech recognition? An HMM can be used to represent a word with internal states representing characteristic acoustic segments, possibly phonemes or allophones. The output of a state is a frame or vector of acoustic parameters or features; this output is probabilistic to allow for variability in pronunciation and hence differences in the acoustic representation. The duration of an acoustic segment is a function of the number of steps in which the model is in the state corresponding to the segment. Staying in the same state, i.e. lengthening a phone, depends on the probability associated with the transition from that state to itself (P_{11} in Figure 7.11). Arcs such as P_{13} may be included to indicate that an intermediate state **S2** is optional during pronunciation, such as the second syllable in "chocolate."

In terms of the basic recognizer components, the "templates" consist of a set of HMMs with one HMM associated with every word. The acoustic representation of an input specimen of speech is *not* a collection of states but rather a set of acous-

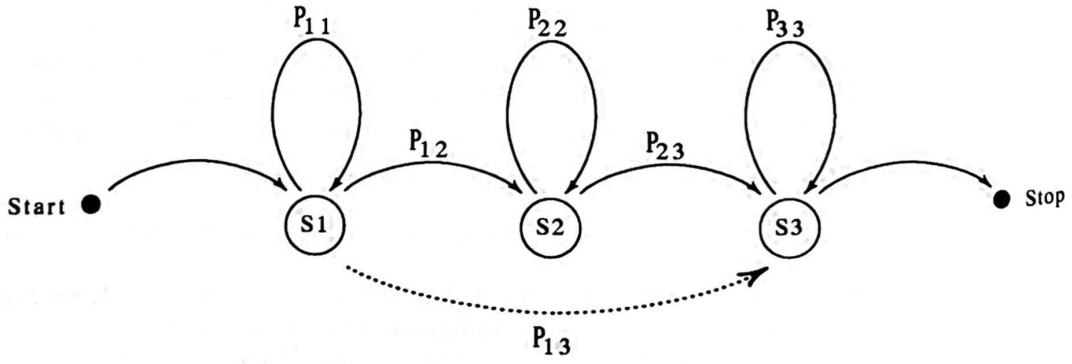


Figure 7.11. A Markov Model.

IPR2023-00033
Apple EX1012 Page 59

tic parameters corresponding to the possible observed sequences from the template HMMs. For the HMM-based recognizer, classification consists of determining which HMM has the highest probability of producing the observed acoustic sequence. The Viterbi algorithm is a commonly used method of solving this classification problem; a more detailed description is beyond the scope of this book but see [Lee and Alleva 1992] for an overview of its theory.

There are two related HMM problems which can also be solved mathematically. One is training the statistically defined HMMs that the recognizer requires as templates. Given some number of observation sequences (i.e., repetitions of the word), an HMM must be derived to represent that particular word. The second problem, not directly relevant during recognition, is to determine for some HMM and a set of observed outputs the internal states of the model that were most likely at each step.

Hidden Markov Models are a powerful representation of speech and consequently are used in many speech recognition systems currently under development. Although this section has described isolated word recognition, HMMs can also be used with connected speech to represent not only the phones of each word but also the probabilities of transitioning from one word to another. For connected speech recognition, the observed sequence would be generated by passing through a sequence of states in a sequence of HMMs with one HMM for each word spoken. The combination of words is also described statistically as another layer of HMMs in which each state corresponds to an entire word thereby encoding syntactic and semantic information.

Vector Quantization

Vector Quantization (VQ) is a technique employed for data reduction in speech coding and speech recognition. VQ is used to reduce a widely ranging input value into a smaller set of numbers representing typical values of the input. Because the input is classified as one of a small number of values instead of the value itself, less storage space is required. More importantly, the input is usually multidimensional (e.g., a set of LPC coefficients), and it is reduced to the single dimension identifying the nearest vector-quantized value.

