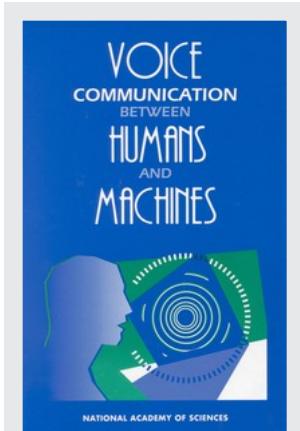


This PDF is available at <http://nap.edu/2308>

SHARE



## Voice Communication Between Humans and Machines (1994)

### DETAILS

560 pages | 6 x 9 | HARDBACK

ISBN 978-0-309-04988-7 | DOI 10.17226/2308

### CONTRIBUTORS

David B. Roe and Jay G. Wilpon, Editors; for the National Academy of Sciences

GET THIS BOOK

FIND RELATED TITLES

### SUGGESTED CITATION

National Research Council 1994. *Voice Communication Between Humans and Machines*. Washington, DC: The National Academies Press.  
<https://doi.org/10.17226/2308>.

Visit the National Academies Press at [NAP.edu](#) and login or register to get:

- Access to free PDF downloads of thousands of scientific reports
- 10% off the price of print titles
- Email or social media notifications of new titles related to your interests
- Special offers and discounts



Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press.  
[\(Request Permission\)](#) Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.

Copyright © National Academy of Sciences. All rights reserved.

# **VOICE COMMUNICATION BETWEEN HUMANS AND MACHINES**

David B. Roe and Jay G. Wilpon, *Editors*

National Academy of Sciences

National Academy Press  
Washington D.C. 1994

Copyright National Academy of Sciences. All rights reserved.

**NATIONAL ACADEMY PRESS 2101 Constitution Avenue, N.W. Washington, D.C. 20418**

This volume is based on the National Academy of Sciences' Colloquium on Human-Machine Communication by Voice. The articles appearing in these pages were contributed by speakers at the colloquium and have not been independently reviewed. Any opinions, findings, conclusions, or recommendations expressed in this volume are those of the authors and do not necessarily reflect the views of the National Academy of Sciences.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

**Library of Congress Cataloging-in-Publication Data**

Voice communication between humans and machines / David B. Roe and Jay G. Wilpon, editors.

p. cm.

Based on a colloquium sponsored by the National Academy of Sciences.

Includes bibliographical references and index.

ISBN 0-309-04988-1

1. Automatic speech recognition. 2. Man-machine systems. I. National Academy of Sciences (U.S.)

TK7882.S65V62 1994

006.4'54—dc20 94-29114

CIP

Copyright 1994 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

## Acknowledgments

The editors would like to thank the many authors who contributed to this book. Without their insights and hard work this book would not have been possible. We also thank Lawrence Rabiner, who inspired and helped organize the NAS-sponsored Colloquium on Human/Machine Communication by Voice. The efforts of Irene Morrongiello, who was instrumental in coordinating every aspect of the manuscript, and Martina Sharp, who made formatting changes to keep the style similar between the diverse contributions of the authors, are gratefully appreciated.

We wish to thank many anonymous reviewers who made generous comments on the chapters.

Finally, this book would not have been possible without the support of our wives, Carol Roe and Sandy Wilpon.



# Contents

<b>Dedication</b>	1
<b>Voice Communication Between Humans and Machines—An Introduction</b>	5
<i>Lawrence R. Rabiner</i>	
<b>SCIENTIFIC BASES OF HUMAN-MACHINE COMMUNICATION BY VOICE</b>	
Scientific Bases of Human-Machine Communication by Voice	15
<i>Ronald W. Schafer</i>	
The Role of Voice in Human-Machine Communication	34
<i>Philip R. Cohen and Sharon L. Oviatt</i>	
Speech Communication—An Overview	76
<i>James L. Flanagan</i>	

<b>SPEECH SYNTHESIS TECHNOLOGY</b>	
Computer Speech Synthesis: Its Status and Prospects <i>Mark Liberman</i>	107
Models of Speech Synthesis <i>Rolf Carlson</i>	116
Linguistic Aspects of Speech Synthesis <i>Jonathan Allen</i>	135
<b>SPEECH RECOGNITION TECHNOLOGY</b>	
Speech Recognition Technology: A Critique <i>Stephen E. Levinson</i>	159
State of the Art in Continuous Speech Recognition <i>John Makhoul and Richard Schwartz</i>	165
Training and Search Methods for Speech Recognition <i>Frederick Jelinek</i>	199
<b>NATURAL LANGUAGE UNDERSTANDING TECHNOLOGY</b>	
The Roles of Language Processing in a Spoken Language Interface <i>Lynette Hirschman</i>	217
Models of Natural Language Understanding <i>Madeleine Bates</i>	238
Integration of Speech with Natural Language Understanding <i>Robert C. Moore</i>	254
<b>APPLICATIONS OF VOICE-PROCESSING TECHNOLOGY I</b>	
A Perspective on Early Commercial Applications of Voice-Processing Technology for Telecommunications and Aids for the Handicapped <i>Chris Seelbach</i>	275

Applications of Voice-Processing Technology in Telecommunications <i>Jay G. Wilpon</i>	280
Speech Processing for Physical and Sensory Disabilities <i>Harry Levitt</i>	311
<b>APPLICATIONS OF VOICE-PROCESSING TECHNOLOGY II</b>	
Commercial Applications of Speech Interface Technology: An Industry at the Threshold <i>John A. Oberteuffer</i>	347
Military and Government Applications of Human-Machine Communication by Voice <i>Clifford J. Weinstein</i>	357
<b>TECHNOLOGY DEPLOYMENT</b>	
Deployment of Human-Machine Dialogue Systems <i>David B. Roe</i>	373
What Does Voice-Processing Technology Support Today? <i>Ryohei Nakatsu and Yoshitake Suzuki</i>	390
User Interfaces for Voice Applications <i>Candace Kamm</i>	422
<b>TECHNOLOGY IN 2001</b>	
Speech Technology in the Year 2001 <i>Stephen E. Levinson and Frank Fallside</i>	445
Toward the Ultimate Synthesis/Recognition System <i>Sadaoki Furui</i>	450
Speech Technology in 2001: New Research Directions <i>Bishnu S. Atal</i>	467

New Trends in Natural Language Processing: Statistical Natural Language Processing <i>Mitchell Marcus</i>	<a href="#">482</a>
The Future of Voice-Processing Technology in the World of Com- puters and Communications <i>Yasuo Kato</i>	<a href="#">505</a>
<b>Author Biographies</b>	<a href="#">515</a>
<b>Index</b>	<a href="#">525</a>

# **VOICE COMMUNICATION BETWEEN HUMANS AND MACHINES**



## Dedication

FRANK FALLSIDE



Frank Fallside was one of the leading authorities in the field of speech technology. His sudden and wholly unexpected death at the age of 61 has robbed all those who worked with him, in whatever capacity, of a highly respected colleague. His loss will be felt throughout the world.

Frank was educated at George Heriot's School in Edinburgh and took his first degree, in electrical engineering, at the University of Edinburgh in 1953. He obtained his Ph.D. from the University of Wales in 1958. He then went to Cambridge University as a member of the Engineering Department, where he spent the rest of his career.

Trained as an electrical engineer, Frank did his early research in cybernetics—on servomechanisms and control systems. By the early 1970s, he was applying the results of his research to a field in which he retained, thereafter, an enduring interest, the analysis and synthesis of human speech by computer.

The field of speech technology has expanded enormously in the past 20 years or so. Frank Fallside played a major role in its expansion. In his own research he maintained his interest in the computational analysis and synthesis of speech, but he also acquired a research interest in the related areas of robotics, vision, and geometrical reasoning. As the cybernetics of the 1960s developed into the more broadly based information technology of the 1970s, and as information technology merged with cognitive science and neuroscience in the 1980s, Frank kept up with new ideas and techniques, making himself familiar with the relevant work in artificial intelligence, computer science, linguistics, neurophysiology, formal logic, and psychology. His academic distinction was recognized by Cambridge University in 1983 when he was appointed as professor of information engineering.

In his last few years Frank developed a specialized interest in the theory of artificial neural networks and built up a large team of researchers to investigate its many applications. At the time of his death he was working with colleagues in the Department of Zoology and the Computer Laboratory to plan an ambitious research program directed at establishing a "bridgehead" between engineering and neurobiology. The theory of neural networks was to play a central role in this research, but the theoretical model was to be confirmed by neurobiological measurements and experiment. He was working on a theory of language acquisition, which, by exploiting the typically cybernetic notion of continuous and corrective on-line feedback, would draw on and integrate recent work in both the analysis and the synthesis of speech.

The interdisciplinary postgraduate "conversion" course that Frank established in 1985, the "MPhil in Computer Speech and Language Processing," is unusual, if not unique, in being based in a department of engineering. It accepts students with first degrees in either arts or science and is taught by specialists from many different departments across the university. As an engineer, however, Frank made sure that, however broadly based and, in parts, theoretical, the teaching was, the students' projects were practically oriented and directly linked to perceived industrial and commercial needs.

As head of the Information Engineering Division in the Engineering Department at Cambridge, Frank carried a very heavy administrative load. He combined this with teaching and the personal supervision of no fewer than 20 research students and, in addition to this and other university work, with the editorship of Computer Speech and Language, with the organization of international conferences, and

with service on several important national and international committees.

As sorely missed as his intellectual presence, however, will be his companionship on social occasions: the range of his interests, cultural and intellectual; his wit and good humor; his ability to put his own strongly held views quietly and persuasively and to listen to opposing views seriously and without condescension; his moral and social commitment; his sympathetic concern for others; his complete lack of envy or malice; and, above all, his sense of fun.

For all of this we dedicate this volume to the memory of Frank Fallside.

Stephen E. Levinson

John Lyons



# Voice Communication Between Humans and Machines—An Introduction

*Lawrence R. Rabiner*

Some great pundit once remarked, "Every time has its technology, and every technology has its time." Although this is a somewhat simplistic view of technology, it is indeed true that when one thinks of a time period (especially in modern times) one always associates with it the key technologies that "revolutionized" the way people lived then. For example, key technologies that "came of age" in the 1970s include the VLSI integrated circuit, the photocopier, the computer terminal, MOS memory, and scanners. In the 1980s we saw the advent and growth of the personal computer, fiber optics, FAX machines, and medical imaging systems of all types. It is not too difficult to predict what some of the key technologies of the 1990s will be; these include voice processing, image processing, wireless communications, and personal information terminals.

If we examine the various technologies noted above and look at the interval between the time the technology was "understood" and the time the technology began to mature and grow, we see a very complex and intricate relationship. For example, the basic principles of FAX were well understood for more than 150 years. However, until there were established worldwide standards for transmission and reception of FAX documents, the technology remained an intellectual curiosity that was shown and discussed primarily in the research laboratory. Similarly, the concept (and realization) of a videophone was demonstrated at the New York World's Fair in 1964 (so-called

Picturephone Service), but the first commercially viable instruments were actually produced and sold in 1992. In this case it took a bandwidth reduction (from 1.5 Mbps down to 19.2 Kbps) and a major cost reduction, as well as algorithm breakthroughs in voice and video coding and in modem design, to achieve this minor miracle.

Other technologies were able to leave the research laboratory rather rapidly, sometimes in response to national imperatives (e.g., miniaturization for the space program), and sometimes in response to business necessities. Hence, when fiber optic lines were first mass produced in the 1980s, it was estimated that it would take about two decades to convert the analog transmission facilities of the old Bell System to digital form. In reality the long-distance telephone network was fully digital by the end of 1989, fully 10 plus years before predicted. Similarly, in the case of cellular telephony, it was predicted that it would be about a decade before there would be 1 million cellular phones in use in the United States. By the end of 1992 (i.e., about 8 years after the beginning of the "cellular revolution"), the 10-millionth cellular phone was already operating in the United States, and the rate of growth of both cellular and wireless telephony was continuing unabated.

Now we come to the decade of the 1990s and we have already seen strong evidence that the key technologies that are evolving are those that support multimedia computing, multimedia communication, ease of use, portability, and flexibility. The vision of the 1990s is ubiquitous, low-cost, easy-to-use communication and computation for everyone. One of the key technologies that must evolve and grow to support this vision is that of voice processing. Although research in voice processing has been carried out for several decades, it has been the confluence of cheap computation (as embodied by modern digital signal processor chips), low-cost memory, and algorithm improvements that has stimulated a wide range of uses for voice processing technology across the spectrum of telecommunications and consumer, military, and specialized applications.

## ELEMENTS OF VOICE PROCESSING TECHNOLOGY

The field of voice processing encompasses five broad technology areas, including:

- voice coding, the process of compressing the information in a voice signal so as to either transmit it or store it economically over a channel whose bandwidth is significantly smaller than that of the uncompressed signal;

- voice synthesis, the process of creating a synthetic replica of a voice signal so as to transmit a message from a machine to a person, with the purpose of conveying the information in the message;
- speech recognition, the process of extracting the message information in a voice signal so as to control the actions of a machine in response to spoken commands;
- speaker recognition, the process of either identifying or verifying a speaker by extracting individual voice characteristics, primarily for the purpose of restricting access to information (e.g., personal/ private records), networks (computer, PBX), or physical premises; and
- spoken language translation, the process of recognizing the speech of a person talking in one language, translating the message content to a second language, and synthesizing an appropriate message in the second language, for the purpose of providing two-way communication between people who do not speak the same language.

To get an appreciation of the progress in each of these areas of voice processing, it is worthwhile to briefly review their current capabilities.

## VOICE CODING

Voice coding technology has been widely used for over two decades in network transmission applications. A key driving factor here has been international standardization of coding algorithms at 64 Kbps ( $\mu$ -law Pulse Code Modulation—G.711), 32 Kbps (Adaptive Differential Pulse Code Modulation—G.721), and 16 Kbps (Low Delay Code Excited Linear Prediction—G.728). Voice coding has also been exploited in cellular systems with the advent of the European GSM standard at 13.2 Kbps, the North American Standard (IS-54, Vector Storage Excitation Linear Prediction) at 8 Kbps, and the promise of the so-called half-rate standards of 6.6 Kbps in Europe and 4 Kbps in North America. Finally, low bit-rate coding for transmission has been a driving force for security applications in the U.S. government, based on standards at 4.8 Kbps (FS 1016, Code Excited Linear Prediction) and 2.4 Kbps (FS 1015, Linear Predictive Coding 10 E).

In the area of voice coding for storage, perhaps the most important application is in the storage of voice messages in voice mailboxes. Typically, most voice mail systems compress the speech to 16 Kbps so as to minimize the total storage requirements of the system while maintaining high-quality voice messages. Another recent application that relies heavily on voice coding is the digital telephone

answering machine in which both voice prompts (including the outgoing message) and voice messages are compressed and stored in the machine's local memory. Current capabilities include tens of seconds of voice prompts and up to about 30 minutes of message storage.

## VOICE SYNTHESIS

Voice synthesis has advanced to the point where virtually any ASCII text message can be converted to speech, providing a message that is completely intelligible, albeit somewhat unnatural (machinelike) in quality. Although the range of applications of voice synthesis is growing rapidly, several key ones have already emerged. One such application is a voice server for accessing electronic mail (e-mail) messages remotely, over a dialed-up telephone line. Such a service is a valuable one for people who travel extensively (especially outside the United States) and who do not have access to computer lines to read their mail electronically. It is also valuable for bridging the "time gap" associated with travel when the working day where you are need not align well with the working day in your home location.

Other interesting and evolving applications of voice synthesis include automated order inquiry (keeping track of the progress of orders); remote student registration (course selection and placement); proofing of text documents ("listening" to your written reports, responses to e-mail, etc.); and providing names, addresses, and telephone numbers in response to directory assistance queries.

## SPEECH RECOGNITION

Although speech recognition technology has made major advancements in the past several years, we are still a long way from the science fiction recognition machines as embodied by Hal in Stanley Kubrick's *2001, A Space Odyssey*, or R2D2 in George Lucas's *Star Wars*. However, our current capability, albeit somewhat limited, has opened up a number of possibilities for improvements in the quality of life in selected areas.

A far-reaching application in speech recognition is the automation of the billing function of operator services whereby all O+ calls that are not dialed directly (e.g., Collect, Person-to-Person, Third-Party Billing, Operated-Assisted, and Calling Card) are handled automatically. Based on calling volumes at the end of 1993, about 4 billion calls per year are handled by speech recognition technology for this application alone. (Most of the remaining 55.5 billion O+ calls

are Calling Card calls, which are normally dialed directly via touch tone responses—rather than voice.)

Other recent applications of speech recognition include toys, cellular voice dialers for automobiles (which promise the ultimate in safety, namely "eyes-free" and "hands-free" communication), voice routing of calls (i.e., replacement for button pushing from touch-tone phones), automatic creation of medical reports (aids to radiologists and medical technicians), order entry (e.g., catalog sales, verification of credit), forms entry (insurance, medical), and even some elementary forms of voice dictation of letters and reports.

## SPEAKER RECOGNITION

Speaker recognition technology is one application where the computer can outperform a human, that is, the ability of a computer to either identify a speaker from a given population or to verify an identity claim from a named speaker, exceeds that of a human trying to perform the same tasks.

Speaker identification is required primarily in some types of forensic applications (e.g., identifying speakers who make obscene phone calls or in criminal investigations). Hence, the number of applications affecting most people's day-to-day lives is limited.

Speaker verification is required for a wide variety of applications that provide secure entree to ATMs (automatic teller machines), PBXs, telecommunications services, banking services, private records, etc. Another interesting (and unusual) application of speaker verification is in keeping track of prison parolees by having an automatic system make calls to the place a parolee is supposed to be and verifying (by voice) that the person answering the call is, in fact, the parolee. Finally, voice verification has been used to restrict entry to buildings, restricted areas, and secure installations, etc., by requiring users to speak "voice passwords" (a throwback to the "open sesame" command in ancient Persia) in order to gain entry.

