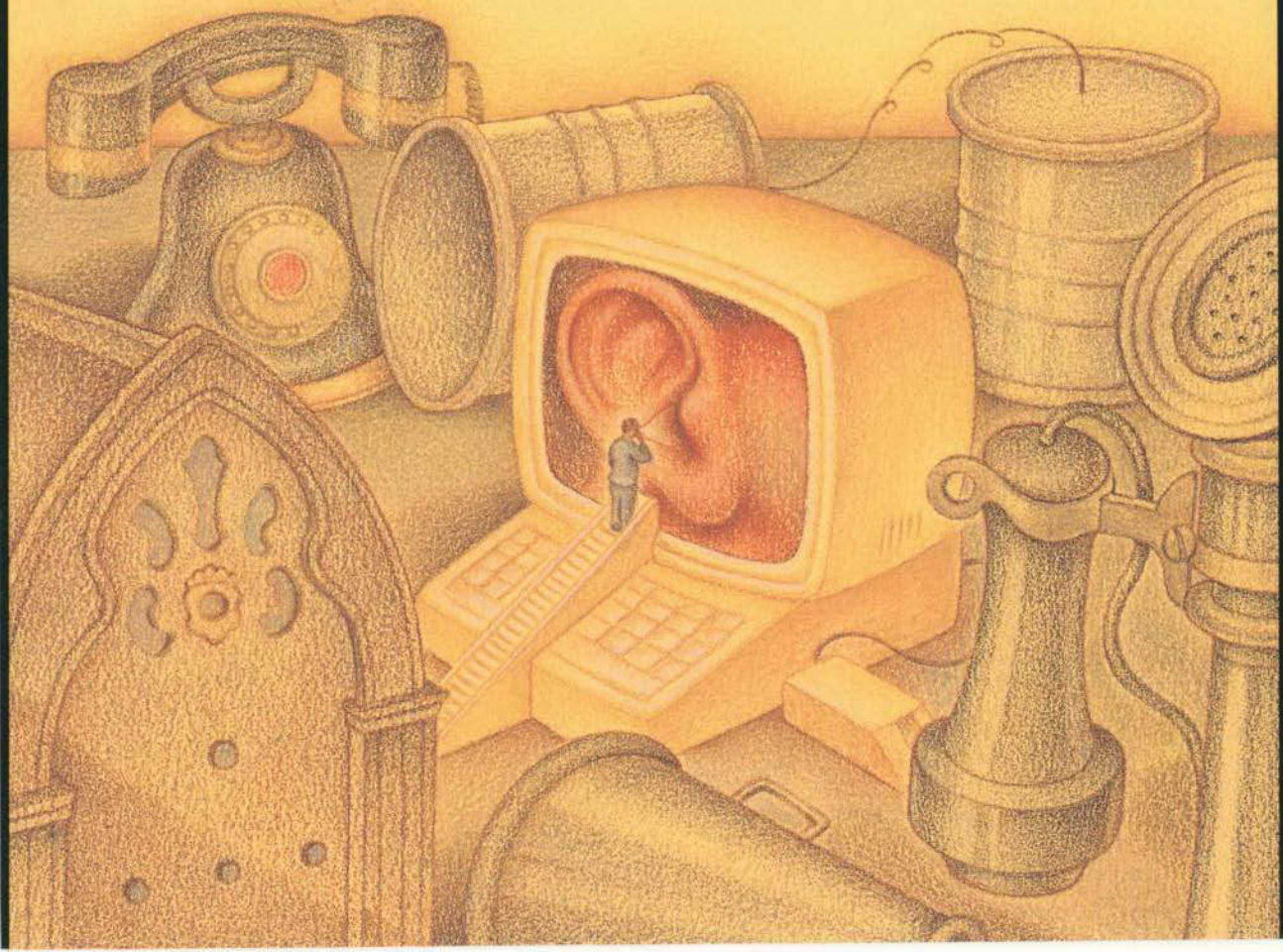
MAS.632 Speech Interfaces and Mobile Devices
Fall 2008

Christopher Schmandt
Foreword by Nicholas Negroponte
MIT Media Lab

Voice Communication with Computers

Conversational Systems



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 speak with a computer.

About the Author

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

VOICE COMMUNICATION WITH COMPUTERS

Conversational Systems

CHRISTOPHER SCHMANDT

VAN NOSTRAND REINHOLD
New York

To Ava and Kaya for the patience to see me through writing this and helping me learn to hear the voice of the desert.