In order to perform Vector Quantization, it is first necessary to decide how to efficiently cluster all possible input values. The analysis of sample data, which must be typical of the data to be encoded, determines how to cluster similar input values. The cluster's "center of gravity," or average value, then represents the entire set of data in the group. Each cluster is represented as one entry in a **codebook,** which lists the average value of each group. Having created a robust VQ codebook from specimen data, new data can be classified by selecting the codebook entry nearest in value to that data. The original value of the input data is lost; it is now represented by the single value associated with the codebook entry.

Building the codebook divides the space of possible values into regions and *typical* values. This is much like the process of selecting word templates for recognition as shown in Figure 7.6. The number of entries required for the codebook

IPR2023-00033
Apple EX1012 Page 60

depends on how well the data clusters into groups and the accuracy with which the data must be represented. If the data points are widely scattered, it is difficult to cluster them efficiently. The range of values subsumed by each cluster determines the accuracy with which vector quantized data can be represented since the codebook entry is used to look up a value for the data and only one value is stored per cluster.

Representing an input as a codebook entry having a similar value significantly reduces both storage and computation during pattern matching. To vector quantize an input value based on a codebook of N entries, the input must be compared with at most N possible values. The error between each of the codebook values can be computed in advance and stored in a two-dimensional array. During pattern matching the error between an input value and a corresponding template value can be found by table lookup instead of direct computation. The sum of the errors for each frame is the error between the input and that template.

As a simple example, vector quantization can be used to represent the number of hours per day that light bulbs are left burning in a home. If daily use is represented as an integral number of hours from zero to twenty-four, five bits of storage are required for each bulb. Higher resolution, such as minutes in addition to hours, requires even more bits of storage per bulb. However, by observing the individual bulbs we discover some interesting patterns of usage. Many bulbs are rarely turned on. Some bulbs are always left on all night like an outdoor lamp or a child's night light. Other bulbs, in the living room and kitchen, may typically be on from dusk until the occupants go to sleep. Still other bulbs such as in a closet may be turned on occasionally for brief periods of a few minutes each day.

This situation can be described by a vector quantization scheme employing a codebook with four entries to cover the four typical cases described in the previous paragraph.[7] Each entry in the codebook (see Figure 7.12) corresponds to one case of light bulb use, and the value stored for that entry indicates how long such a bulb is likely to be used. Using this codebook, light bulb use is encoded in two bits at a savings of three bits per bulb.

Vector quantization is usually employed to represent multidimensional quantities so the compression to a single-dimensioned codebook offers further data compression. To continue with the light bulb example, we might observe that many of the lamps used rarely or left on all night are low wattage, e.g., 60 or 75 watts. The living room and kitchen lights are more likely to be 100 or 150 watts to illuminate work areas. So we might vector quantize this information as well as shown in Figure 7.13. This new codebook indicates that a bulb assigned value two is typically 120 watts and on for five hours a day. Note that a 150 watt bulb used seven hours a day and a 75 watt bulb used four hours a day will both be represented by codebook value 2 as well. It may be that there is actually no single bulb of 120 watts; this is an average that represents the least error when compared to all the bulbs in the sample set with this codebook value.

[7]Such a codebook is said to be "of size four."

IPR2023-00033
Apple EX1012 Page 61

codebook entry	codebook value
0	0 hours
1	15 minutes
2	5 hours
3	10 hours

Figure 7.12. A simple VQ codebook representing light bulb use.

Once quantized, detailed information about a particular light bulb is lost, but we still have a general idea of how much each is used. This would be useful for identifying which bulbs should be replaced by more expensive energy-efficient models. While the memory savings are small in this example, vector quantization can be used for much greater data reduction in speech coding applications where it represents complex acoustical information about an utterance.

How is vector quantization used in speech representation? Consider the output of an LPC coder; it includes precise indications of the locations of each of the formants, or resonances, in the vocal tract as conveyed by filter parameters. But for gross categorization of the position of the vocal tract, e.g., to differentiate the vowels by their F1/F2 ratio, such precision is not necessary. Furthermore, not all combinations of F1 and F2 are possible since they are constrained by the physical characteristics of the vocal tract. Vector quantizing these filter coefficients can capture the overall vocal tract configuration in a very concise form when great precision of each parameter is not necessary.