## SPOKEN LANGUAGE TRANSLATION

Since spoken language translation relies heavily on speech recognition, speech synthesis, and natural language understanding, as well as on text-to-text (or message-to-message) translation, it is an ambitious long-term goal of voice processing technology. However, based on extensive research by NEC and ATR in Japan, AT&T, Carnegie-Mellon University, and IBM in the United States, and Siemens and Telefonica in Europe (among other laboratories), a number of inter-

esting laboratory systems for language translation have evolved. Such systems, although fairly limited in scope and capability, point the way to what "might be" in the future.

One example of such a laboratory language translation system is VEST, the Voice English-Spanish Translator, developed jointly by AT&T and Telefonica; it can handle a limited set of banking and currency transactions. This speaker-trained system, with a vocabulary of about 450 words, was demonstrated continuously at the Seville World's Fair in Spain in 1992 for about 6 months.

The NEC language translation system, which has been successfully demonstrated at two Telecom meetings, deals with tourist information on activities in Japan for foreign visitors. The ATR system is geared to "interpreting telephony," that is, voice translation over dialed-up telephone lines. The ATR system was demonstrated in 1993 via a three-country international call (among Japan, the United States, and Germany) for the purpose of registering for, and getting information about, an international conference.

## NATURAL LANGUAGE PROCESSING

As voice processing technology becomes more capable and more sophisticated, the goals become more far reaching, namely to provide human-like intelligence to every voice transaction with a machine. As such, the machine will ultimately have to go beyond "recognizing or speaking the words"; it will have to understand the meaning of the words, the talker's intent, and perhaps even the talker's state of mind and emotions.

Although we are a long way from being able to understand the nuances of spoken language, or to provide such nuances in synthetic speech, we are able to go "beyond the words" toward understanding the meaning of spoken input via evolving methods of natural language understanding. These include well established methods of syntactic and semantic analysis, pragmatics, and discourse and dialogue analysis. Such natural language understanding provides the bridge between words and concepts, thereby enabling the machine to act properly based on a spoken request and to respond properly with an appropriate action taken by the machine.

## COLLOQUIUM THEME

This volume, *Voice Communication Between Humans and Machines*, follows a colloquium on Human/Machine Communication by Voice held in February 1993. This colloquium, sponsored by the

National Academy of Sciences, examined two rapidly evolving voice processing technologies—voice synthesis and speech recognition along with the natural language understanding needed to integrate these technologies into a speech understanding system. The major purpose of the colloquium was to bring together researchers, system developers, and technologists in order to better understand the strengths and limitations of current technology and to think about both business and technical opportunities for applying the technology. The contents and structure of this book are the same as that of the colloquium.

The colloquium was organized into four sessions each day. The talks on the first day provided a perspective on our current understanding of voice processing in general, and on speech synthesis, speech recognition, and natural language understanding specifically. The talks on the second day discussed applications of the technology in the telecommunications area, in the government and military, in the consumer area, and in aids for handicapped persons. In addition, one session examined the hardware and software constraints in implementing speech systems and at the user interface, which is critical to the deployment and user acceptance of voice processing technology. The last session of the colloquium concentrated on trying to both predict the future and provide a roadmap as to how we might achieve our vision for the technology.

Each session (with the exception of the final one on the second day) consisted of two 30-minute presentations followed by a lively 30-minute discussion period among the session chairman, the speakers, and the audience.

The first session, chaired by Ron Schafer (Georgia Tech), dealt with the scientific bases of human-machine communication by voice. Phil Cohen (SRI International) discussed the role of voice in human-machine communication and argued for interface designs that integrate multiple modes of communication. Jim Flanagan (Rutgers University) provided a comprehensive overview of the history of speech communications from the ancient Greeks (who were fascinated by talking statues) to modern voice communication systems.

The second session, chaired by Mark Liberman (University of Pennsylvania), discussed current understanding of the acoustic and linguistic aspects of speech synthesis. Rolf Carlson (Royal Institute of Technology, Sweden) argued for using speech knowledge from a wide range of disciplines. Jon Allen (MIT) presented the case for merging linguistic models with statistical knowledge obtained from exhaustive analysis of a large tagged database.

The third session, chaired by Steve Levinson (AT&T Bell Laboratories), was on speech recognition technology and consisted of an

overview of current speech recognition techniques by John Makhoul (BBN) and a talk on training and search methods by Fred Jelinek (IBM). Makhoul's talk emphasized the rate of progress in improving performance (as measured in terms of word accuracy) in continuous speech recognition over the past several years and the factors that led to these performance improvements. Jelinek concentrated on the mathematical procedures used to train and decode speech recognizers.

The final session of the first day, chaired by Lynette Hirschman (Massachusetts Institute of Technology), dealt with natural language understanding. Madeleine Bates (BBN) discussed models of natural language understanding and reviewed our current understanding in the areas of syntax, semantics, pragmatics, and discourse. Bob Moore (SRI) discussed the way in which speech can be integrated with natural language as the basis for a speech understanding system.

The two morning sessions on the second day were devoted to applications of the technology and were chaired by Chris Seelbach (Seelbach Associates) and John Oberteuffer (*ASR News*). Excellent overviews of key applications in the areas of telecommunications (Jay Wilpon, AT&T Bell Laboratories), aids for the handicapped (Harry Levitt, CUNY), the military (Cliff Weinstein, MIT Lincoln Laboratory), and consumer electronics (George Doddington, SISTO/DARPA) were given and stimulated lively discussion.

The next session, chaired by David Roe (AT&T Bell Laboratories), concentrated on technical and human requirements for successful technology deployment. Ryohei Nakatsu (NTT) discussed the hardware/software issues, and Candace Kamm (Bellcore) discussed the user interface issues that needed to be addressed and understood.

The final session, titled "Technology 2001," was chaired by Frank Fallside (University of Cambridge) and consisted of three views of where the technology is headed over the next decade and how each speaker thought it would get there. Bishnu Atal (AT&T Bell Laboratories) looked at fundamentally new research directions. Sadaoki Furui (NTT) predicted the directions of research in synthesis and recognition systems. Finally, Mitch Marcus (University of Pennsylvania) discussed new developments in statistical modeling of semantic concepts.

A highlight of the colloquium was an after-dinner talk by Yasuo Kato (NEC) on the future of voice processing technology in the world of computers and communications. Kato, who has contributed to the field for close to 40 years, looked back at how far we have come and gave glimpses of how far we might go in the next decade.

## **SCIENTIFIC BASES OF HUMAN-MACHINE COMMUNICATION BY VOICE**



# Scientific Bases of Human-Machine Communication by Voice

*Ronald W. Schafer\**

## SUMMARY

The scientific bases for human-machine communication by voice are in the fields of psychology, linguistics, acoustics, signal processing, computer science, and integrated circuit technology. The purpose of this paper is to highlight the basic scientific and technological issues in human-machine communication by voice and to point out areas of future research opportunity. The discussion is organized around the following major issues in implementing human-machine voice communication systems: (1) hardware/software implementation of the system, (2) speech synthesis for voice output, (3) speech recognition and understanding for voice input, and (4) usability factors related to how humans interact with machines.

## INTRODUCTION

Humans communicate with other humans in many ways, including body gestures, printed text, pictures, drawings, and voice. But surely voice communication is the most widely used in our daily affairs. Flanagan (1972) succinctly summarized the reasons for this with a pithy quote from Sir Richard Paget (1930):

---

\* Supported by the John and Mary Franklin Foundation.

What drove man to the invention of speech was, as I imagine, not so much the need of expressing his thoughts (for that might have been done quite satisfactorily by bodily gesture) as the difficulty of "talking with his hands full."

Indeed, speech is a singularly efficient way for humans to express ideas and desires. Therefore, it is not surprising that we have always wanted to communicate with and command our machines by voice. What may be surprising is that a paradigm for this has been around for centuries. When machines began to be powered by draft animals, humans discovered that the same animals that provided the power for the machine also could provide enough intelligence to understand and act appropriately on voice commands. For example, the simple vocabulary of *gee*, *haw*, *back*, *giddap*, and *whoa* served nicely to allow a single human to control the movement of a large farm machine. Of course, voice commands were not the only means of controlling these horse or mule-powered machines. Another system of more direct commands was also available through the reins attached to the bit in the animal's mouth. However, in many cases, voice commands offered clear advantages over the alternative. For example, the human was left completely free to do other things, such as walking alongside a wagon while picking corn and throwing it into the wagon. This eliminated the need for an extra person to drive the machine, and the convenience of not having to return to the machine to issue commands greatly improved the efficiency of the operation. (Of course, the reins were always tied in a conveniently accessible place just in case the voice control system failed to function properly!)

Clearly, this reliance on the modest intelligence of the animal source of power was severely limiting, and even that limited voice control capability disappeared as animal power was replaced by fossil fuel power. However, the allure of voice interaction with machines remained and became stronger as technology became more advanced and complex. The obvious advantages include the following:

- Speech is the natural mode of communication for humans.
- Voice control is particularly appealing when the human's hands or eyes are otherwise occupied.
- Voice communication with machines is potentially very helpful to handicapped persons.
- The ubiquitous telephone can be an effective remote terminal for two-way voice communication with machines that can also speak, listen, and understand.

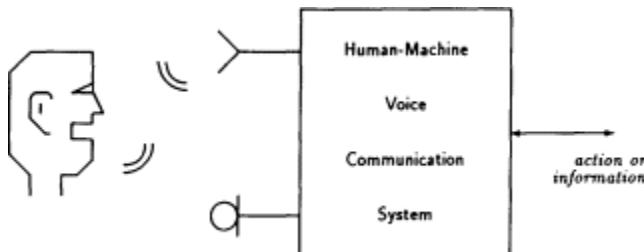


FIGURE 1 Human-machine communication by voice.

Figure 1 depicts the elements of a system for human-machine communication by voice. With a microphone to pick up the human voice and a speaker or headphones to deliver a synthetic voice from the system to the human ear, the human can communicate with the system, which in turn can command other machines or cause desired actions to occur. In order to do this, the voice communication system must take in the human voice input, determine what action is called for, and pass information to other systems or machines. In some cases "recognition" of the equivalent text or other symbolic representation of the speech is all that is necessary. In other cases, such as in natural language dialogue with a machine, it may be necessary to "understand" or extract the meaning of the utterance. Such a system can be used in many ways. In one class of applications, the human controls a machine by voice; for example, the system could do something simple like causing switches to be set or it might gather information to complete a telephone call or it might be used by a human to control a wheelchair or even a jet plane. Another class of applications involves access and control of information; for example, the system might respond to a request for information by searching a database or doing a calculation and then providing the answers to the human by synthetic voice, or it might even attempt to understand the voice input and speak a semantically equivalent utterance in another language.

What are the important aspects of the system depicted in Figure 1? In considering almost all the above examples and many others that we can imagine, there is a tendency to focus on the speech recognition aspects of the problem. While this may be the most challenging and glamorous part, the ability to recognize or understand speech is still only part of the picture. The system also must have the capability to produce synthetic voice output. Voice output can be used to provide feedback to assure the human that the machine has

correctly understood the input speech, and it also may be essential for returning any information that may have been requested by the human through the voice transaction. Another important aspect of the problem concerns usability by the human. The system must be designed to be easy to use, and it must be flexible enough to cope with the wide variability that is common in human speech. Finally, the technology available for implementing such a system must be an overarching concern.

Like many new ideas in technology, voice communication with machine in its modern form may appear to be a luxury that is not essential to human progress and well-being. Some have questioned both the *ultimate feasibility* and the *need* for voice recognition by machine, arguing that anything short of the full capabilities of a native speaker would not be useful or interesting and that such capabilities are not feasible in a machine. The questions raised by Pierce (1969) concerning the goals, the value, and the potential for success of research in speech recognition stimulated much valuable discussion and thought in the late 1960s and may even have dampened the enthusiasm of engineers and scientists for a while, but ultimately the research community answered with optimistic vigor. Although it is certainly true that the ambitious goal of providing a machine with the speaking and understanding capability of a native speaker is still far away, the past 25 years have seen significant progress in both speech synthesis and recognition, so that effective systems for human-machine communication by voice are now being deployed in many important applications, and there is little doubt that applications will increase as the technology matures.

The progress so far has been due to the efforts of researchers across a broad spectrum of science and technology, and future progress will require an even closer linkage between such diverse fields as psychology, linguistics, acoustics, signal processing, computer science, and integrated circuit technology. The purpose of this paper is to highlight the basic scientific and technological issues in human-machine communication by voice and to set the context for the next two papers in this volume, which describe in detail some of the important areas of progress and some of the areas where more research is needed. The discussion is organized around the following major issues in implementing human-machine voice communication systems: (1) hardware/software implementation of the system, (2) speech synthesis for voice output, (3) speech recognition and understanding for voice input, and (4) usability factors related to how humans interact with machines.

## DIGITAL COMPUTATION AND MICROELECTRONICS

Scientists and engineers have systematically studied the speech signal and the speech communication process for well over a century. Engineers began to use this knowledge in the first part of the twentieth century to experiment with ways of conserving bandwidth on telephone channels. However, the invention and rapid development of the digital computer were key to the rapid advances in both speech research and technology. First, computers were used as tools for simulating analog systems, but it soon became clear that the digital computer would ultimately be the only way to realize complex speech signal processing systems. Computer-based laboratory facilities quickly became indispensable tools for speech research, and it is not an exaggeration to say that one of the strongest motivating forces in the modern field of digital signal processing was the need to develop digital filtering, spectrum analysis, and signal modeling techniques for simulating and implementing speech analysis and synthesis systems (Gold and Rader, 1969; Oppenheim and Schafer, 1975; Rabiner and Gold, 1975; Rabiner and Schafer, 1978).

In addition to its capability to do the numerical computations called for in analysis and synthesis of speech, the digital computer can provide the intelligence necessary for human-machine communication by voice. Indeed, any machine with voice input/output capability will incorporate or be interfaced to a highly sophisticated and powerful digital computer. Thus, the disciplines of computer science and engineering have already made a huge impact on the field of human-machine communication by voice, and they will continue to occupy a central position in the field.

Another area of technology that is critically intertwined with digital computation and speech signal processing is microelectronics technology. Without the mind-boggling advances that have occurred in this field, digital speech processing and human-machine communication by voice would still be languishing in the research laboratory as an academic curiosity. As an illustration, [Figure 2](#) shows the number of transistors per chip for several members of a popular family of digital signal processing (DSP) microcomputers plotted as a function of the year the chip was introduced. The upper graph in this figure shows the familiar result that integrated circuit device densities tend to increase exponentially with time, thereby leading inexorably to more powerful systems at lower and lower cost. The lower graph shows the corresponding time required to do a single multiply-accumulate operation of the form (*previous sum + cx[n]*), which is a ubiquitous operation in DSP. From this graph we see that currently avail

able chips can do this combination of operations in 40 nanoseconds or less or the equivalent of 50 MFLOPS (million floating-point operations per second). Because of multiple busses and parallelism in the architecture, such chips can also do hundreds of millions of other operations per second and transfer hundreds of millions of bytes per second. This high performance is not limited to special-purpose microcomputers. Currently available workstations and personal computers also are becoming fast enough to do the real-time operations required for human-machine voice communication without any coprocessor support.

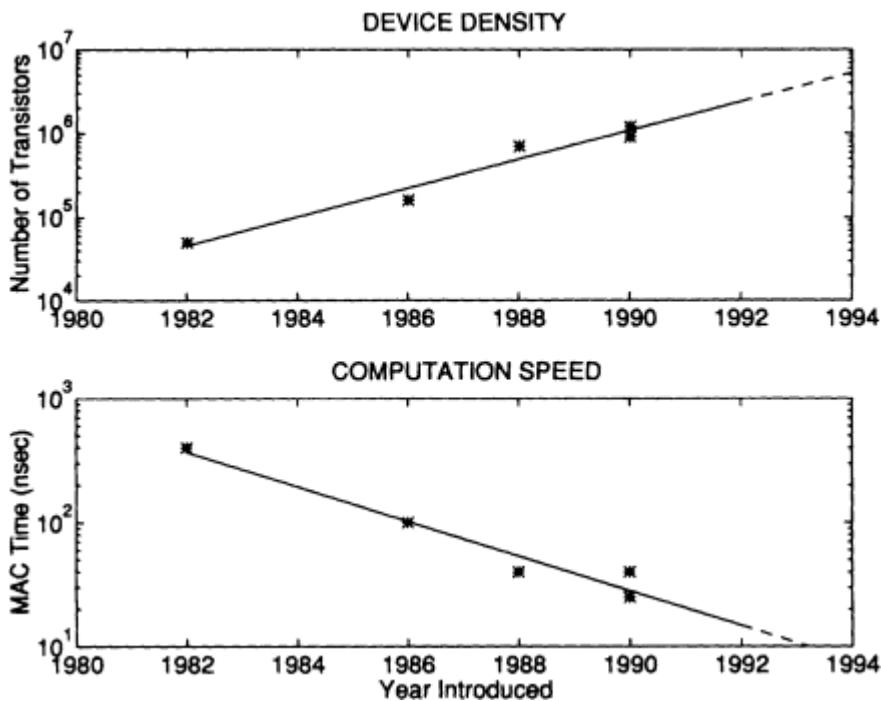


FIGURE 2 Device density and computation speed for a family of DSP microcomputers (data courtesy of Texas Instruments, Inc).

Thus, there is a tight synergism between speech processing, computer architecture, and microelectronics. It is clear that these areas will continue to complement and stimulate each other; indeed, in order to achieve a high level of success in human-machine voice communication, new results must continue to be achieved in areas of computer science and engineering, such as the following:

*Microelectronics.* Continued progress in developing more powerful and sophisticated general-purpose and special-purpose computers is necessary to provide adequate inexpensive computer power for human-machine voice communication applications. At this time, many people in the microelectronics field are confidently predicting chips with a billion transistors by the end of the decade. This presents significant challenges and opportunities for speech researchers to learn how to use such massive information processing power effectively.