Contents

Speaking of Talk

As we enter the next millennium, speech unquestionably will become the primary means of communication between human beings and machines. By primary I mean the one most often used. The reasons are simple and go beyond our intuitive desire to use native spoken language, which comes naturally to us and accounts for so much of human-to-human discourse.

Namely, the value of speech is not just the use of "natural language," as such, but includes other valuable properties. Speech works in the dark. It allows you to communicate with small objects. It lets you transmit a message beyond arm's reach and around corners. It is a subcarrier of meaning through prosody, tone and, of course, volume.[1]

How does speech play into today's trends in computing?

Miniaturization is surely one such trend. I expect to have more computing on my wrist in ten years than I have on my desk today. Another is ubiquity. I expect computing to be woven into the entire fabric of my life.[2] Another is concurrence. More and more, I see the parallel use of our auditory channels, when our hands and eyes are occupied.[3] Finally, I anticipate the increasing use of computers "in passing." By that I mean the ability to voice a remark, pose a question, or give a

[1]Any parent knows that it is not what you say, but how you say it.

[2]In the literal sense, as well, I expect apparel to be part of my private area network: PAN.

[3]Schmandt illustrates this in Chapter 6 with his example of the Back Seat Driver.

salutation without bringing all of one's other faculties to a grinding halt, as we do now when we sit down to a session with our computer. Computer usage today is solely in the fovea of our attention, while tomorrow we will enjoy it equally at the periphery of thoughts, actions, and consciousness, more like air than headlines.

Consider these trends and the features intrinsic to the audio domain, and there is little question of where we are headed. Talk will be the most common means of interaction with all machines, large and small.[4]

This book is not only timely but different. Chris Schmandt brings a unique and important perspective to speech. Let me illustrate, with a story, the difference I see between his work and the work of others. The story is old and has gained some color with time, but illustrates an important point about breaking engineering deadlocks.

In 1978, the so-called Architecture Machine Group received the first working model of NEC's connected speech recognizer—the top of the line for its day, designed by Dr. Yasuo Kato, now a member of the Board of Directors of that same corporation. It was a continuous speech recognition system, speaker-dependent, and able to handle fifty utterances.[5]

In the celebrated program called Put-That-There (discussed in Chapter 8) this machine worked with unmatched elegance and became the objective and site for technical tourists from all over the world. But it had a problem; one that turned out not to be unique to that specific device but, in fact, quite general.

If a user of the system was tense or nervous, his or her voice would be affected. Although this condition was not necessarily audible to another human being, it was distinctly noticeable to the machine. This seemingly harmless feature turned into a disastrous bug, since the enormous importance of presenting one's results to sponsors often creates tension.

The Advanced Research Projects Agency (ARPA) makes periodic visits to MIT, frequently with very high-ranking military personnel. Of course, ARPA was very interested in this elegant prototype of the future of command and control systems.[6] The importance of these visits was so great that graduate students, sometimes Chris Schmandt himself, became sufficiently nervous that a perfectly working system (ten minutes before) functioned improperly in the presence of these august, easily disappointed, and frequently disbelieving visitors. There is something particularly frustrating and stupid looking about a speech recognition system when it does not work, which only exacerbates the problem and makes the person demonstrating the system even more anxious.

[4]I do not limit my remarks to computers, as we know them, but consider all machines: home appliances, entertainment equipment, telecommunication apparati, transportation machinery, engines for work, play, and learning—as well as some not yet imagined.

[5]The difference between a word and an utterance is crucial, as an utterance can be a part of a word or a concatenation of them (up to five seconds in this particular case).

[6]Command and control is deeply rooted in voice; my own limited experience in Army exercises witnessed a great deal of yelling.

The short-term fix was to train students and staff who demonstrated the system to be cool, calm, and collected and to ignore the stature of their guests. Guests were encouraged to wear blue jeans versus uniforms.

The longer-term solution resides in a different view of speech, as communication and interaction, not just as a string of words in the acoustic domain. This perspective results in a variety of ideas, of which one is to find the pauses in the speaker's speech[7] and judiciously issue the utterance "ah ha," taken from a suite of "ah ha's" ranging from the mousy to the elephantine. Namely, as one speaks to a system it should listen and, now and again, say "ah ha"—"ah ha"—"ah ha." The user will relax and the system's performance will skyrocket.

Discussions of ideas such as this have been highly criticized as yet another example of how members of the Media Lab are mere charlatans. How dare they give a user the false impression that the machine has understood when, in fact, it may not have?