Employing Constraints

A recognizer's task becomes increasingly difficult as the number of words to be recognized increases. To improve recognition large vocabulary systems apply constraints to the set of all possible words that can make up an input utterance.

codebook entry	codebook value	
	time	wattage
0	0 hours	65
1	15 minutes	75
2	5 hours	120
3	10 hours	70

Figure 7.13. Vector quantization is usually used to represent multidimensional quantities. Here it encodes both the average duration of illumination as well as the wattage of bulbs.

IPR2023-00033
Apple EX1012 Page 62

Constraints limit the search space during word classification by ruling out improper combinations of words. Although we readily employ general syntactic and semantic constraints to rule out many ill-formed or nonsense sentences while trying to understand a talker, it is difficult to express the knowledge that underlies such decisions in computationally tractable terms. Current recognition systems incorporate limited constraints specific to a given task.

Some of these constraints are specific to the the input; for example, North American telephone numbers have seven digits (ten if the area code is included). Other sets of constraints can be derived from a specification of all the legal utterances for a set of words and a task. For example, simple rules of syntax tell us that every sentence must have a verb, and common sense dictates that an application for ordering pizza should expect to encounter the word "large" adjacent to "pepperoni." To define these constraints the application designer must specify every legal sentence to be recognized. This is facilitated by lumping groups of words into classes which have meaning in a particular task context, such as "digit," "size," or "topping." Sentences can then be described as series of words combined from particular classes instead of listing every possible word sequence.

Because they are derived by observing people perform a task and computing the likelihood of particular sequences of words, constraints can also be probabilistic. This requires a large body of data from which to derive the probabilities as there may be many possible sentences and statistical representations that require a large quantity of training data. Some language models have been based on analysis of a large corpus of text documents. Although spoken and written language differ, the advantage of this approach is that the text is both readily available and easily analyzed by a computer.

The effectiveness of constraints is related to the degree to which they limit the possible input utterances; this can be stated as the predictive ability of a string of n words on the $n + 1st$ word. Alternatively, the constraints may be quantized as the **branching factor** associated with any node in a string of input words; this number indicates how many different words can follow one or more previous words. A question requiring a yes-or-no answer has a branching factor of two, for example, while one requiring a single digit has a branching factor of ten. **Perplexity** is a measure of the branching factor averaged across all possible word junctures in the set of legal utterances. The lower the perplexity, the more effectively lexical constraints can improve recognition.

Recognizers based on either *ad hoc* or statistical models of language and the task at hand may be subject to limited portability. The words a talker uses change as a function of the topic of discourse, and a recognizer may not be able to make such a shift. For example, a carefully constructed grammar for a task that allows the user to make inquiries into the state of ships in a navy's fleet will be rather different from one used in general business correspondence. Recognizers do not yet make use of language representations general enough to make specification of the task a simple problem. A more general discussion of methods of specifying both syntactic and semantic constraints is found in Chapter 9.

IPR2023-00033
Apple EX1012 Page 63

ADVANCED RECOGNITION SYSTEMS

This section briefly describes details of three advanced speech recognizers that are considerably more sophisticated than the simplistic example presented early in this chapter. These recognizers are research projects, not yet commercial products. This is hardly an exhaustive list of speech recognition research, but it is intended to be representative of the approaches being used in modern, large vocabulary recognition systems. These descriptions will almost certainly be out of date by the time this book is actually published so they should be taken as little more than a snapshot of a rapidly evolving field.

IBM's Tangora

IBM's Tangora speech recognizer is designed to recognize 5000 to 20,000 words of isolated speech [Jelinek 1985]. A real-time version of this research project was implemented on a single-slot personal computer board. The Tangora recognizer is speaker adaptive with training based on the user's reading a few paragraphs of text to calibrate its acoustic model.