*Algorithms.* New algorithms can improve performance and increase speed just as effectively as increased computer power. Current research on topics such as wavelets, artificial neural networks, chaos, and fractals is already finding application in speech processing applications. Researchers in signal processing and computer science should continue to find motivation for their work in the problems of human-machine voice communication.

*Multiprocessing.* The problems of human-machine communication by voice will continue to challenge the fastest computers and the most efficient algorithms. As more sophisticated systems evolve, it is likely that a single processor with sufficient computational power may not exist or may be too expensive to achieve an economical solution. In such cases, multiple parallel processors will be needed. The problems of human-machine voice communication are bound to stimulate many new developments in parallel computing.

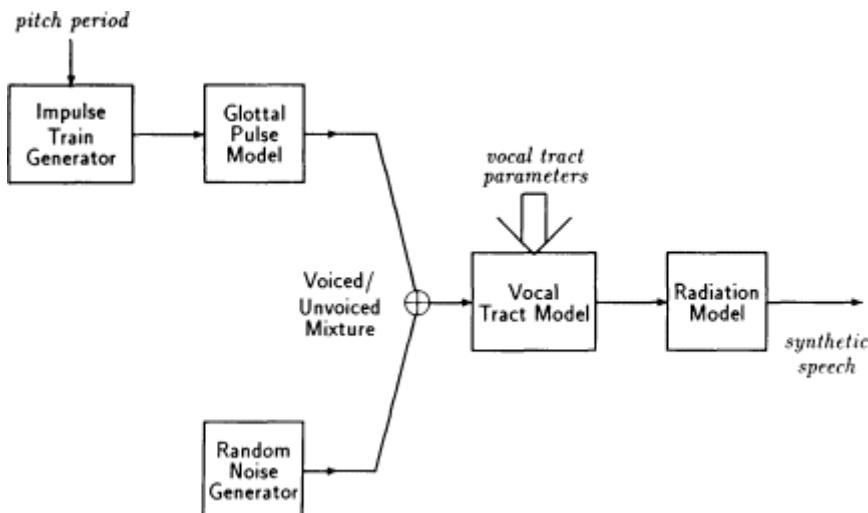
*Tools.* As systems become more complex, the need for computer-aided tools for system development continues to increase. What is needed is an integrated and coordinated set of tools that make it easy to test new ideas and develop new systems concepts, while also making it easy to move a research system through the prototype stage to final implementation. Such tools currently exist in rudimentary form, and it is already possible in some applications to do the development of a system directly on the same DSP microprocessor or workstation that will host the final implementation. However, much more can be done to facilitate the development and implementation of voice processing systems.

## SPEECH ANALYSIS AND SYNTHESIS

In human-machine communication by voice, the basic information-carrying medium is speech. Therefore, fundamental knowledge of the speech signal—how it is produced, how information is encoded in it, and how it is perceived—is critically important.

Human speech is an acoustic wave that is generated by a well-defined physical system. Hence, it is possible using the laws of phys

ics to model and simulate the production of speech. The research in this area, which is extensive and spanning many years, is described in the classic monographs of Fant (1960) and Flanagan (1972), in more recent texts by Rabiner and Schafer (1978) and Deller et al. (1993), and in a wealth of scientific literature. Much of this research has been based on the classic source/system model depicted in [Figure 3](#).



[Figure 3](#): Source/system model for speech production.

In this model the different sounds of speech are produced by changing the mode of excitation between quasi-periodic pulses for voiced sounds and random noise for fricatives, with perhaps a mixture of the two sources for voiced fricatives and transitional sounds. The vocal tract system response also changes with time to shape the spectrum of the signal to produce appropriate resonances or *formants*. With such a model as a basis, the problem of *speech analysis* is concerned with finding the parameters of the model given a speech signal. The problem of *speech synthesis* then can be defined as obtaining the output of the model, given the time-varying control parameters of the model.

A basic speech processing problem is the representation of the analog acoustic waveform of speech in digital form. [Figure 4](#) depicts a general representation of a system for digital speech coding and processing.

Speech, like any other band-limited analog waveform can be sampled and quantized with an analog-to-digital (A-to-D) converter to pro

duce a sequence of binary numbers. These binary numbers represent the speech signal in the sense that they can be converted back to an analog signal by a digital-to-analog (D-to-A) converter, and, if enough bits are used in the quantization and the sampling rate is high enough, the reconstructed signal can be arbitrarily close to the original speech waveform. The information rate (bit rate) of such a digital waveform representation is simply the number of samples per second times the number of bits per sample. Since the bit rate determines the channel capacity required for digital transmission or the memory capacity required for storage of the speech signal, the major concern in digital speech coding is to minimize the bit rate while maintaining an acceptable perceived fidelity to the original speech signal.

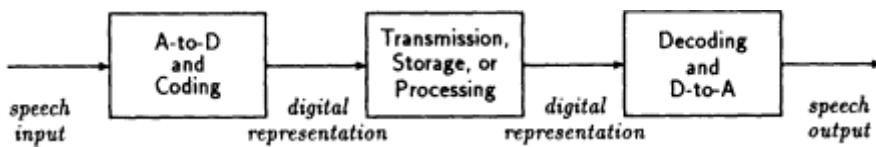


FIGURE 4 Digital speech coding.

One way to provide voice output from a machine is simply to prerecord all possible voice responses and store them in digital form so that they can be played back when required by the system. The information rate of the digital representation will determine the amount of digital storage required for this approach. With a bandwidth of 4000 Hz (implying an 8000-Hz sampling rate) and eight bits per sample (with  $\mu$ -law or A-law compression), speech can be represented by direct sampling and quantization with a bit rate of 64,000 bits/second and with a quality comparable to a good long-distance telephone connection (often called "toll quality"). To further reduce the bit rate while maintaining acceptable quality and fidelity, it is necessary to incorporate knowledge of the speech signal into the quantization process. This is commonly done by an *analysis/synthesis* coding system in which the parameters of the model are estimated from the sampled speech signal and then quantized for digital storage or transmission. A sampled speech waveform is then synthesized by controlling the model with the quantized parameters, and the output of the model is converted to analog form by a D-to-A converter.

In this case the first block in Figure 4 contains the analysis and coding computations as well as the A-to-D, and the third block would contain the decoding and synthesis computations and the D-to-A converter. The output of the discrete-time model of Figure 3 satisfies a linear difference equation; that is, a given sample of the output de

pends linearly on a finite number of previous samples and the excitation. For this reason, linear predictive coding (LPC) techniques have enjoyed huge success in speech analysis and coding. Linear predictive analysis is used to estimate parameters of the vocal tract system model in [Figure 3](#), and, either directly or indirectly, this model serves as the basis for a digital representation of the speech signal. Variations on the LPC theme include adaptive differential PCM (ADPCM), multipulse-excited LPC (MPLPC), code-excited LPC (CELP), self-excited LPC (SEV), mixed-excitation LPC (MELP), and pitch-excited LPC (Deller et al., 1993; Flanagan, 1972; Rabiner and Schafer, 1978). With the exception of ADPCM, which is a waveform coding technique, all the other methods are *analysis/synthesis* techniques. Coding schemes like CELP and MPLPC also incorporate frequency-weighted distortion measures in order to build in knowledge of speech perception along with the knowledge of speech production represented by the synthesis model. Another valuable approach uses frequency-domain representations and knowledge of auditory models to distribute quantization error so as to be less perceptible to the listener. Examples of this approach include sinusoidal models, transform coders, and subband coders (Deller et al., 1993; Flanagan, 1972; Rabiner and Schafer, 1978).

In efforts to reduce the bit rate, an additional trade-off comes into play—that is, the complexity of the analysis/synthesis modeling processes. In general, any attempt to lower the bit rate while maintaining high quality will increase the complexity (and computational load) of the analysis and synthesis operations. At present, toll quality analysis/synthesis representations can be obtained at about 8000 bits/second or an average of about one bit per sample (see Flanagan, in this volume). Attempting to lower the bit rate further leads to degradation in the quality of the reconstructed signal; however, intelligible speech can be reproduced with bit rates as low as 2000 bits/second (see Flanagan, in this volume).

The waveform of human speech contains a significant amount of information that is often irrelevant to the message conveyed by the utterance. An estimate under simple assumptions shows that the fundamental information transmission rate for a human reading text is on the order of 100 bits/second. This implies that speech can in principle be stored or transmitted an order of magnitude more efficiently if we can find ways of representing the phonetic/linguistic content of the speech utterance in terms of the parameters of a speech synthesizer. [Figure 5](#) shows this approach denoted as *text-to-speech synthesis*.

The text of a desired speech utterance is analyzed to determine its phonetic and prosodic variations as a function of time. These in

turn are used to determine the control parameters for the speech model, which then computes the samples of a synthetic speech waveform. This involves literally a pronouncing dictionary (along with rules for exceptions, acronyms, and irregularities) for determining phonetic content as well as extensive linguistic rules for producing durations, intensity, voicing, and pitch. Thus, the complexity of the synthesis system is greatly increased while the bit rate of the basic representation is greatly reduced.

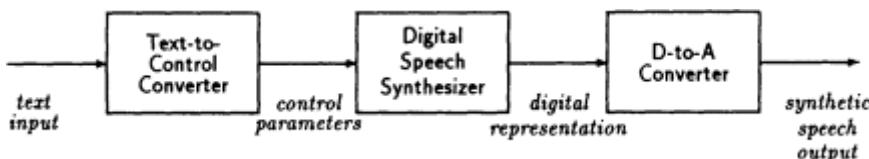


FIGURE 5 Text-to-speech synthesis.

In using digital speech coding and synthesis for voice response from machines, the following four considerations lead to a wide range of trade-off configurations: (1) complexity of analysis/synthesis operations, (2) bit rate, (3) perceived quality, and (4) flexibility to modify or make new utterances. Clearly, straightforward playback of sampled and quantized speech is the simplest approach, requiring the highest bit rate for good quality and offering almost no flexibility other than that of simply splicing waveforms of words and phrases together to make new utterances. Therefore, this approach is usually only attractive where a fixed and manageable number of utterances is required. At the other extreme is text-to-speech synthesis, which, for a single investment in program, dictionary, and rule base storage, offers virtually unlimited flexibility to synthesize speech utterances. Here the text-to-speech algorithm may require significant computational resources. The usability and perceived quality of text-to-synthetic speech has progressed from barely intelligible and "machine-like" in the early days of synthesis research to highly intelligible and only slightly unnatural today. This has been achieved with a variety of approaches ranging from concatenation of diphone elements of natural speech represented in analysis/synthesis form to pure computation of synthesis parameters for physical models of speech production.

Speech analysis and synthesis have received much attention from researchers for over 60 years, with great strides occurring in the 25 years since digital computers became available for speech research. Synthesis research has drawn support from many fields, including acoustics, digital signal processing, linguistics, and psychology. Fu

ture research will continue to synthesize knowledge from these and other related fields in order to provide the capability to represent speech with high quality at lower and lower information rates, leading ultimately to the capability of producing synthetic speech from text that compares favorably with that of an articulate human speaker. Some specific areas where new results would be welcome are the following:

*Language modeling.* A continuing goal must be to understand how linguistic structure manifests itself in the acoustic waveform of speech. Learning how to represent phonetic elements, syllables, stress, emphasis, etc., in a form that can be effectively coupled to speech modeling, analysis, and synthesis techniques should continue to have high priority in speech research. Increased knowledge in this area is obviously essential for text-to-speech synthesis, where the goal is to ensure that linguistic structure is correctly introduced into the synthetic waveform, but more effective application of this knowledge in speech analysis techniques could lead to much improved analysis/synthesis coders as well.

*Acoustic modeling.* The linear source/system model of [Figure 3](#) has served well as the basis for speech analysis and coding, but it cannot effectively capture many subtle nonlinear phenomena in speech. New research in modeling wave propagation in the vocal tract (see Flanagan, in this volume) and new models based on modulation theory, fractals, and chaos (Maragos, 1991; Maragos et al., 1993) may lead to improved analysis and synthesis techniques that can be applied to human-machine communication problems.

*Auditory modeling.* Models of hearing and auditory perception are now being applied with dramatic results in high-quality audio coding (see Flanagan, in this volume). New ways of combining both speech production and speech perception models into speech coding algorithms should continue to be a high priority in research.

*Analysis by synthesis.* The analysis-by-synthesis approach to speech analysis is depicted in [Figure 6](#), which shows that the parametric representation of the speech signal is obtained by adjusting the parameters of the model until the synthetic output of the model matches the original input signal accurately enough according to some error criterion. This principle is the basis for MPLPC, CELP, and SEV coding systems. In these applications the speech synthesis model is a standard LPC source/system model, and the "perceptual comparison" is a frequency-weighted mean-squared error. Although great success has already been achieved with this approach, it should be possible to apply the basic idea with more sophisticated comparison mechanisms based on auditory models and with other signal models.

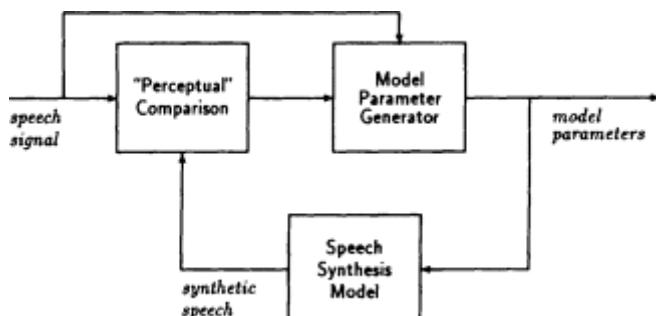


FIGURE 6 Speech analysis by synthesis.

Success will depend on the development of appropriate computationally tractable optimization approaches (see Flanagan, in this volume).

### SPEECH RECOGNITION AND UNDERSTANDING

The capability of recognizing or extracting the text-level information from a speech signal (speech recognition) is a major part of the general problem of human-machine communication by voice. As in the case of speech synthesis, it is critical to build on fundamental knowledge of speech production and perception and to understand how linguistic structure of language is expressed and manifested in the speech signal. Clearly there is much that is common between speech analysis, coding, synthesis, and speech recognition.

[Figure 7](#) depicts the fundamental structure of a typical speech

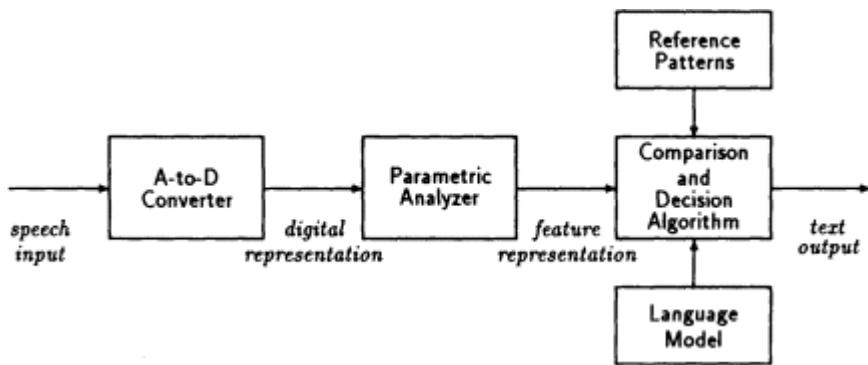


FIGURE 7 Speech recognition.

recognition system. The "front-end" processing extracts a parametric representation or input pattern from the digitized input speech signal using the same types of techniques (e.g., linear predictive analysis or filter banks) that are used in speech analysis/synthesis systems. These acoustic features are designed to capture the linguistic features in a form that facilitates accurate linguistic decoding of the utterance. Cepstrum coefficients derived from either LPC parameters or spectral amplitudes derived from FFT or filter bank outputs are widely used as features (Rabiner and Juang, 1993). Such analysis techniques are often combined with vector quantization to provide a compact and effective feature representation. At the heart of a speech recognition system is the set of algorithms that compare the feature pattern representation of the input to members of a set of stored reference patterns that have been obtained by a training process. Equally important are algorithms for making a decision about the pattern to which the input is closest. Cepstrum distance measures are widely used for comparison of feature vectors, and dynamic time warping (DTW) and hidden Markov models (HMMs) have been shown to be very effective in dealing with the variability of speech (Rabiner and Juang, 1993). As shown in [Figure 7](#), the most sophisticated systems also employ grammar and language models to aid in the decision process.

Speech recognition systems are often classified according to the scope of their capabilities. Speaker-dependent systems must be "trained" on the speech of an individual user, while speaker-independent systems attempt to cope with the variability of speech among speakers. Some systems recognize a large number of words or phrases, while simpler systems may recognize only a few words, such as the digits 0 through 9. Finally, it is simpler to recognize isolated words than to recognize fluent (connected) speech. Thus, a limited-vocabulary, isolated-word, speaker-dependent system would generally be the simplest to implement, while to approach the capabilities of a native speaker would require a large-vocabulary, connected-speech, speaker-independent system. The accuracy of current speech recognition systems depends on the complexity of the operating conditions. Recognition error rates below 1 percent have been obtained for highly constrained vocabulary and controlled speaking conditions; but for large-vocabulary, connected-speech systems, the word error rate may exceed 25 percent.

Clearly, different applications will require different capabilities. Closing switches, entering data, or controlling a wheelchair might very well be achieved with the simplest system. As an example where high level capabilities are required, consider the system depicted in

**Figure 8**, which consists of a speech recognizer producing a text or symbolic representation, followed by storage, transmission, or further processing, and then text-to-speech synthesis for conversion back to an acoustic representation. In this case it is assumed that the output of the text-to-speech synthesis system is sent to a listener at a remote location, such that the machine is simply an intermediary between two humans. This system is the ultimate speech compression system since the bit rate at the text level is only about 100 bits/second. Also shown in **Figure 8** is the possibility that the text might be processed before being sent to the text-to-speech synthesizer. An example of this type of application is when processing is applied to the text to translate one natural language such as English into another such as Japanese. Then the voice output is produced by a Japanese text-to-speech synthesizer, thereby resulting in automatic interpretation in the second language.