The criticism is a perfect example of why computer science is so sensory deprived and getting duller by the day. Critics miss the point totally. Listen to yourself on the telephone. As someone is talking to you, you will naturally issue an "ah ha" every fifteen to thirty seconds. In fact, if you don't, after a while your telephone partner will be forced to ask, "Are you still there?" The point is that the "ah ha" *has no lexical value*. It is purely a communications protocol, no different from the electronic hand-shaking with which network architects are so familiar.

How did this escape researchers in the speech recognition community, itself a mature pursuit for over twenty-five years? The same question can be rephrased: Why was speech production and speech recognition so utterly decoupled?[8]

My answer is not popular with most members of the speech community, but it is at the root of why this book is so special. That is, the traditional speech recognition community was fundamentally disinterested in "communication." Its interest was in "transcription"; namely, transcoding from ASCII into utterances or spoken language into ASCII. It was not interested in the tightly coupled interaction of human dialogue, typical of face-to-face communication. That kind of speech, Schmandt's kind of speech, is filled with interruption, retort, and paralinguals (the "ah ha's"), wherein the verbal back-channel is no less important than speech output. This concept is what *Voice Communication with Computers* brings to the table.

Voice Communication with Computers is not just a textbook or tutorial, it is a perspective. Its point of view underscores the need to interact and communicate with computers in much the same way we communicate with human beings. Only when we are able to talk to machines the way we talk to people will the concept of "ease of use" be seriously addressed.

[7]We all pause, even those of us who speak like machine guns must come up for air.

[8]Literally, in almost all labs world-wide, researchers in speech production and speech recognition were in different groups, different labs, or different divisions.

Today "ease of use" means simple, direct manipulation, with little or no delegation. The future of computing is discourse and delegation and, for that, speech is mandatory—not a process of transcription but conversation. This is the area pioneered by Schmandt and reflected in the following pages.

Nicholas Negroponte
MIT Media Lab

Preface

This book began as graduate class notes to supplement technical readings from a number of disciplines. Over the years as I have worked on various projects, I kept running against the question, "What can I read to better understand this?" Although materials are available, texts have been too focused and detailed and papers are spread across too many conference proceedings for either students or specialists in other areas to make good use of them. This book is meant to open the more specialized disciplines to a wider audience while conveying a vision of speech communication integrated with computing.

The primary message of this book is an approach to incorporating speech into computing environments; it is unique in its focus on the *user* of speech systems and interaction techniques to enhance the utility of voice computing. But a complete evaluation of speech at the interface is grounded in the understanding of speech technologies, the computational means of expressing linguistic and discourse knowledge, and the software architectures that support voice in multimedia computer environments.

This interdisciplinary approach is the most valuable asset of this book but simultaneously a limitation. I attempt to open technical doors for readers who span many fields and professions. For example, I discuss speech coding without reference to the standard notations of discrete time digital signal processing. This is a little bit like trying to teach high school physics to students who have not yet learned calculus: terribly inelegant in retrospect *if* one pursues studies in physics. But this is not a speech coding text; this book examines how and when to *use* speech coding, speech recognition, text-to-speech synthesis, computer-telephone interfaces, and natural language processing. Anyone trying to under-

stand how to use voice in emerging computing environments should benefit from this information.

Technology has changed rapidly in the five years this book has been in progress. While it has been reassuring to see many of my expectations fulfilled—and technological developments only increase the potential for pervasive deployment of speech-capable applications—these changes have necessitated continual updating in this text. This pace of change almost guarantees that some portion of this book will be out of date by the time it appears in print. Five years ago speech technology was awkward, expensive, and difficult to integrate into work environments. Speech systems research almost invariably required interfacing to additional peripheral devices including complete components such as text-to-speech synthesizers or internal add-in hardware such as digital signal processors. No computer was equipped with a microphone back then, and while computer speakers usually were connected to tone-generating hardware to emit various beep sounds, these systems could not play a digital audio file. At that time, large-scale consumer access to digitized audio via the compact disc was only just beginning to make serious inroads on analog audio tape and vinyl LP music distribution.