The Tangora recognizer uses a vector quantization front end to classify acoustic input, which is then matched to words using discrete Hidden Markov Models of phoneme realization. A linguistic decoder then classifies the output of the acoustic front end producing a word for output. Tangora is targetted at automatic transcription of speech to text in a dictation context.

The linguistic constraints employed by the second stage of this recognizer are based on the probabilities of groups of two or three words occurring in sequence. This statistical model, called a **bigram** or **trigram grammar,** was derived through analysis of a large quantity of text from business correspondence. It should be readily apparent that such a representation of the language includes both syntactic and semantic information as well as indirectly encoded world knowledge. Not only is "late I slept" an unusual sequence due to constraints, but similarly "I slept furiously" is improbable due to its ill-formed semantics; "the green cat" is unlikely due to world knowledge that cats do not have green fur. But this model cannot distinguish the source of knowledge; it merely represents the end result of all factors that contribute to likely word sequences.

CMU's Sphinx

Carnegie-Mellon University's Sphinx recognizer is designed for speaker-independent connected speech recognition [Lee and Hon 1988, Lee 1988]. At the time of this writing, it operates in nearly real time using hardware multiprocessor peripherals to aid in the search process. Sphinx operates with a 1000 word vocabulary and achieves a recognition accuracy percentage in the mid-nineties when provided with a bigram language model; however, accuracy drops to the mid-fifties without the grammar model.

IPR2023-00033
Apple EX1012 Page 64

Sphinx makes extensive use of Hidden Markov Models: each phone is represented by an HMM, each word is a network of phones, and the language is a network of words. Phones in function words are modeled separately from other phones because of the higher degree of coarticulation and distortion in function words. Because function words are common to all vocabularies, they can be modeled using more thoroughly trained HMMs.

Sphinx uses multiple codebook vector quantization as its acoustic front end. Separate codebooks are used to encode energy, cepstral, and differential cepstral parameters. The cepstrum is a signal analysis construct gaining popularity with speech recognition because it differentiates voicing and vocal tract aspects of the speech signal. Differential cepstrum measurements indicate how the cepstrum changes from one sampling period to the next; this measure is particularly useful for analysis of the most dynamic aspects of speech such as many of the consonants.

MIT's SUMMIT

The SUMMIT [Zue *et al.* 1989a] system is a phoneme-based connected speech recognizer being developed by the Spoken Language Systems Group at M.I.T. In recognizing speech, SUMMIT first transforms the speech signal into a representation modeled after the auditory processing that occurs in our ears. This is a nonlinear transformation based on Seneff's auditory model [Seneff 1988]; it enhances acoustic information crucial for recognizing speech and suppresses irrelevant detail. Part of this analysis models the transduction between the hairs in the cochlea and their associated nerve cells to enhance temporal variation in the input emphasizing onsets and offsets crucial to detecting consonants. Another portion, which models the nerve cell firing rate related to the characteristic frequency of the cell, emphasizes spectral peaks; this is useful for vowel identification.

The next stage of recognition segments the transformed speech by finding acoustic landmarks in the signal. Regions between the landmarks are grouped into segments on the basis of similarity. A set of phonetic classifiers assigns probabilistic phonetic labels to each of these portions of the speech signal. Different classifiers may be invoked for differentiating among distinct classes of phonemes. Words are represented by a network of phonemes; several networks may be used to encode alternate pronunciations. A pattern matching algorithm similar to dynamic programming matches the phoneme labels of the input against the word networks, deducing word boundaries in the process. Coarticulation rules compensate for some phonemic aspects of connected speech.

SUMMARY

This chapter has detailed the process of recognizing speech by computers. It began with an overview of the recognition process as a set of three basic components: an acoustical representation, a vocabulary or set of templates, and a pattern matching process to measure the similarity of an input utterance with each

IPR2023-00033
Apple EX1012 Page 65

of the set of templates. This process was illustrated by considering a simplistic, hypothetical recognizer. The chapter then discussed a full range of possible recognizer configurations differentiated by the specificity of the talker, the style of speaking, and the size of the vocabulary to be recognized. Several techniques important for connected and large vocabulary recognition were then introduced; these included Dynamic Time Warping, Hidden Markov Models, Vector Quantization, and language constraints. The chapter concluded with a descriptive overview of three well-known research projects on large vocabulary speaker independent recognizers.