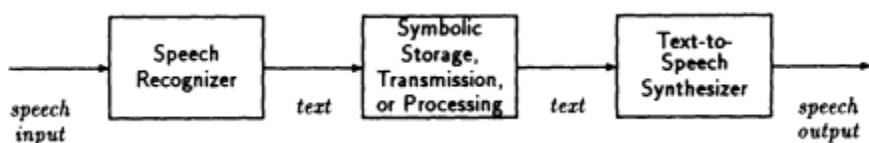


FIGURE 8 Speech recognition/synthesis.

If the goal is to create a machine that can speak and understand speech as well as a human being, then speech synthesis is probably further along than recognition. In the past, synthesis and recognition have been treated as separate areas of research, generally carried out by different people and by different groups within research organizations. Obviously, there is considerable overlap between the two areas, and both would benefit from closer coupling. The following are topics where both recognition and synthesis would clearly benefit from new results:

*Language modeling.* As in the case of speech synthesis, a continuing goal must be to understand how linguistic structure is encoded in the acoustic speech waveform, and, in the case of speech recognition, to learn how to incorporate such models into both the pattern analysis and pattern matching phases of the problem.

*Robustness.* A major limitation of present speech recognition systems is that their performance degrades significantly with changes in the speaking environment, transmission channel, or the condition of the speaker's voice. Solutions to these problems may involve the development of more robust feature representations having a basis in auditory models, new distance measures that are less sensitive to

nonlinguistic variations, and new techniques for normalization of speakers and speaking conditions.

*Computational requirements.* Computation is often a dominant concern in speech recognition systems. Search procedures, hidden Markov model training and analysis, and new feature representations based on detailed auditory models all require much computation. All these aspects and more will benefit from increased processor speed and parallel computation.

*Speaker identity and normalization.* It is clear that speaker identity is represented in the acoustic waveform of speech, but much remains to be done to quantify the acoustic correlates of speaker identity. Greater knowledge in this area would be useful for normalization of speakers in speech recognition systems, for incorporation of speaker characteristics in text-to-speech synthesis, and for its own sake as a basis for speaker identification and verification systems.

*Analysis-by-synthesis.* The analysis-by-synthesis paradigm of [Figure 6](#) may also be useful for speech recognition applications. Indeed, if the block labeled "Model Parameter Generator" were a speech recognizer producing text or some symbolic representation as output, the block labeled "Speech Synthesis Model" could be a text-to-speech synthesizer. In this case the symbolic representation would be obtained as a by-product of the matching of the synthetic speech signal to the input signal. Such a scheme, although appealing in concept, clearly presents significant challenges. Obviously, the matching metric could not simply compare waveforms but would have to operate on a higher level. Defining a suitable metric and developing an appropriate optimization algorithm would require much creative research, and the implementation of such a system would challenge present computational resources.

## USABILITY ISSUES

Given the technical feasibility of speech synthesis and speech recognition, and given adequate low-cost computational resources, the question remains as to whether human-machine voice communication is useful and worthwhile. Intuition suggests that there are many situations where significant improvements in efficiency and performance could result from the use of voice communication/control even of a limited and constrained nature. However, we must be careful not to make assumptions about the utility of human-machine voice communication based on conjecture or our personal experience with human-human communication. What is needed is hard experimental data from which general conclusions can be drawn. In some very

special cases, voice communication with a machine may allow something to be done that cannot be done any other way, but such situations are not the norm. Even for what seem to be obvious areas of application, it generally can be demonstrated that some other means of accomplishing the task either already exists or could be devised. Therefore, the choice usually will be determined by such factors as convenience, accuracy, and efficiency. If voice communication with machines is more convenient or accurate, it may be considered to be worth the extra cost even if alternatives exist. If it is more efficient, its use will be justified by the money it saves.

The scientific basis for making decisions about such questions is at best incomplete. The issues are difficult to quantify and are not easily encapsulated in a neat theory. In many cases even careful experiments designed to test the efficacy of human-machine communication by voice have used humans to simulate the behavior of the machine. Some of the earliest work showed that voice communication capability significantly reduced the time required to perform tasks involving simulated human-computer interaction (Chapanis, 1975), and subsequent research has added to our understanding. However, widely applicable procedures for the design of human-machine voice communication systems are not yet available. The paper by Cohen and Oviatt in this volume is a valuable contribution because it summarizes the important issues and current state of knowledge on human-machine interaction and points the way to research that is needed as a basis for designing systems.

The paradigm of the voice-controlled team and wagon has features that are very similar to those found in some computer-based systems in use today—that is, a limited vocabulary of acoustically distinct words, spoken in isolation, with an alternate communication/control mechanism conveniently accessible to the human in case it is necessary to override the voice control system. Given a computer system with such constrained capabilities, we could certainly go looking for applications for it. In the long-term, however, a much more desirable approach would be to determine the needs of an application and then specify the voice communication interface that would meet the needs effectively. To do this we must be in a better position to understand the effect on the human's performance and acceptance of the system of such factors as:

- vocabulary size and content,
- fluent speech vs. isolated words,
- constraints on grammar and speaking style,
- the need for training of the recognition system,
- the quality and naturalness of synthetic voice response,

- the way the system handles its errors in speech understanding, and
- the availability and convenience of alternate communication modalities.

These and many other factors come into play in determining whether humans can effectively use a system for voice communication with a machine and, just as important, whether they will prefer using voice communication over other modes of communication that might be provided.

*Mixed-mode communication.* Humans soon tire of repetition and welcome anything that saves steps. Graphical interfaces involving pointing devices and menus are often tedious for repetitive tasks, and for this reason most systems make available an alternate shortcut for entering commands. This is no less true for voice input; humans are also likely to tire of talking to their machines. Indeed, sometimes we would even like for the machine to anticipate our next command—for example, something like the well-trained team of mules that automatically stopped the corn-picking wagon as the farmer fell behind and moved it ahead as he caught up (William H. Schafer, personal communication, 1993). While mind reading may be out of the question, clever integration of voice communication with alternative sensing mechanisms, alternate input/output modalities, and maybe even machine learning will ultimately lead to human-machine interfaces of greatly improved usability.

*Experimental capabilities.* With powerful workstations and fast coprocessors readily available, it is now possible to do real-time experiments with real human-machine voice communication systems. These experiments will help answer questions about the conditions under which these systems are most effective, about how humans learn to use human-machine voice communication systems, and about how the interaction between human and machine should be structured; then the new "theory of modalities" called for by Cohen and Oviatt (in this volume) may begin to emerge.

## CONCLUSION

Along the way to giving machines human-like capability to speak and understand speech there remains much to be learned about how structure and meaning in language are encoded in the speech signal and about how this knowledge can be incorporated into usable systems. Continuing improvement in the effectiveness and naturalness of human-machine voice communication systems will depend on cre

ative synthesis of concepts and results from many fields, including microelectronics, computer architecture, digital signal processing, acoustics, auditory science, linguistics, phonetics, cognitive science, statistical modeling, and psychology.

## REFERENCES

- Chapanis, A., "Interactive Human Communication," *Scientific American*, vol. 232, pp. 36-49, March 1975.
- Deller, J. R., Jr., Proakis, J. G., and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Co., New York, 1993.
- Fant, G., *Acoustic Theory of Speech Production*, Mouton & Co. N.V., The Hague, 1960.
- Flanagan, J. L., *Speech Analysis, Synthesis, and Perception*, Springer Verlag, New York, 1972.
- Gold, B., and C. M. Rader, *Digital Processing of Signals*, McGraw-Hill, New York, 1969.
- Maragos, P., "Fractal Aspects of Speech Signals: Dimension and Interpolation," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 417-420, Toronto, May 1991.
- Maragos, P.A., J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Transaction on Signal Processing*, in press.
- Oppenheim, A.V., and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
- Paget, R., *Human Speech*, Harcourt, New York, 1930.
- Pierce, J.R., "Whither Speech Recognition?," *Journal of the Acoustical Society of America*, vol. 47, no. 6 (part 2), pp. 1049-1050, 1969.
- Rabiner, L.R., and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- Rabiner, L.R., and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
- Rabiner, L.R., and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

# The Role of Voice in Human-Machine Communication\*

*Philip R. Cohen and Sharon L. Oviatt*

## SUMMARY

Optimism is growing that the near future will witness rapid growth in human-computer interaction using voice. System prototypes have recently been built that demonstrate speaker-independent real-time speech recognition and understanding of naturally spoken utterances in moderately sized vocabularies (1000 to 2000 words), and larger-vocabulary speech recognition systems are on the horizon. Already, computer manufacturers are building speech recognition subsystems into their new product lines. However, before this technology will be broadly useful, a substantial knowledge base about human spoken language and performance during computer-based interaction needs to be gathered and applied. This paper reviews application areas in which spoken interaction may play a significant role, assesses potential benefits of spoken interaction with machines, and attempts to compare voice with alternative and complementary modalities of human-computer interaction. The paper also discusses information that will be needed to build a firm empirical foundation for future designing of human-computer interfaces. Finally, it argues for a more systematic and scientific approach to understanding human language and performance with voice interactive systems.

---

\* The writing of this paper was supported in part by a grant from the National Science Foundation (No. IRI-9213472) to SRI International.

## INTRODUCTION

From the beginning of the computer era, futurists have dreamed of the conversational computer—a machine that we could engage in spoken natural language conversation. For instance, Turing's famous "test" of computational intelligence imagined a computer that could conduct such a fluent English conversation that people could not distinguish it from a human. However, despite prolonged research and many notable scientific and technological achievements, until recently there have been few human-computer dialogues, none spoken. This situation has begun to change, as steady progress in speech recognition and natural language processing technologies, supported by dramatic advances in computer hardware, has made possible laboratory prototype systems with which one can engage in simple question-answer dialogues. Although far from human-level conversation, this initial capability is generating considerable interest and optimism for the future of human-computer interaction using voice.

This paper aims to identify applications for which spoken interaction may be advantageous, to situate voice with respect to alternative and complementary modalities of human-computer interaction, and to discuss obstacles that exist to the successful deployment of spoken language systems because of the nature of spoken language interaction.

Two general sorts of speech input technology are considered. First, we survey a number of existing applications of speech *recognition* technologies, for which the system identifies the words spoken, but need not understand the meaning of what is being said. Second, we concentrate on applications that will require a more complete *understanding* of the speaker's intended meaning, examining future spoken dialogue systems. Finally, we discuss how such speech understanding will play a role in future human-computer interactions, particularly those involving the coordinated use of multiple communication modalities, such as graphics, handwriting, and gesturing. It is argued that progress has been impeded by the lack of adequate scientific knowledge about human spoken interactions, especially with computers. Such a knowledge base is essential to the development of well-founded human-interface guidelines that can assist system designers in producing successful applications incorporating spoken interaction. Given recent technological developments, the field is now in a position to systematically expand that knowledge base.

## Background and Definitions

Human-computer interaction using voice may involve speech input or speech output, perhaps in combination with each other or with other modalities of communication.

### Speech Analysis

The speech analysis task is often characterized along five dimensions:

*Speaker dependence.* Speech recognizers are described as speaker-dependent/trained, speaker-adaptive, and speaker-independent. For speaker-dependent recognition, samples of a given user's speech are collected and used as models for his/her subsequent utterances. For speaker-adaptive recognition, parameterized acoustical models are initially available, which can be more finely tuned for a given user through pronunciation of a limited set of specified utterances. Finally, speaker-independent recognizers are designed to handle any user's speech, without training, in the given domain of discourse (see Flanagan, in this volume).

*Speech continuity.* Utterances can be spoken in an isolated manner, with breaks between words, or as continuous natural speech.

*Speech type.* To develop initial algorithms, researchers typically first use read speech as data, in which speakers read random sentences drawn from some corpus, such as the *Wall Street Journal*. Subsequent to this stage of algorithm development, speech recognition research attempts to handle spontaneous speech, in which speakers construct new utterances in the chosen domain of discourse.

*Interactivity.* Certain speech recognition tasks, such as dictation, can be characterized as noninteractive, in that the speaker is receiving no feedback from the intended listener(s). Other systems are designed to process interactive speech, in which speakers construct utterances as part of an exchange of turns with a system or with another speaker.

*Vocabulary and grammar.* The user can speak words from a tightly constrained vocabulary and grammar or from larger vocabularies and grammars that more closely approximate those of a natural language. The system's vocabulary and grammar can be chosen by the system designer or application developer, or they can be compiled from data based on actual users speaking either to a simulated system or to an early system prototype. Current speech recognition technologies require an estimate of the probability of occurrence of each word in the

context of the other words in the vocabulary. Because these probabilities are typically approximated from the distribution of words in a given corpus, it is currently difficult to expand a system's vocabulary, although research is proceeding on vocabulary-independent recognition (Hon and Lee, 1991).

Vendors often describe their speech recognition hardware as offering very high recognition accuracy, but it is only in the context of a quantitative understanding of the recognition task that one can meaningfully compare the performance of recognizers. To calibrate the difficulty of a given recognition task for a given system, researchers have come to use a measure of the *perplexity* of that system's language model, which measures, roughly speaking, the average number of word possibilities at each state of the grammar (Bahl et al., 1983; Baker, 1975; Jelinek, 1976). Word recognition accuracy has been found, in general, to be inversely proportional to perplexity. Most commercial systems offer speech recognition systems claiming >95 percent word recognition accuracy given a perplexity on the order of 10. At least one vendor offers a 1000 to 5000 word, speaker-independent system, with perplexities in the range of 66 to 433, and a corresponding word-recognition error of 3 to 15 percent for recognition of isolated words (Baker, 1991). Current laboratory systems support real-time, speaker-independent recognition of continuously spoken utterances drawn from a vocabulary of approximately 1500 words, with a perplexity of 50 to 70, resulting in word recognition error rates between 4 and 8 percent (Pallett et al., 1993). The most ambitious speaker-independent systems are currently recognizing, in real-time, read speech drawn from a 5000-word vocabulary of *Wall Street Journal* text, with a perplexity of 120, resulting in a word recognition error rate of 5 percent (Pallett et al., 1993). Larger vocabularies are now being attempted.

The end result of voice recognition is the highest-ranking string(s) of words, or often lattice of words, that covers the signal. For small vocabularies and tightly constrained grammars, a simple interpreter can respond to the spoken words directly. However, for larger vocabularies and more natural grammars, *natural language understanding* must be applied to the output of the recognizer in order to recover the intended meaning of the utterance.<sup>1</sup> Because this natural language understanding process is complex and open ended, it is often constrained by the application task (e.g., retrieving information from a data base) and by the domain of discourse (e.g., a data base

---

<sup>1</sup> See Moore (in this volume) for a discussion of how these components can be integrated.

about airline flights). Here the combination of speech recognition and language understanding will be termed *speech understanding*, and the systems that use such input will be termed *spoken language systems*. This paper reviews earlier work on the uses of speech recognition but concentrates on the uses of spoken language.

## Speech Synthesis

Three forms of speech synthesis technology exist:

*Digitized speech.* To produce an utterance, the machine assembles and plays back previously recorded and compressed samples of human speech. Although a noticeable break between samples can often be heard, and the overall intonation may be inaccurate, such a synthesis process can offer human-sounding speech of high intelligibility. This process is, however, limited to producing combinations of the recorded samples.

*Text-to-speech.* Text-to-speech synthesis involves an automated analysis of the structure of words into their morphological constituents. By combining the pronunciations of those subword units according to letter- and morph-to-sound rules, coupled with a large list of exceptional pronunciations (for English), arbitrary text can be rendered as speech. Because this technology can handle open-ended text, it is suitable for large-scale applications, such as reading text aloud to blind users or reading electronic mail over the telephone. Text-to-speech science and technology are covered at length elsewhere in this volume (see Allen, in this volume, and Carlson, in this volume).

*Concept-to-speech.* With text-to-speech systems, the text to be converted is supplied from a human source. Future dialogue systems will require computers to decide for themselves what to say and how to say it in order to arrive at a meaningful and contextually appropriate dialogue contribution. Such systems need to determine what speech action(s) to perform (e.g., request, suggestion), how to refer to entities in the utterance, what to say about them, what grammatical forms to use, and what intonation to apply. Moreover, the utterance should contribute to the course of the dialogue, so the system should keep a representation of what it has said in order to analyze and understand the user's subsequent utterances.

The research areas of speech synthesis and language generation have received considerably less attention than speech recognition and understanding but will increase in importance as the possibility of developing spoken dialogue systems becomes realizable.

The remainder of this paper explores current and future applica

tion areas in which spoken interaction may be a preferred modality of communication with computers. First, factors that may influence the desirability and efficiency of voice-based interaction with computers are identified, independent of whether a simple command language or a quasi-natural language is being spoken. Then, we discuss spoken language interaction, comparing it both to keyboard-based interaction and to the currently dominant graphical user-interface paradigm. After identifying circumstances that favor spoken language interaction, gaps in the scientific knowledge base of spoken communication are identified that present obstacles to the development of spoken language-based systems. It is observed that future systems will be multimodal, with voice being only one of the communication modalities available. We conclude with suggestions for further research that needs to be undertaken to support the development of voice-based unimodal and multimodal systems and argue that there is a pressing need to create empirically based human interface guidelines for system developers before voice-based technology can fulfill its potential.