Much has changed while I wrote and rewrote this book. Most computers are now audio capable, equipped with microphones and the ability to play back digital audio at qualities ranging from that of the telephone to that of the CD. The capacity of disk drives is now measured in Gigabytes, not Megabytes, minimizing the difficulty of storing audio data, which requires much more disk space than text. Some computers have been shipped with digital signal microprocessors included on the mother board. Others ship with digital-telephone network interfaces (ISDN) as standard equipment. The explosion of cellular telephony and palmtop computers has created a new community of potential computer audio users, who need the access and the means to communicate with computers as well as with other people while on the move.

Perhaps the most significant development, however, is the more than ten-fold increase in the speed of general-purpose processors found in all classes of computers. This speed now enables the majority of computers to perform basic real-time voice processing and simple speech recognition. The more powerful desktop workstations are capable of the highest quality text-to-speech synthesis, large vocabulary speech recognition, and sophisticated voice processing. Speech technology is rapidly becoming a commodity that can be distributed as shrink-wrapped software or included with a computer operating system.

During these five years, applications have also followed the enabling technology although with a frustrating lag. Basic speech recognition to control computer window systems is available in several products. A number of isolated applications allow voice annotations to spreadsheets and other text-oriented data. Several competing formats for voice attachments to electronic mail have also appeared, and now progress is being made to converge on a single standard format. Recently a new venture initiated a digital audio "radio" program distribution over the Internet. Speech recognition is now being used in the telephone network, to help complete collect call attempts, for example.

This is just the beginning of ubiquitous voice technology. Voice is so useful and such an important part of our lives that, by necessity, it will become pervasive in the ways computers are used. The trends of the last five years will continue only to accelerate over the next five. While we must accept that we are still a long, long way from the science fiction theme of the ever-present listening computer (with a pleasant voice as well!), we will find ourselves holding an increasing number of conversations with machines. Some of the readers of this book will be among those researchers or developers making this a reality.

Acknowledgments

This book originated with sketchy class notes coauthored with Jim Davis, and Pascal Chesnais triggered our note-writing by encouraging us to give students almost anything. Jim wisely decided to complete his dissertation instead of joining me to write this book and graduated years ago.

I have had help from many people both in editing my technical writing as well as reviewing portions of this manuscript for accuracy. I have benefitted tremendously from their help; the errors that remain are my own. This list includes Barry Arons, Janet Cahn, Jim Davis, Kate Gordon, Harry Hersh, Lisa Stifelman, Ben Stoltz, Tony Vitale, and Nicole Yankelovich.

I have also benefitted tremendously from both editorial help as well as immeasurable assistance with the production of this book from Gayle Sherman and Elaine McCarthy. I think I would have given up months from completion without Gayle's help. Without the patience of Dianne Littwin at Van Nostrand Reinhold, I'm sure I would have backed out of this project years ago. And most important has been the support of my family, Ava and Kaya, who let me add the time to write this book to an already overly crowded work schedule.

This book includes descriptions of many research projects we have implemented at M.I.T, both at the Media Laboratory as well as one of its predecessors, the Architecture Machine Group. I would never have been in the position to create the group that performed this work without the unwavering support of my laboratory director, Nicholas Negroponte, for the last thirteen years. And I probably would have been unable to deal with the politics of a new laboratory without the comradeship of Walter Bender, even though we have worked on completely independent projects since jointly debugging network code a decade ago.

Finally and most importantly, my own ideas as well as these projects have evolved over years of interacting with students. They have written much of the software, helped me debug my class notes, and suffered through my attempts to put these thoughts into a coherent whole in my courses. Some have been friends, some have been peers; yet we have had plenty of disagreements. But our graduates are more valuable than anything else we do, and I have enjoyed working with them immensely. It would be impossible to rank-order their contributions, so I list them alphabetically: Mark Ackerman, Barry Arons, Derek Atkins, Lorne Berman, Jim Davis, Lorin Jurow, Debby Hindus, Angie Hinrichs, Chris Horner, Eric Hulteen, Andrew Kass, Kevin Landel, Eric Ly, Sanjay Manandhar, Mike McKenna, Atty Mullins, Sheldon Pacotti, Charles Simmons, Jordan Slott, Lisa Stifelman, Tom Trobaugh, Mark Vershel, Todd Wayne, Chi Wong, and Jim Zamiska.

During the course of writing this book I discovered the deserts of the American Southwest. Over the years I find myself repeatedly drawn back to their open spaces and infinite vistas. Although I have edited a significant portion of this book in their midst I have missed them badly while sitting at desks trying to put this project to rest. Although the two are somehow inextricably intermingled in my consciousness, I look forward to getting back to the purity of the desert without carrying this project along.

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.

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-

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

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.