This chapter was intended to introduce the technology of speech recognition and some of the difficulties it must overcome to be effective. Speech recognition algorithms are under continual development and are useful for an ever-expanding range of applications. Even more important for its deployment, the steady increases in microprocessor speeds enable even the more sophisticated algorithms to be implemented entirely as software thus encouraging the more widespread distribution of applications that make use of recognition.

The next chapter explores the utility of speech recognition, the classes of applications for which it may be well suited, and interaction techniques to make effective use of recognition. Of central concern is the issue of speech recognition errors; error rates remain the primary deterrent to the successful application of speech recognition. But in some specialized application niches speech recognition is already being used to improve personal productivity.

FURTHER READING

O'Shaughnessy covers many topics in speech recognition thoroughly. Furui and Sondhi is a collection of particularly current papers about all aspects of speech processing with particular emphasis on recognition. Although rather technical, it has chapters describing most of the concepts and some of the speech-recognition research projects discussed in this chapter. Waibel and Lee present a collection of previously published research papers about speech recognition. Dixon and Martin is a similar collection of key papers from the 1970s. Another good source of very good working papers are conference proceedings of the DARPA Speech and Natural Language Workshops.

IPR2023-00033
Apple EX1012 Page 66

Voice Communication with Computers

Conversational Systems

Christopher Schmandt

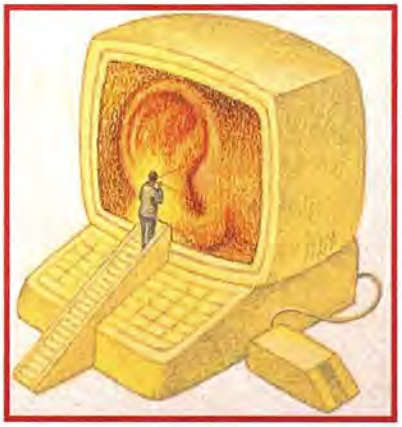
This book discusses how human language techniques can be embedded in human-computer dialogue to make "talking computers" conversational. It uses an interdisciplinary approach to explain application of speech technologies to computer interfaces.

Chapters progress from physiological and psychological components of speech, language, and hearing to technical principles of speech synthesis and voice recognition to user interaction techniques to employ such technologies effectively. Areas covered include:

- Computational means of expressing linguistic and discourse knowledge
- Software architectures that support voice in multimedia computer environments
- The operations of digital recording, speech synthesis, and speech recognition
- Coding schemes based on data rate, intelligibility, and flexibility
- Applications and editing of stored voice in multimedia computer documents
- Integration of telephone functionality into computer workstations

Case studies demonstrate how to relate utterances to intention and real-world objects, use discourse knowledge to carry on the thread of a conversation across multiple exchanges, and integrate speech into many aspects of computer related work. Other topics addressed are:

• Desktop audio • Interactive voice response • ISDN—digital telephony • Text-to-speech algorithms • Interaction between audio and window systems • Operating system support and run-time support for voice • Applications that provide a means of accessing personal databases while away from the office



Interactive speech systems represent some of the newest technology available from several computer manufacturers. *Voice Communication with Computers* will provide vital information to application developers, user interface designers or researchers, human factors developers, groupware developers, and all professionals interested in taking advantage of the ability to

About the Auth

Christopher Schmandt is Director of the Speech Research Group, and a Principal Research Scientist at the Massachusetts Institute of Technology Media Laboratory.

Cover illustrations by Thomas Sciacca
Cover design by Angelo Papadopoulos

VAN NOSTRAND REINHOLD
115 Fifth Avenue, New York, NY 10003

ISBN 0-442-23935-1

9 780442 239350

IPR2023-00033
Apple EX1012 Page 67