### **WHEN IS SPOKEN INTERACTION WITH COMPUTERS USEFUL?**

As yet there is no theory or categorization of tasks and environments that would predict, all else being equal, when voice would be a preferred modality of human-computer communication. Still, a number of situations have been identified in which spoken communication with machines may be advantageous:

- when the user's hands or eyes are busy,
- when only a limited keyboard and/or screen is available,
- when the user is disabled,
- when pronunciation is the subject matter of computer use, or
- when natural language interaction is preferred.

We briefly examine the present and future roles of spoken interaction with computers for these environments. Because spoken natural language interaction is the most difficult to implement, we discuss it extensively in the section titled "Natural Language Interaction."

### **Voice Input**

#### **Hands/Eyes-Busy Tasks**

The classic situation favoring spoken interaction with machines is one in which the user's hands and/or eyes are busy performing

some other task. In such circumstances, by using voice to communicate with the machine, people are free to pay attention to their task, rather than breaking away to use a keyboard. Field studies suggest that, for example, F-16 pilots who can attain a high speech recognition rate can perform missions, such as formation flying or low-level navigation, faster and more accurately when using spoken control over various avionics subsystems, as compared with keyboard and multifunction-button data entry (Howard, 1987; Rosenhoover et al., 1987; Williamson, 1987). Similar results have been found for helicopter pilots in noisy environments during tracking and communications tasks (Simpson et al., 1982, 1985; Swider, 1987).<sup>2</sup>

Commercial hands/eyes-busy applications also abound. For instance, wire installers, who spoke a wire's serial number and then were guided verbally by computer to install that wire achieved a 20 to 30 percent speedup in productivity, with improved accuracy and lower training time, over their prior manual method of wire identification and installation (Marshall, 1992). Parcel sorters who spoke city names instead of typing destination-labeled keys attained a 37 percent improvement in entry time during hands/eyes-busy operations (Visick et al., 1984). However, when the hands/eyes-busy component of parcel sorting was removed, spoken input offered no distinct speed advantages. In addition, VLSI circuit designers were able to complete 24 percent more tasks when spoken commands were available than when they only used a keyboard and mouse interface (see the section titled "Direct Manipulation") (Martin, 1989). Although individual field studies are rarely conclusive, many field studies of highly accurate speech recognition systems with hands/eyes-busy tasks have found that spoken input leads to higher task productivity and accuracy.

Not only does spoken input offer efficiency gains for a given hands/eyes-busy task, it also offers the potential to change the nature of that task in beneficial ways. For example, instead of having to remember and speak or type the letters "YYZ" to indicate a destination airport, a baggage handler could simply say "Toronto," thereby using an easy-to-remember name (Martin, 1989; Nye, 1982). Similar potential advantages are identified for voice-based telephone dialers, to which one can say "Call Tom," rather than having to remember and input a phone number (Rabiner et al., 1980). Other hands/eyes-busy applications that might benefit from voice interaction include data entry and machine control in factories and field applications

---

<sup>2</sup> Further discussion of speech recognition for military environments can be found in (Weinstein, 1991, in this volume).

(Martin, 1976), access to information for military command-and-control, astronauts' information management during extravehicular access in space, dictation of medical diagnoses (Baker, 1991), maintenance and repair of equipment, control of automobile equipment (e.g., radios, telephones, climate control), and navigational aids (Streeter et al., 1985).

A major factor determining success for speech input applications is speech recognition accuracy. For example, the best task performance reported during F-16 test flights was obtained once pilots attained isolated word recognition rates greater than 95 percent. Below 90 percent, the effort needed to correct recognition errors was said to outweigh the benefits gained for the user (Howard, 1987). Similar results showing the elimination of benefits once error correction is considered also have been found in tasks as simple as entering connected digits (Hauptmann and Rudnick, 1990).

To attain a sufficiently high level of recognition accuracy in field tests, spoken input has been severely constrained to allow only a small number of possible words at any given time. Still, even with such constraints, accuracy in the field often lags that of laboratory tests because of many complicating factors, such as the user's physical and emotional state, ambient noise, microphone equipment, the demands of real tasks, methods of the user and system training, and individual differences encountered when an array of real users is sampled. However, it is claimed that most failures of speech technology have been the result of human factors engineering and management (Lea, 1992), rather than low recognition accuracy per se. Human factors issues are discussed further below and by Kamm (in this volume).

### Limited Keyboard/Screen Option

The most prevalent current uses of speech synthesis and recognition are telephone-based applications. Speech synthesizers are commonly used in the telecommunications industry to support directory assistance by speaking the desired telephone number to the caller, thereby freeing the operator to handle another call. Speech recognizers have been deployed to replace or augment operator services (e.g., collect calls), handling hundreds of millions of callers each year and resulting in multimillion dollar savings (Lennig, 1989; Nakatsu, in this volume; Wilpon, in this volume). Speech recognizers for telecommunications applications accept a very limited vocabulary, perhaps spotting only certain key words in the input, but they need to function with high reliability for a broad spectrum of the general

public. Although not as physically severe as avionic or manufacturing applications, telecommunications applications are difficult because callers receive little or no training about use of the system and may have low-quality equipment, noisy telephone lines, and unpredictable ambient noise levels. Moreover, caller behavior is difficult to predict and channel (Basson, 1992; Kamm, in this volume; Spitz, 1991).<sup>3</sup>

The considerable success at automating the simpler operator services opens the possibility for more ambitious telephone-based applications, such as information access from remote databases. For example, the caller might inquire about airline and train schedules (Advanced Research Projects Agency, 1993; Proceedings of the Speech and Natural Language Workshop, 1991; Peckham, 1991), yellow pages information, or bank account balances (Nakatsu, in this volume), and receive the answer auditorily. This general area of human-computer interaction is much more difficult to implement than simple operator services because the range of caller behavior is quite broad and because speech understanding and dialogue participation are required rather than just word recognition. When even modest quantities of data need to be conveyed, a purely vocal interaction may be difficult to conduct, although the advent of "screen phones" may well improve such cases.

Perhaps the most challenging potential application of telephone-based spoken language technology is the interpretation of telephony (Kurematsu, 1992; Roe et al., 1991) in which two callers speaking different languages can engage in a dialogue mediated by a spoken language translation system (Kitano, 1991; Yato et al., 1992). Such systems are currently designed to incorporate speech recognition, machine translation, and speech synthesis subsystems and to interpret one sentence at a time. A recent initial experiment organized by ATR International (Japan), with Carnegie-Mellon University (USA) and Siemens A.G. (Germany) involved Japanese-English and Japanese-German machine-interpreted dialogues (Pollack, 1993; Yato et al., 1992). Utterances in one language were recognized and translated by a local computer, which sent a translated textual rendition to the foreign site, where text-to-speech synthesis took place. AT&T has demonstrated a limited-domain spoken English-Spanish translation system (Roe et al., 1991), although not a telephone-based one, and Nippon Electric Corporation has demonstrated a similar Japanese-English system.

Apart from the use of telephones, a second equipment-related

---

<sup>3</sup> An excellent review of the human factors and technical difficulties encountered in telecommunications applications of speech recognition can be found in Karis and Dobroth (1991).

factor favoring voice-based interaction is the ever-decreasing size of portable computers. Portable computing and communications devices will soon be too small to allow for use of a keyboard, implying that the input modalities for such machines will most likely be digitizing pen and voice (Crane, 1991; Oviatt, 1992), with screen and voice providing system output. Given that these devices are intended to supplant both computer and telephone, users will already be speaking *through* them. A natural evolution of the devices will offer the user the capability to speak *to* them as well.

Finally, an emerging use of voice technology is to replace the many control buttons on consumer electronic devices (e.g., VCRs, receivers). As the number of user-controllable functions on these devices increases, the user interface becomes overly complex and can lead to confusion over how to perform even simple tasks. Products have recently been announced that allow users to program their devices using simple voice commands.

## Disability

A major potential use of voice technology will be to assist deaf users in communicating with the hearing world using a telephone (Bernstein, 1988). Such a system would recognize the hearing person's speech, render it as text, and synthesize the deaf person's textual reply (if using a computer terminal) as a spoken utterance. Another use of speech recognition in assisting deaf users would be captioning television programs or movies in real-time. Speech recognition could also be used by motorically impaired users to control suitably augmented household appliances, wheelchairs, and robotic prostheses. Text-to-speech synthesis can assist users with speech and motor impediments; can assist blind users with computer interaction; and, when coupled with optical character recognition technology, can read printed materials to blind users. Finally, given sufficiently capable speech recognition systems, spoken input may become a prescribed therapy for repetitive stress injuries, such as carpal tunnel syndrome, which are estimated to afflict approximately 1.5 percent of office workers in occupations that typically involve the use of keyboards (Tanaka et al., 1993), although speech recognizers may themselves lead to different repetitive stress injuries (Markinson, personal communication, 1993).<sup>4</sup>

---

<sup>4</sup> The general subject of "assistive technology" is covered at length by H. Levitt (in this volume), and a survey of speech recognition for rehabilitation can be found in Bernstein (1988).

## Subject Matter Is Pronunciation

Speech recognition will become a component of future computer-based aids for foreign language learning and for the teaching of reading (Bernstein and Rtischev, 1991; Bernstein et al., 1990; Mostow et al., 1993). For such systems, speakers' pronunciation of computer-supplied texts would be analyzed and given as input to a program for teaching reading or foreign languages. Whereas these may be easier applications of speech recognition than some because the words being spoken are supplied by the computer, the recognition system will still be confronted with mispronunciations and slowed pronunciations, requiring a degree of robustness not often considered in other applications of speech recognition. Substantial research will also be needed to develop and field test new educational software that can take advantage of speech recognition and synthesis for teaching reading. This is perhaps one of the most important potential applications of speech technology because the societal implications of raising literacy levels on a broad scale are enormous.

## Voice Output

As with speech input, the factors favoring voice output are only informally understood. Just as tasks with a high degree of visual or manual activity may be more effectively accomplished using spoken input, such tasks may also favor spoken system output. A user could concentrate on a task rather than altering his or her gaze to view a system display. Typical application environments include flying a plane, in which the pilot could receive information about the status of the plane's subsystems during critical phases of operation (e.g., landing, high-speed maneuvering), and driving a car, in which the driver would be receiving navigational information in the course of driving. Other factors thought to favor voice output include remote access to information services over the telephone, lack of reading skills, darkened environments, and the need for omnidirectional information presentation, as in the issuing of warnings in cockpits, control rooms, factories, etc. (Simpson et al., 1985; Thomas et al., 1984).

There are numerous studies of speech synthesis, but no clear picture has emerged of when computer-human communication using speech output is most effective or preferred. Psychological research has investigated the intelligibility, naturalness, comprehensibility, and recallability of synthesized speech (Luce et al., 1983; Nusbaum and Schwab, 1983; Simpson et al., 1985; Thomas et al., 1984). Intelligibil

ity and naturalness are orthogonal dimensions in that synthetic speech present in an environment of other human voices may be intelligible but unnatural. Conversely, human speech in a noisy environment may be natural but unintelligible (Simpson et al., 1985). Many factors influence the intelligibility of synthesized speech in an actual application environment, including the baseline phoneme intelligibility, speaking rate, signal-to-noise level, and presence of other competing voices, as well as the linguistic and pragmatic contexts (Simpson and Navarro, 1984; Simpson et al., 1985).

The desirability of voice output depends on the application environment. Pilots prefer to hear warnings with synthetic speech rather than digitized speech, as the former is more easily distinguished from other voices, such as radio traffic (Voorhees et al., 1983). However, in simulations of air traffic control systems, in which pilots would expect to interact with a human, digitized human speech was preferred to computer synthesized speech (Simpson et al., 1985). Users may prefer to receive information visually, either on a separate screen or on a heads-up display (Swider, 1987), reserving spoken output for critical warning messages (Simpson et al., 1985). Much more research is required in order to determine those types of information processing environments for which spoken output is beneficial and preferred. Furthermore, rather than just concentrating on the benefits of speaking an utterance as compared with other modes of presenting the same information, future research needs to evaluate user performance and preferences as a function of the *content* of what is being communicated, especially if the computer will be determining that content (e.g., the generation of navigational instructions for drivers). Finally, research is critically necessary to develop algorithms for determining the appropriate intonation contours to use during a spoken human-computer dialogue.

### Summary

There are numerous existing applications of voice-based human-computer interaction, and new opportunities are developing rapidly. In many applications for which the user's input can be constrained sufficiently to allow for high recognition accuracy, voice input has been found to lead to faster task performance with fewer errors than keyboard entry. Unfortunately, no principled method yet exists to predict when voice input will be the most effective, efficient, or preferred modality of communication. Similarly, no comprehensive analysis has identified the circumstances when voice will be the preferred or most efficient form of computer output, though again hands/eyes-

busy tasks may also be among the leading candidates for voice output.

One important circumstance favoring human-computer communication by voice is when the user wishes to interact with the machine in a natural language, such as English. The next section discusses such spoken language communication.

## **COMPARISON OF SPOKEN LANGUAGE WITH OTHER COMMUNICATION MODALITIES**

A user who will be speaking to a machine may expect to be able to speak in a natural language, that is, to use ordinary linguistic constructs such as noun and verb phrases. Conversely, if natural language interaction is chosen as a modality of human-computer communication, users may prefer to speak rather than type. In either case, users may expect to be able to engage in a dialogue, in which each party's utterance sets the context for interpreting subsequent utterances. We first discuss the status of the development of spoken language systems and then compare spoken language interaction with typed interaction.

### **Spoken Language System Prototypes**

Research is progressing on the development of spoken language question answering systems—systems that allow users to speak their questions freely and which then understand those questions and provide an accurate reply. The Advanced Research Projects Agency-supported air travel information systems, ATIS (Advanced Research Projects Agency, 1993), developed at Bolt, Beranek, and Newman (Kubala et al., 1992), Carnegie-Mellon University (Huang et al., 1993), the Massachusetts Institute of Technology (Zue et al., 1992), SRI International (Appelt and Jackson, 1992), and other institutions, allow novice users to obtain information in real-time from the Official Airline Guide database, through the use of speaker-independent, continuously spoken English questions. The systems recognize the words in the user's utterance, analyze the meaning of those utterances, often in spite of word recognition errors, retrieve information from (a subset of) the Official Airline Guide's database, and produce a tabular set of answers that satisfy the question. These systems respond with the correct table of flights for over 70 percent of context-independent questions, such as "Which flights depart from San Francisco for Washington after 7:45 a.m.?" Rapid progress has been made in the development of these systems, with a 4-fold reduction in weighted error rates rec

ognition over a 20-month period for speech recognition, a 3.5-fold reduction over a 30-month period for natural language understanding, and a 2-fold reduction over a 20-month period for their combination as a spoken language understanding system. Other major efforts to develop spoken dialogue systems are ongoing in Europe (Mariani, 1992; Peckham, 1991) and Japan (Yato et al., 1992).

Much of the language processing technology used for spoken language understanding has been based on techniques for keyboard-based natural language systems.<sup>5</sup> However, spoken input presents qualitatively different problems for language understanding that have no analog in keyboard interaction.

### Spoken Language vs. Typed Language

#### Research Methodology

In our review of findings about linguistic communication relevant to spoken human-computer interaction, some results are based on analyses of human-human interaction, some are based on human-to-simulated-computer interaction, and some are based on human-computer interaction. Studies of human-human communication can identify the communicative capabilities that people bring to their interactions with computers and can show what could be achieved were computers adequate conversationalists. However, because this level of conversational competence will be unachievable for some time, scientists have developed techniques for simulating computer systems that interact via spoken language (Andry et al., 1990; Fraser and Gilbert, 1991; Gould et al., 1983; Guyomard and Siroux, 1988; Leiser, 1989; Oviatt et al., 1992, 1993a; Pavan and Pelletti, 1990; Price, 1990) by using a concealed human assistant who responds to the spoken language. With this method, researchers can analyze people's language, dialogue, task performance, and preferences before developing fully functional systems.

Important methodological issues for such simulations include providing accurate and rapid response, and training the simulation assistant to function appropriately. Humans engage in rapid spoken interaction and bring expectations for speed to their interaction with computers. Slow interactions can cause users to interrupt the system.

---

<sup>5</sup> For a discussion of the state of research and technology of natural language processing, see Bates (in this volume).

with repetitions while the system is processing their earlier input (VanKatwijk et al., 1979) and, it is conjectured, can also elicit phenomena characteristic of noninteractive speech (Oviatt and Cohen, 1991a). One technique used to speed up such voice-in/voice-out simulations is the use of a vocoder, which transforms the assistant's naturally spoken response into a mechanical-sounding utterance (Fraser and Gilbert, 1991; Guyomard and Siroux, 1988). The speed of the "system" is thus governed by the assistant's knowledge and reaction time, as well as the task at hand, but not by speech recognition, language understanding, and speech synthesis. However, because people speak differently to a computer than they do to a person (Fraser and Gilbert, 1991), even to prompts for simple yes/no answers (Basson, 1992; Basson et al., 1989), the assistant should not provide too *intelligent* a reply, as this might reveal the "system" as a simulation. A second simulation method, which both constrains the simulation assistant and supports a rapid response, is to provide the assistant with certain predefined fields and structures on the screen that can be selected to reply to the subject (Andry et al., 1990; Dahlback et al., 1992; Leiser, 1989; Oviatt et al., 1992). More research is needed into the development of simulation methodologies that can accurately model spoken language systems, such that patterns of interaction with the simulator are predictive of interaction patterns with the actual spoken language system.

### Comparison of Language-Based Communication Modalities

In a series of studies of interactive human-human communication, Chapanis and colleagues (Chapanis et al., 1972, 1977; Kelly and Chapanis, 1977; Michaelis et al., 1977; Ochsmann and Chapanis, 1974) compared the efficiency of human-human communication when subjects used any of 10 communication modalities (including face-to-face, voice-only, linked teletypes, interactive handwriting). The most important determinant of a team's problem-solving speed was found to be the presence of a voice component. Specifically, a variety of tasks were solved two to three times faster using a voice modality than a hardcopy one, as illustrated in [Figure 1](#). At the same time, speech led to an 8-fold increase in the number of messages and sentences and a 10-fold increase in the rate of communicating words. These results indicate the substantial *potential* for efficiency advantages that may result from use of spoken language communication.

Research by the authors confirmed these efficiency results in human-human dialogues to perform equipment assembly tasks (Cohen, 1984; Oviatt and Cohen, 1991b), finding a 3-fold speed advantage for

interactive telephone speech over keyboard communication. Furthermore, the structure of telephone dialogues differed from that of keyboard dialogues. Among the differences, spoken dialogues exhibited more cue phrases that signaled the structure of the dialogue (such as "next," "ok now"), and speakers interacted in a more "fine-grained" fashion than did keyboard users. Specifically, in order to achieve a subtask, speakers often made two requests, one for object identification and one for action, whereas keyboard users typically integrated both into one imperative utterance. Similar findings of a fine-grained approach during spoken interaction versus a more syntactically integrated approach for keyboard interaction have been found in a study.

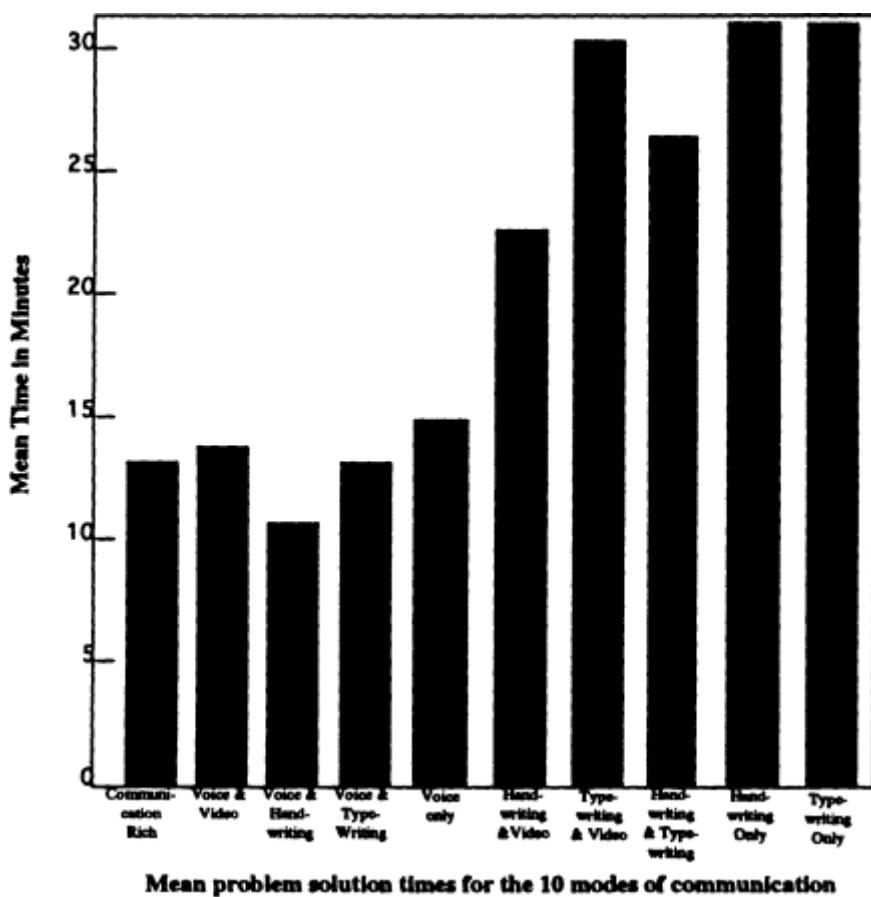


FIGURE 1 Voice determines task efficiency (from Ochsman and Chapanis, 1974).

of simulated human-computer interaction (Zoltan-Ford, 1991). Finally, spoken input was more "indirect" than keyboard input. That is, unlike keyboard interaction, spoken utterances did not literally convey the speaker's intention that the listener perform an action (Cohen, 1984). Future research needs to address the extent to which such results generalize to spoken human-computer interaction for comparable tasks.

One benefit of voice input is the elimination of typing, which could offer potential office productivity savings (Baker, 1991; Jelinek, 1985). In a study of a simulated "listening typewriter," Gould et al. (Gould, 1978, 1982; Gould et al., 1983) examined how novice and expert users of dictation would use a machine that could recognize and type the user's dictation of a business letter, as compared with dictating and editing the letter to a human or handwriting and editing the letter. The listening typewriter system was simulated, and the subjects were informed that they were in fact speaking to a person. It was claimed that users of a listening typewriter were as satisfied with that mode of communication as with the others and that dictating to a listening typewriter could potentially be as fast a mode of letter composition as typing. There is, however, countervailing evidence from a number of simulation studies (Murray et al., 1991; Newell et al., 1990) that speech-only word processors are less efficient and less preferred than composition methods based on writing or typing. Moreover, a combined method of using speech for text input and a touch screen for cursor control was more efficient than speech alone, though still less efficient than composition and editing using keyboards or handwriting.

Neither series of studies examined in detail the linguistic and discourse structure of the dictated material that might explain why spoken composition and editing are less efficient than other modalities. In a study of human-human communication it was found that inexperienced "dictators" providing instructions for a human listener produced more discourse structures that would require editing in order to make acceptable text, such as repetitions, elaborations, and unusual uses of referring expressions, than did users of interactive speech or interactive keyboard (Oviatt and Cohen, 1991a, 1991b). Thus, lack of interaction with a listener may contribute to poorly formulated input, placing a larger burden on the postediting phase where speech input is less efficient (Newell et al., 1990). In summary, though automatic dictation devices have been much touted as an important product concept for speech technology, their potential benefit remains a question.

The space of modality studies has not yet been systematically explored. We do not know precisely how results from human-human communication studies can predict results for studies of human-simulation or human-computer interactions. Also, more studies comparing the structure and content of spoken human-computer language with typed human-computer language need to be conducted in order to understand how to adapt technology developed for keyboard interaction to spoken language systems.

Common to many successful applications of voice-based technology is the lack of an adequate alternative to voice, given the task and environment of computer use. Major questions remain as to the applications where voice will be favored when other modalities of communication are possible. Some studies report a decided preference for speech when compared to other modalities (Rudnicky, 1993), yet other studies report an opposite conclusion (Murray et al., 1991; Newell et al., 1990). Thus, despite the aforementioned potential benefits of human-computer interaction using voice, it is not obvious why people should want to speak to their computers in performing their daily office work. To provide a framework for answering this question, the discussion below compares the currently dominant direct manipulation user interface with typed or spoken natural language.

### **Comparison of Natural Language Interaction with Alternative Modalities**

Numerous alternative modalities of human-computer interaction exist, such as the use of keyboards for transmitting text, pointing and gesturing with devices such as the mouse, a digitizing pen, trackballs, touchscreens, and digitizing gloves. It is important to understand what role speech, and specifically spoken language, can play in supporting human interaction, especially when these other modalities are available. To begin this discussion, we need to identify properties of successful interfaces. Ideally, such an interface should be:

*Error free.* The interface should prevent the user from formulating erroneous commands, should minimize misinterpretations of the user's intent, and should offer simple methods for error correction.

*Transparent.* The functionality of the application system should be obvious to the user.

*High-level.* The user should not have to learn the underlying computer structures and languages but rather should be able to state simply his or her desires and have the system handle the details.

*Consistent.* Strategies that work for invoking one computer function should transfer to the invocation of others.

*Easy to learn.* The user should not need formal training, but rather a brief process of exploration should suffice for learning how to use a given system.

*Expressive.* The user should be able to perform easily any combination of tasks in mind, within the bounds of the system's intended functionality.

Using this set of properties, we discuss the use of direct manipulation and natural language technologies.

## Direct Manipulation

The graphical user-interface paradigm involves a style of interaction that offers the user menus, icons, and pointing devices (e.g., the "mouse" [English et al., 1967]) to invoke computer commands, as well as multiple windows in which to display the output. These graphical user interfaces (GUIs), popularized by the Apple Macintosh and by Microsoft Windows, use techniques pioneered at SRI International and at Xerox's Palo Alto Research Center in the late 1960s and 1970s (Englebart, 1973; Kay and Goldberg, 1977). With GUIs, users perform actions by selecting objects and then choosing the desired action from a menu, rather than by typing commands.

In addition, with many GUIs a user can directly manipulate graphical objects in order to perform actions on the objects they represent. For example, a user can copy a file from one disk to another by selecting its icon with the pointing device and "dragging" it from the list of files on the first disk to the second. Other direct manipulation actions include using a "scroll bar" to view different sections of a file and dragging a file's icon on top of the "trash" icon to delete it. Apart from the mouse, numerous pointing devices exist, such as trackballs and joysticks, and some devices offer multiple capabilities, such as the use of pens for pointing, gesturing, and handwriting. Finally, to generalize along a different dimension, users now can directly manipulate virtual worlds using computer-instrumented gloves and bodysuits (Fisher, 1990; Kreuger, 1977; Rheingold, 1991), allowing for subtle effects of body motion to affect the virtual environment.

*Strengths* Many writers have identified virtues of well-designed graphically based direct manipulation interfaces (DMIs) (e.g., Hutchins et al., 1986; Shneiderman, 1983), claiming that

- Direct manipulation interfaces based on familiar metaphors are intuitive and easy to use.
- Graphical user interfaces can have a consistent "look and feel" that enables users of one program to learn another program quickly.
- Menus make the available options clear, thereby curtailing user errors in formulating commands and specifying their arguments.
- GUIs can shield the user from having to learn underlying computer concepts and details.

It is no exaggeration to say that graphical user interfaces supporting direct manipulation interaction have been so successful that no serious computer company would attempt to sell a machine without one.

*Weaknesses* Direct manipulation interfaces do not suffice for all needs. One clear expressive weakness is the paucity of means available for identifying entities. Merely allowing users to select currently displayed entities provides them with little support for identifying objects not on the screen (such as a file name in a list of 200 files), for specifying temporal relations that denote future or past events, for identifying and operating on large sets of entities, and for using the context of interaction. At most, developers of GUIs have provided simple string-matching routines that find objects based on exact or partial matches of their names. What is missing is a way for users to *describe* entities using some form of linguistic expression in order to denote or pick out an individual object, a set of objects, a time period, and so forth.<sup>6</sup> At a minimum, a description language should include some way to find entities that have a given set of properties, to say which properties are of interest as well as which are not, to say how many entities are desired, to supply temporal constraints on actions involving those properties, and so forth. Moreover, a useful feature of a description language is the ability to reuse the referents of previous descriptions. Some of these capabilities are found in formal query languages, and all are found in natural languages.

Although shielding a user from implementation details, direct manipulation interfaces are often not high level. For example, one common way to request information from a relational database is to select certain fields from tables that one wants to see. To do this correctly, the user needs to learn the structure of the database—for

---

<sup>6</sup> Of course, the elimination of descriptions was a conscious design decision by the originators of GUIs.

example, that the data are represented in one or more tables, comprised of numerous fields, whose meanings may not be obvious. Thus, the underlying tabular implementation has become the user interface metaphor. An alternative is to develop systems and interfaces that translate between the user's way of thinking about the problem and the implementation. In so doing, the user might perhaps implicitly retrieve information but need not know that it is kept in a database, much less learn the structure of that database. By engaging in such a high level interaction, users may be able to combine information access with other information processing applications, such as running a simulation, without first having to think about database retrieval, and then switching "applications" mentally to think about simulation.

When numerous commands are possible, GUIs usually present a hierarchical menu structure. As the number of commands grows, the casual user may have difficulty remembering in which menu they are located. However, the user who knows where the desired action is located in a large action hierarchy still needs to navigate the hierarchy. Software designers have attempted to overcome this problem by providing different menu sets for users of different levels of expertise, by preselecting the most recently used item in a menu, and by providing direct links to commonly used commands through special key combinations. However, in doing the latter, GUIs are borrowing from keyboard-based interfaces and command languages.

Because direct manipulation emphasizes rapid graphical response to actions (Shneiderman, 1983), the time of system action in DMIs is literally the time at which the action was invoked. Although some systems can delay actions until specific future times, DMIs and GUIs offer little support for users who want to execute actions at an unknown but describable future time.

Finally, DMIs rely heavily on a user's hands and eyes. Given our earlier discussion, certain tasks would be better performed with speech. So far, however, there is little research comparing graphical user interfaces with speech. Early laboratory results of a direct manipulation VLSI design system augmented with speaker-dependent speech recognition indicate that users were as fast at speaking single-word commands as they were at invoking the same commands with mouse-button clicks or by typing a single letter command abbreviation (Martin, 1989). That is, no loss of efficiency occurred due to use of speech for simple tasks at which DMIs typically excel. Note that a 2- to 3-fold advantage in speed is generally found when speaking is compared to typing full words (Chapanis et al., 1977, Oviatt and Cohen, 1991b). In a recent study of human-computer interaction to retrieve information

from a small database (240 entries), it was found that speech was substantially preferred over direct manipulation use of scrolling, even though the overall time to complete the task with voice was longer (Rudnicky, 1993). This study suggests that, for simple risk-free tasks, user preference may be based on time to input rather than overall task completion times or overall task accuracy.

## Natural Language Interaction

*Strengths* Natural language is the paradigmatic case of an expressive mode of communication. A major strength is the use of psychologically salient and mnemonic descriptions. English, or any other natural language, provides a set of finely honed descriptive tools such as the use of noun phrases for identifying objects, verb phrases for identifying events, and verb tense and aspect for describing time periods. By the very nature of sentences, these capabilities are deployed simultaneously, as sentences must be about something, and most often describe events situated in time.

Coupled with this ability to describe entities, natural languages offer the ability to avoid extensive redescription through the use of pronouns and other "anaphoric" expressions. Such expressions are usually intended to denote the same entities as earlier ones, and the recipient is intended to infer the connection. Thus, the use of anaphora provides an economical benefit to the speaker, at the expense of the listener's having to draw inferences.

Furthermore, natural language commands can offer a direct route to invoking an action or making selections that would be deeply embedded in the hierarchical menu of actions or would require multiple menu selections, such as font and type style and size in a word processing program. In using such commands, a user could avoid having to select numerous menu entries to isolate the desired action. Moreover, because the invocation of an action may involve a description of its arguments, information retrieval is intimately woven into the invocation of actions.

Ideally, natural language systems should require only a minimum of training on the domain covered by the target system. Using natural language, people should be able to interact immediately with a system of known content and functionality, without having to learn its underlying computer structures. The system should have sufficient vocabulary, as well as linguistic, semantic, and dialogue capabilities, to support interactive problem solving by casual users—that is, users who employ the system infrequently. For example, at its present state of development, many users can successfully solve trip

planning problems with one of the ATIS systems (Advanced Research Projects Agency, 1993), within a few minutes of introduction to the system and its coverage. To develop systems with this level of robustness, the system must be trained and tested on a substantial amount of data representing input from a broad spectrum of users.<sup>7</sup> Currently, the level of training required to achieve a given level of proficiency in using these systems is unknown.

*Weaknesses* In general, various disadvantages are apparent when natural language is incorporated into an interface. Pure natural language systems suffer from opaque linguistic and conceptual coverage—the user knows the system cannot interpret every utterance but does not know precisely what it *can* interpret (Hendrix and Walter, 1987; Murray et al., 1991; Small and Weldon, 1983; Turner et al., 1984). Often, multiple attempts must be made to pose a query or command that the system can interpret correctly. Thus, such systems can be error prone and, as some claim (Shneiderman, 1980), lead to frustration and disillusionment. One way to overcome these problems was suggested in a menu-based language processing system in which users composed queries in a quasi-natural language by selecting phrases from a menu (Tennant et al., 1983). Although the resulting queries are guaranteed to be analyzable, when there is a large number of menu choices to make, the query process becomes cumbersome.

Many natural language sentences are ambiguous, and parsers often find more ambiguities than people do. Hence, a natural language system often engages in some form of clarification or confirmation subdialogue to determine if its interpretation is the intended one. Current research is attempting to handle the ambiguity of natural language input by developing probabilistic parsing algorithms for which analyses would be ranked by their probability of occurrence in the given domain (see Marcus, this volume). Also, research is beginning to investigate the potential for using prosody to choose among ambiguous parses (Bear and Price, 1990; Price et al., 1991). A third research direction involves minimizing ambiguities through multimodal interface techniques to channel the user's language (Cohen, 1991b; Cohen et al., 1989; Oviatt et al., 1993).

---

<sup>7</sup> The ATIS effort has required the collection and annotation of over 10,000 user utterances, some of which is used for system development and the rest for testing during comparative evaluations conducted by the National Institute of Standards and Technology.

Another disadvantage of natural language interaction is that reference resolution algorithms do not always supply the correct answer, in part because systems have underdeveloped knowledge bases and in part because the system has little access to the discourse situation, even if the system's prior utterances and graphical presentations have created that discourse situation. To complicate matters, systems currently have difficulty following the context shifts inherent in dialogue. These contextual and world knowledge limitations undermine the search for referents and provide another reason that natural language systems are usually designed to confirm their interpretations.

It is not clear where typed natural language interaction will be a modality of choice. Studies comparing typed natural language database question answering with database querying using an artificial query language (e.g., SQL) (Chamberlin and Boyce, 1974) have given equivocal results, with some studies concluding that natural language interaction offers faster and more compact query formulation (Jarke et al., 1985), while others conclude that database querying using SQL is more accurate and easier to learn (Jarke et al., 1985; Shneiderman, 1980a). However, these studies are flawed by the use of prototype natural language systems rather than product quality systems. When a product quality natural language database retrieval system (INTELLECT; Harris, 1977) was studied in the field, users reported efficiency gains and a clear preference for natural language interaction as compared with a previous query language method of database interaction (Capindale and Crawford, 1990). Another difficulty in many laboratory studies is the lack of adequate controls on subject training. In one study comparing the utility of natural versus query language usage for database access (Shneiderman, 1980b), users in the natural language condition were given virtually no training on the content of a database, with the rationale that natural language systems should require no training, while users of SQL were trained on the file and field names of that database. Not surprisingly, under these conditions, natural language users made more "overshoot" errors, in the sense of asking for information not contained in the database.

### **Summary: Circumstances Favoring Spoken Language Interaction with Machines**

Theoretically, direct manipulation should be beneficial when the objects to be manipulated are on the screen, their identity is known, and there are not too many objects from which to select. In addition, graphical user interfaces limit users' options, preventing them from

making errors in formulating commands. Natural language interaction with computers offers potential benefits when users need to identify objects, actions, and events from sets too large to be displayed and/or examined individually and when users need to invoke actions at future times that must be described. Furthermore, natural language allows users to think about their problems and express their goals in their own terms rather than those of the computer. However, in allowing users to do so, systems need to have sufficient reasoning and interpretive capabilities to solve the problems of translating between the user's conceptual model and the system's implementation.

Combining the empirical results on circumstances favoring voice-based interaction with the foregoing analysis of interactions for which natural language may be most appropriate, it appears that applications requiring speedy user input of complex descriptions will favor spoken natural language communication. Moreover, this preference is likely to be stronger when a minimum of training about the underlying computer structures is possible. Examples of such an application area are asking questions of a database or creating rules for action (e.g., "If I am late for a meeting, notify the meeting participants"). Because of the recency of usable spoken language systems, there are very few studies comparing spoken language interaction with direct manipulation for accomplishing real tasks.

So far, we have contrasted spoken interaction with other modalities. It is worth noting that these modalities have complementary advantages and disadvantages, which can be leveraged to develop multimodal interfaces that compensate for the weaknesses of one interface technology via the strengths of another (Cohen, 1991; Cohen et al., 1989). (See section titled "[Multimodal Systems](#)."

## HUMAN FACTORS OBSTACLES TO SPOKEN LANGUAGE SYSTEMS

Although there are numerous technical challenges to building spoken language systems, many of which are detailed in this volume, interface and human factors knowledge is especially needed about such systems. We consider below information needed about spontaneous speech, spoken natural language, and spoken interaction.

### Spontaneous Speech

When an utterance is *spontaneously spoken*, it may well involve false starts, hesitations, filled pauses, repairs, fragments, and other types of technically "ungrammatical" utterances. These phenomena

disrupt both speech recognizers and natural language parsers and must be detected and corrected before techniques based on present technology can be deployed robustly. Current research has begun to investigate techniques for detecting and handling disfluencies in spoken human-computer interaction (Bear et al., 1992; Hindle, 1983; Nakatani and Hirschberg, 1993), and robust processing techniques have been developed that enable language analysis routines to recover the meaning of an utterance despite recognition errors (Dowding et al., 1993; Huang et al., 1993; Jackson et al., 1991; Stallard and Bobrow, 1992).

Assessment of different types of human-human and human-computer spoken language has revealed that people's rate of spontaneous disfluencies and self-repairs is substantially lower when they speak to a system, rather than another person (Oviatt, 1993). A strong predictive relationship also has been demonstrated between the rate of spoken disfluencies and an utterance's length (Oviatt, 1993). Rather than having to resolve disfluencies, interface research has revealed that form-based techniques can reduce up to 70 percent of all disfluencies that occur during human-computer interaction (Oviatt, 1993). In short, research suggests that some difficult types of input, such as disfluencies, may be avoided altogether through strategic interface design.

### Natural Language

In general, because the human-machine communication in spoken language involves the system understanding a natural language but not the entire language, users will employ constructs outside the system's coverage. However, it is hoped that given sufficient data on which to base the development of grammars and templates, the likelihood will be small that a cooperative user will generate utterances outside the coverage of the system. Still, it is not currently known:

- how to select relatively "closed" domains, whose vocabulary and linguistic constructs can be acquired through iterative training and testing on a large corpus of user input,
- how well users can discern the system's communicative capabilities,
- how well users can stay within the bounds of those capabilities,
- what level of task performance users can attain
- what level of misinterpretation users will tolerate, and what level is needed for them to solve problems effectively, and
- how much training is acceptable.

Systems are not adept at handling linguistic coverage problems,

other than responding that given words are not in the vocabulary or that the utterance was not understood. Even recognizing that an out-of-vocabulary word has occurred is itself a difficult issue (Cole et al., 1992). If users can discern the system's vocabulary, we can be optimistic that they can adapt to that vocabulary. In fact, human-human communication research has shown that users communicating by typing can solve problems as effectively with a constrained task-specific vocabulary (500 to 1000 words) as with an unlimited vocabulary (Kelly and Chapanis, 1977; Michaelis et al., 1977). User adaption to vocabulary restrictions has also been found for simulated human-computer interaction (Zoltan-Ford, 1983, 1991), although these results need to be verified for spoken human-computer interaction.

For interactive applications, the user may begin to *imitate* or *model* the language observed from the system, and the opportunity is present for the system to play an active role in *shaping* or *channeling* the user's language to match that coverage more closely. Numerous studies of human communication have shown that people will adopt the speech styles of their interlocutors, including vocal intensity (Welkowitz et al., 1972), dialect (Giles et al., 1987), and tempo (Street et al., 1983). Explanations for this convergence of dialogue styles include social factors such as the desire for approval (Giles et al., 1987), and psycholinguistic factors associated with memory limitations (Levett and Kelter, 1982). Similar results have been found in a study of typed and spoken communication to a simulated natural language system (Zoltan-Ford, 1983, 1984), which showed that people will model the vocabulary and length of the system's responses. For example, if the system's responses are terse, the user's input is more likely to be terse as well. In a simulation study of typed natural language database interactions, subjects modeled simple syntactic structures and lexical items that they observed in the system's paraphrases of their input (Leiser, 1989). However, it is not known if the modeling of syntactic structures occurs in spoken human-computer interaction. If users of *spoken* language systems do learn to adopt the grammatical structures they observe, then new forms of user training may be possible by having system designers adhere to the principle that any messages supplied to a user must be analyzable by the system's parser. One way to guarantee such system behavior would be to require the system to generate its utterances, rather than merely reciting canned text, employing a bidirectional grammar. Any utterances the system could generate using that grammar would thus be guaranteed to be parseable.

A number of studies have investigated methods for shaping user's language into the system's coverage. For telecommunications appli

cations, the phrasing of system prompts for information spoken over the telephone dramatically influences the rate of caller compliance for the expected words and phrases (Basson, 1992; Rubin-Spitz and Yashchin, 1989; Spitz, 1991). For systems with screen-based feedback, human spoken language can be effectively channeled through the use of a form that the user fills out with speech (Oviatt et al., 1993). Form-based interactions reduce the syntactic ambiguity of the user's speech by 65 percent, measured as the number of parses per utterance, thereby leading to user language that is simpler to process. At the same time, for the service transactions analyzed in this study, users were found to prefer forms-based spoken and written interaction over unconstrained ones by a factor of 2 to 1. Thus, not only can people's language be channeled, there appear to be cases where they prefer the guidance and sense of completion provided by a form.

### **Interaction and Dialogue**

When given the opportunity to interact with systems via spoken natural language, users will attempt to engage in dialogues, expecting prior utterances and responses to set a context for subsequent utterances, and expecting their conversational partner to make use of that context to determine the referents of pronouns. Although pronouns and other context-dependent constructs sometimes occur less frequently in dialogues with machines than they do in human-human dialogues (Kennedy et al., 1988), context dependence is nevertheless a cornerstone of human-computer interaction. For example, contextually dependent utterances comprise 44 percent of the ATIS corpus collected for the Advanced Research Projects Agency spoken language community (MADCOW Working Group, 1992). In general, a solution to the problem of understanding context dependent utterances will be difficult, as it may require the system to deploy an arbitrary amount of world knowledge (Charniak, 1973; Winograd, 1972). However, it has been estimated that a simple strategy for referent determination employed in text processing, and one that uses only the syntactic structure of previous utterances, can suffice to identify the correct referent for pronouns in over 90 percent of cases (Hobbs, 1978). Whether such techniques will work as well for spoken human-computer dialogue is unknown. One way to mitigate the inherent difficulty of referent determination when using a multimodal system may be to couple spoken pronouns and definite noun phrases with pointing actions (Cohen, 1991; Cohen et al., 1989).

Present spoken language systems have supported dialogues in which the user asks multiple questions, some of which request fur

ther refinement of the answers to prior questions (Advanced Research Projects Agency, 1993), or dialogues in which the user is prompted for information (Andry, 1992; Peckham, 1991). Much more varied dialogue behavior is likely to be required by users, such as the ability to engage in advisory, clarificatory, and confirmatory dialogues (Codd, 1974; Litman and Allen, 1987). With respect to dialogue confirmations, spoken communication is tightly interactive and speakers expect rapid confirmation of understanding through backchannels (e.g., "uh huh") and other signals. Studies have shown that communication delays as brief as 0.25 seconds can disrupt conversation patterns (Krauss and Bricker, 1967), leading speakers to elaborate and rephrase their utterances (Krauss and Weinheimer, 1966; Oviatt and Cohen, 1991a), and that telephone communications are especially sensitive to delays. The need for timely confirmations will challenge most applications of spoken language processing, particularly those involving telephony.

To support a broader range of dialogue behavior, more general models of dialogue are being investigated, both mathematically and computationally, including plan-based models of dialogue and dialogue grammars. Plan-based models are founded on the observation that utterances are not simply strings of words but rather are the observable performance of communicative actions, or speech acts (Searle, 1969), such as requesting, informing, warning, suggesting, and confirming. Moreover, humans do not just perform actions randomly; they plan their actions to achieve various goals, and, in the case of communicative actions, those goals include changes to the mental states of listeners. For example, speakers' requests are planned to alter the intentions of their addressees. Plan-based theories of communicative action and dialogue (Allen and Perrault, 1980; Appelt, 1985; Cohen and Levesque, 1990; Cohen and Perrault, 1979; Perrault and Allen, 1980; Sidner and Israel, 1981) assume that the speaker's speech acts are part of a plan, and the listener's job is to uncover and respond appropriately to the underlying plan, rather than just to the utterance. For example, in response to a customer's question of "Where are the steaks you advertised?", a butcher's reply of "How many do you want?" is appropriate because the butcher has discovered that the customer's plan of getting steaks himself is going to fail. Being cooperative, he attempts to execute a plan to achieve the customer's higher-level goal of having steaks (Cohen, 1978). Current research on this model is attempting to incorporate more complex dialogue phenomena, such as clarifications (Litman and Allen, 1987, 1990; Yamaoka and Iida, 1991), and to model dialogue more as a *joint* enterprise,

something the participants are doing together (Clark and Wilkes-Gibbs, 1986; Cohen and Levesque, 1991; Grosz and Sidner, 1990).

The dialogue grammar approach models dialogue simply as a finite-state transition network (Dahlback and Jonsson, 1992; Polanyi and Scha, 1984; Winograd and Flores, 1986), in which state transitions occur on the basis of the type of communicative action that has taken place (e.g., a request). Such automata might be used to predict the next dialogue "states" that are likely and thus could help speech recognizers by altering the probabilities of various lexical, syntactic, semantic, and pragmatic information (Andry, 1992; Young et al., 1989). However, a number of drawbacks to the model are evident (Cohen, 1993; Levinson, 1981). First, it requires that the communicative action(s) being performed by the speaker in issuing an utterance be identified, which itself is a difficult problem, for which prior solutions have required plan recognition (Allen and Perrault, 1980; Kautz, 1990; Perrault and Allen, 1980). Second, the model assumes that only one state results from a transition. However, utterances are multifunctional. An utterance can be, for example, both a rejection and an assertion. The dialogue grammar subsystem would thus need to be in multiple states simultaneously, a property typically not allowed. Finally, and most importantly, the model does not say how systems should choose among the next moves, that is, the states currently reachable, in order for it to play its role as a cooperative conversant. Some analog of planning is thus also likely to be required.

Dialogue research is currently the weakest link in the research program for developing spoken language systems. First and foremost, dialogue technology is in need of a specification methodology, in which a theorist could state formally what a dialogue system should *do* (i.e., what would count as acceptable dialogue behavior). As in other branches of computer science, such specifications may then lead to methods for mathematically and empirically evaluating whether a given system has met the specifications. However, to do this will require new theoretical approaches. Second, more implementation experiments need to be carried out, ranging from the simpler state-based dialogue models to the more comprehensive plan-based approaches. Research aimed at developing computationally tractable plan recognition algorithms is critically needed.

## MULTIMODAL SYSTEMS

There is little doubt that voice will figure prominently in the array of potential interface technologies available to developers. Except for conventional telephone-based applications, however, human-

computer interfaces incorporating voice will probably be multimodal, in the sense of combining voice with screen feedback use of a pointing device, gesturing, handwriting, etc. (Cohen et al., 1989; Hauptmann and McAvinney, 1993; Oviatt, 1992; Wahlster, 1991). Many application systems require multimodal communication, such as inherently map-based interactions. Such systems can involve coordinated speaking, gesturing, pointing, or writing on the map during input, and speech synthesis coordinated with graphics for output. From the previous discussion, it is apparent that each interface technology has strengths and weaknesses, and it may be strategic to attempt to develop interfaces that capitalize on the strengths of one to overcome weaknesses in another (Cohen, 1991). That is, users should be able to speak when desired, supplemented with other modalities as needed.

There are many advantages to multimodal interfaces:

*Error avoidance and robust performance.* Multimodal interfaces can offer the potential to avoid errors that otherwise would be made in a unimodal interface. For example, it is estimated that 86 percent of the task-critical human performance errors that occurred during a study of an interpreted telephony could have been avoided by opening up a screen-based handwriting channel (Oviatt, in press). Multimodal recognition also offers the possibility of enhanced recognition in adverse conditions. For example, simultaneous use of lip-reading speech recognizers may increase the recognition rate in high-noise environments (Garcia et al., 1992; Petajan et al., 1988) that otherwise would impair acoustic speech recognizers. Alternatively, in such environments, users of multimodal interfaces could simply switch modes, for example, to use handwriting.

*Error correction.* Multimodal interfaces offer more options for correcting errors that do occur. Recognition errors present a problem to users, partly because their source is not apparent. Users frequently respond to speech recognition errors by hyperarticulating. But since recognizers are typically not trained on hyperarticulated speech, this repair strategy leads to a lower likelihood of successful recognition for that content (Shriberg et al., 1992). Recognition problems can thus repeat numerous times on the same content, leading to a "degradation spiral" that is frustrating to users and may cause them to abort the application (Oviatt, 1992). By providing the option of using another modality, such as handwriting, a user can simply switch modes in order to correct an error in the first modality.

*Situational and user variation.* The various circumstances in which portable computers will be used are likely to alter people's preferences for one modality of communication or another. For example,

the user may at times encounter noisy environments or desire privacy and would therefore rather not speak. Also, people may prefer to speak for some task content but not for others. Finally, different types of users may systematically prefer to use one modality rather than another. In all these cases a multimodal system offers the needed flexibility.

Even as we investigate multimodal interaction for potential solutions to problems arising in speech-only applications, many implementation obstacles need to be overcome in order to integrate and synchronize modalities. For example, multimodal systems could present information graphically or in multiple coordinated modalities (Feiner and McKeown, 1991; Wahlster, 1991) and permit users to refer linguistically to entities introduced graphically (Cohen, 1991; Wahlster, 1991). Techniques need to be developed to synchronize input from simultaneous data streams, so that, for example, gestural inputs can help resolve ambiguities in speech processing and vice versa. Research on multimodal interfaces needs to examine not only the techniques for forging a productive synthesis among modalities but also the effect that specific integration architectures will have on human-computer interaction. Much more empirical research on the human use of multimodal systems needs to be undertaken, as we yet know relatively little about how people use multiple modalities in communicating with other people, let alone with computers, or about how to support such communication most effectively.

### **SCIENTIFIC RESEARCH ON COMMUNICATION MODALITIES**

The present research and development climate for speech-based technology is more active than it was at the time of the 1984 National Research Council report on speech recognition in severe environments (National Research Council, 1984). Significant amounts of research and development funding are now being devoted to building speech-understanding systems, and the first speaker-independent, continuous, real-time spoken language systems have been developed. However, some of the same problems identified then still exist today. In particular, few answers are available on how people will interact with systems using voice and how well they will perform tasks in the target environments as opposed to the laboratory. There is little research on the dependence of communication on the modality used, or the types of tasks, in part because there have not been principled taxonomies or comprehensive research addressing these factors. In

particular, the use of multiple communication modalities to support human-computer interaction is only now being addressed.

Fortunately, the field is now in a position to fill gaps in its knowledge base about spoken human-machine communication. Using existing systems that understand real-time, continuously spoken utterances, which allow users to solve real problems, a number of vital studies can now be undertaken in a more systematic manner. Examples include:

- longitudinal studies of users' linguistic and problem-solving behavior that would explore how users adapt to a given system;
- studies of users' understanding of system limitations, and of their performance in observing the system's bounds;
- studies of different techniques for revealing a system's coverage, and for channeling user input;
- studies comparing the effectiveness of spoken language technology with alternatives, such as the use of keyboard-based natural language systems, query languages, or existing direct manipulation interfaces; and
- studies analyzing users' language, task performance, and preferences to use different modalities, individually and within an integrated multimodal interface.

The information gained from such studies would be an invaluable addition to the knowledge base of how spoken language processing can be woven into a usable human-computer interface. Sustained efforts need to be undertaken to develop more adequate spoken language simulation methods, to understand how to build limited but robust dialogue systems based on a variety of communication modalities, and to study the nature of dialogue.

A vital and underappreciated contribution to the successful deployment of voice technology for human-computer interaction will come from the development of a principled and empirically validated set of human-interface guidelines for interfaces that incorporate speech (cf. Lea, 1992). Graphical user-interface guidelines typically provide heuristics and suggestions for building "usable" interfaces, though often without basing such suggestions on scientifically established facts and principles. Despite the evident success of such guidelines for graphical user interfaces, it is not at all clear that a simple set of heuristics will work for spoken language technology, because human language is both more variable and creative than the behavior allowed by graphical user interfaces. Answers to some of the questions posed earlier would be valuable in laying a firm empirical founda

tion for developing effective guidelines for a new generation of language-oriented interfaces.

Ultimately, such a set of guidelines embodying the results of scientific theory and experimentation should be able to predict, given a specified communicative situation, task, user population, and a set of component modalities, what the user-computer interaction will be like with a multimodal interface of a certain configuration. Such predictions could inform the developers in advance about potential trouble spots and could lead to a more robust, usable, and satisfying human computer interface. Given the complexities of the design task and the considerable expense required to create spoken language applications, if designers are left to their intuitions, applications will suffer. Thus, for scientific, technological, and economic reasons, a concerted effort needs to be undertaken to develop a more scientific understanding of communication modalities and how they can best be integrated in support of successful human-computer interaction.

## ACKNOWLEDGMENTS

Many thanks to Jared Bernstein, Clay Coler, Carol Simpson, Ray Perrault, Robert Markinson, Raja Rajasekharan, and John Vester for valuable discussions and source materials.

## REFERENCES

- Advanced Research Projects Agency. ARPA Spoken Language Systems Technology Workshop. Massachusetts Institute of Technology, Cambridge, Mass. 1993.
- Allen, J. F., and C. R. Perrault. Analyzing intention in dialogues. *Artificial Intelligence*, 15 (3):143-178, 1980.
- Andry, F. Static and dynamic predictions: A method to improve speech understanding in cooperative dialogues. In Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, Oct. University of Alberta, 1992.
- Andry, F., E. Bilange, F. Charpentier, K. Choukri, M. Ponamale, and S. Soudoplatoff. Computerised simulation tools for the design of an oral dialogue system. In Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218). Commission of the European Communities, 1990.
- Appelt, D. Planning English Sentences. Cambridge University Press, Cambridge, U.K., 1985.
- Appelt, D. E., and E. Jackson. SRI International February 1992 ATIS benchmark test results. In Fifth DARPA Workshop on Speech and Natural Language, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1992.
- Bahl, L., F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5 (2):179-190, March 1983.
- Baker, J. F. Stochastic modeling for automatic speech understanding. In D. R. Reddy, ed., *Speech Recognition*, pp. 521-541. Academic Press, New York, 1975.

- Baker, J. M. Large-vocabulary speaker-adaptive continuous speech recognition research overview at Dragon systems. In Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology, pp. 29-32, Genova, Italy, 1991.
- Basson, S. Prompting the user in ASR applications. In Proceedings of COST232 Workshop—European Cooperation in Science and Technology, November 1992.
- Basson, S., O. Christie, S. Levas, and J. Spitz. Evaluating speech recognition potential in automating directory assistance call completion. In AVIOS Proceedings. American Voice I/O Society, 1989.
- Bear, J., J. Dowding, and E. Shriberg. Detection and correction of repairs in human-computer dialog. In D. Walker, ed., Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, Newark, Delaware, June 1992.
- Bear, J., and P. Price. Prosody, syntax and parsing. In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, pp. 17-22, Pittsburgh, Pa., 1990.
- Bernstein, J. Applications of speech recognition technology in rehabilitation. In J. E. Harkins and B. M. Virvan, eds., *Speech to Text: Today and Tomorrow*. GRI Monograph Series B., No. 2. Gallaudet University Research Institute, Washington, D.C., 1988.
- Bernstein, J., M. Cohen, H. Murveit, D. Rtishev, and M. Weintraub. Automatic evaluation and training in English pronunciation. In Proceedings of the 1990 International Conference on Spoken Language Processing, pp. 1185-1188, The Acoustical Society of Japan, Kobe, Japan, 1990.
- Bernstein, J., and D. Rtishev. A voice interactive language instruction system. In Proceedings of Eurospeech '91, pp. 981-984, Genova, Italy. IEEE, 1991.
- Capindale, R. A., and R. C. Crawford. Using a natural language interface with casual users. *International Journal of Man-Machine Studies*, 32:341-362, 1990.
- Chamberlin, D. D., and R. F. Boyce. Sequel: A structured English query language. In Proceedings of the 1974 ACM SIGMOD Workshop on Data Description, Access and Control, May 1974.
- Chapanis, A., R. B. Ochsman, R. N. Parrish, and G. D. Weeks. Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem solving. *Human Factors*, 14:487-509, 1972.
- Chapanis, A., R. N. Parrish, R. B. Ochsman, and G. D. Weeks. Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving. *Human Factors*, 19(2):101-125, April 1977.
- Charniak, E., Jack and Janet in search of a theory of knowledge. In Advance Papers of the Third Meeting of the International Joint Conference on Artificial Intelligence, Los Altos, Calif. William Kaufmann, Inc., 1973.
- Clark, H. H., and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1-39, 1986.
- Codd, E. F. Seven steps to rendezvous with the casual user. In Proceedings IFIP TC-2 Working Conference on Data Base Management Systems, pp. 179-200. North-Holland Publishing Co., Amsterdam, 1974.
- Cohen, P. R. On Knowing What to Say: Planning Speech Acts. PhD thesis, University of Toronto, Toronto, Canada. Technical Report No. 118, Department of Computer Science, 1978.
- Cohen, P. R. The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10(2):97-146, April-June 1984.
- Cohen, P. R. The role of natural language in a multimodal interface. In The 2nd FRIEND21

- International Symposium on Next Generation Human Interface Technologies, Tokyo, Japan, November 1991. Institute for Personalized Information Environment.
- Cohen, P. R. Models of dialogue. In M. Nagao, ed., *Cognitive Processing for Vision and Voice: Proceedings of the Fourth NEC Research Symposium*. SIAM, 1993.
- Cohen, P. R., and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Intentions in Communication*. MIT Press, Cambridge, Mass., 1990.
- Cohen, P. R., and H. J. Levesque. Confirmations and joint action. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pp. 951-957, Sydney, Australia, Morgan Kaufmann Publishers, Inc. 1991.
- Cohen, P. R., and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177-212, 1979.
- Cohen, P. R., M. Dalrymple, D. B. Moran, F. C. N. Pereira, J. W. Sullivan, R. A. Gargan, J. L. Schlossberg, and S. W. Tyler. Synergistic use of direct manipulation and natural language. In *Human Factors in Computing Systems: CHI'89 Conference Proceedings*, pp. 227-234, New York, Addison-Wesley Publishing Co. 1989.
- Cole, R., L. Hirschman, L. Atlas, M. Beckman, A. Bierman, M. Bush, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorain, and V. Zue. NSF Workshop on Spoken Language Understanding. Technical Report CS/E 92-014, Oregon Graduate Institute, September 1992.
- Crane, H. D. Writing and talking to computers. Business Intelligence Program Report D91-1557, SRI International, Menlo Park, Calif., July 1991.
- Dahlback, N., and A. Jonsson. An empirically based computationally tractable dialogue model. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society (COGSCI-92)*, Bloomington, Ind., July 1992.
- Dahlback, N., A. Jonsson, and L. Ahrenberg. Wizard of Oz studies—why and how. In L. Ahrenberg, N. Dahlback, and A. Jonsson, eds., *Proceedings from the Workshop on Empirical Models and Methodology for Natural Language Dialogue Systems*, Trento, Italy, April. Association for Computational Linguistics, 1992.
- Dowding, J., J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran. Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 54-61, Columbus, Ohio, June 1993.
- Englebart, D. Design considerations for knowledge workshop terminals. In *National Computer Conference*, pp. 221-227, 1973.
- English, W. K., D. C. Englebart, and M. A. Berman. Display-selection techniques for text manipulation. *IEEE Transactions on Human Factors in Electronics*, HFE-8(1):515, March 1967.
- Feiner, S. K., and K. R. McKeown. COMET: Generating coordinated multimedia explanations. In *Human Factors in Computing Systems (CHI'91)*, pp. 449-450, New York, April. ACM Press, 1991.
- Fisher, S. Virtual environments, personal simulation, and telepresence. *Multimedia Review: The Journal of Multimedia Computing*, 1(2), 1990.
- Fraser, N. M., and G. N. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5 (1):81-99, 1991.
- Garcia, O. N., A. J. Goldschen, and E. D. Petajan. Feature Extraction for Optical Speech Recognition or Automatic Lipreading. Technical Report, Institute for Information Science and Technology, Department of Electrical Engineering and Computer Science. The George Washington University, Washington, D.C., November 1992.

- Giles, H., A. Mulac, J. J. Bradac, and P. Johnson. Speech accommodation theory: The first decade and beyond. In M. L. McLaughlin, ed., *Communication Yearbook 10*, pp. 13-48. Sage Publishers, Beverly Hills, California, 1987.
- Gould, J. D. How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4):648-661, 1978.
- Gould, J. D. Writing and speaking letters and messages. *International Journal of Man-Machine Studies*, 16(1):147-171, 1982.
- Gould, J. D., J. Conti, and T. Hovanyecz. Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26(4):295-308, April 1983.
- Grosz, B., and C. Sidner. Plans for discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Intentions in Communication*, pp. 417-444. MIT Press, Cambridge, Mass., 1990.
- Guyomard, M., and J. Siroux. Experimentation in the specification of an oral dialogue. In H. Niemann, M. Lang, and G. Sagerer, eds., *Recent Advances in Speech Understanding and Dialogue Systems*. NATO ASI Series, vol. 46. Springer Verlag, Berlin, 1988.
- Harris, R. User oriented data base query with the robot natural language query system. *International Journal of Man-Machine Studies*, 9:697-713, 1977.
- Hauptmann, A. G., and P. McAvinney. Gestures with speech for direct manipulation. *International Journal of Man-Machine Studies*, 38:231-249, 1993.
- Hauptmann, A. G., and A. I. Rudnick. A comparison of speech and typed input. In *Proceedings of the Speech and Natural Language Workshop*, pp. 219-224, San Mateo, Calif., June. Morgan Kaufmann, Publishers, Inc., 1990.
- Hendrix, G. G., and B. A. Walter. The intelligent assistant. *Byte*, pp. 251-258, December 1987.
- Hindle, D. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pp. 123-128, Cambridge, Mass., June 1983.
- Hobbs, J. R., Resolving pronoun reference. *Lingua*, 44, 1978.
- Hon, H.-W., and K.-F. Lee. Recent progress in robust vocabulary-independent speech recognition. In *Proceedings of the Speech and Natural Language Workshop*, pp. 258-263, San Mateo, Calif., October. Morgan Kaufmann, Publishers, Inc., 1991.
- Howard, J. A., Flight testing of the AFTI/F-16 voice interactive avionics system. In *Proceedings of Military Speech Tech 1987*, pp. 76-82, Arlington, Va., Media Dimensions., 1987.
- Huang, X., F. Alleva, M.-Y. Hwang, and R. Rosenfeld. An overview of the SPHINX-II speech recognition system. In *Proceedings of the ARPA Workshop on Human Language Technology*, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1993.
- Hutchins, E. L., J. D. Hollan, and D. A. Norman. Direct manipulation interfaces. In D. A. Norman and S. W. Draper, eds., *User Centered System Design*, pp. 87-124. Lawrence Erlbaum Publishers, Hillsdale, N.J., 1986.
- Jackson, E., D. Appelt, J. Bear, R. Moore, and A. Podlozny. A template matcher for robust NL interpretation. In *Proceedings of the 4th DARPA Workshop on Speech and Natural Language*, pp. 190-194, San Mateo, Calif., February. Morgan Kaufmann Publishers, Inc., 1991.
- Jarke, M., J. A. Turner, E. A. Stohr, Y. Vassiliou, N. H. White, and K. Michielsen. A field evaluation of natural language for data retrieval. *IEEE Transactions on Software Engineering*, SE-11 (1):97-113, 1985.
- Jelinek, F. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532-536, April 1976.

- Jelinek, F. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, 73(11):1616-1624, November 1985.
- Karis, D., and K. M. Dobroth. Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal of Selected Areas in Communications*, 9(4):574-585, 1991.
- Kautz, H. A circumscription theory of plan recognition. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Intentions in Communication*. MIT Press, Cambridge, Mass., 1990.
- Kay, A., and A. Goldberg. Personal dynamic media. *IEEE Computer*, 10(1):31-42, 1977.
- Kelly, M. J., and A. Chapanis. Limited vocabulary natural language dialogue. *International Journal of Man-Machine Studies*, 9:479-501, 1977.
- Kennedy, A., A. Wilkes, L. Elder, and W. S. Murray. Dialogue with machines. *Cognition*, 30 (1):37-72, 1988.
- Kitano, H. o dm-dialog. *IEEE Computer*, 24(6):36-50, June 1991.
- Krauss, R. M., and P. D. Bricker. Effects of transmission delay and access delay on the efficiency of verbal communication. *Journal of the Acoustical Society of America*, 41(2):286-292, 1967.
- Krauss, R. M., and S. Weinheimer. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343-346, 1966.
- Kreuger, M. Responsive environments. In *Proceedings of the National Computer Conference*, 1977.
- Kubala, F., C. Barry, M. Bates, R. Bobrow, P. Fung, R. Ingria, J. Makhoul, L. Nguyen, R. Schwartz, and D. Stallard. BBN BYBLOS and HARC February 1992 ATIS benchmark results. In *Fifth DARPA Workshop on Speech and Natural Language*, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1992.
- Kurematsu, A. Future perspective of automatic telephone interpretation. *Transactions of IEICE*, E75 (1):14-19, January 1992.
- Lea, W. A. Practical lessons from configuring voice I/O systems. In *Proceedings of Speech Tech/Voice Systems Worldwide*, New York. Media Dimensions, Inc., 1992.
- Leiser, R. G. Exploiting convergence to improve natural language understanding. *Interacting with Computers*, 1(3):284-298, December 1989.
- Lennig, M. Using speech recognition in the telephone network to automate collect and third-number-billed calls. In *Proceedings of Speech Tech'89*, pp. 124-125, Arlington, Va. Media Dimensions, Inc., 1989.
- Levelt, W. J. M., and S. Kelter. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78-106, 1982.
- Levinson, S. Some pre-observations on the modeling of dialogue. *Discourse Processes*, 4(1), 1981.
- Litman, D. J., and J. F. Allen. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163-200, 1987.
- Litman, D. J., and J. F. Allen. Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Intentions in Communication*, pp. 365-388. MIT Press, Cambridge, Mass., 1990.
- Luce, P. A., T. C. Feustel, and D. B. Pisoni. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25(1):17-32, 1983.
- MADCOW Working Group. Multi-site data collection for a spoken language corpus. In *Proceedings of the Speech and Natural Language Workshop*, pp. 7-14, San Mateo, Calif., February. Morgan Kaufmann Publishers, Inc., 1992.
- Mariani, J. Spoken language processing in the framework of human-machine commu

- nication at LIMSI. In Proceedings of Speech and Natural Language Workshop, pp. 55-60, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1992.
- Marshall, J. P. A manufacturing application of voice recognition for assembly of aircraft wire harnesses. In Proceedings of Speech Tech/Voice Systems Worldwide, New York. Media Dimensions, Inc., 1992.
- Martin, G. L. The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies*, 30(4):355-375, 1989.
- Martin, T. B. Practical applications of voice input to machines. *Proceedings of the IEEE*, 64(4):487-501, April 1976.
- Michaelis, P. R., A. Chapanis, G. D. Weeks, and M. J. Kelly. Word usage in interactive dialogue with restricted and unrestricted vocabularies. *IEEE Transactions on Professional Communication*, PC-20(4), December 1977.
- Mostow, J., A. G. Hauptmann, L. L. Chase, and S. Roth. Towards a reading coach that listens: Automated detection of oral reading errors. In Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93), Menlo Park, Calif., AI Press/The MIT Press, 1993.
- Murray, I. R., J. L. Arnott, A. F. Newell, G. Cruickshank, K. E. P. Carter, and R. Dye. Experiments with a Full-Speed Speech-Driven Word Processor. Technical Report CS 91/09, Mathematics and Computer Science Department, University of Dundee, Dundee, Scotland, April 1991.
- Nakatani, C., and J. Hirschberg. A speech-first model for repair detection and correction. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 46-53, Columbus, Ohio, June 1993.
- National Research Council. Automatic Speech Recognition in Severe Environments. National Academy Press, Washington, D.C., 1984.
- Newell, A. F., J. L. Arnott, K. Carter, and G. Cruickshank. Listening typewriter simulation studies. *International Journal of Man-Machine Studies*, 33(1):1-19, 1990.
- Nusbaum, H. C., and E. C. Schwab. The effects of training on intelligibility of synthetic speech: II. The learning curve for synthetic speech. In Proceedings of the 105th meeting of the Acoustical Society of America, Cincinnati, Ohio, May 1983.
- Nye, J. M. Human factors analysis of speech recognition systems. In *Speech Technology* 1, pp. 50-57, 1982.
- Ochsman, R. B., and A. Chapanis. The effects of 10 communication modes on the behaviour of teams during co-operative problem-solving. *International Journal of Man-Machine Studies*, 6(5):579-620, Sept. 1974.
- Oviatt, S. L. Pen/voice: Complementary multimodal communication. In *Proceedings of Speech Tech'92*, pp. 238-241, New York, February 1992.
- Oviatt, S. L. Predicting spoken disfluencies during human-computer interaction. In K. Shirai, ed., *Proceedings of the International Symposium on Spoken Dialogue: New Directions in Human-Machine Communication*, Tokyo, Japan, November 1993.
- Oviatt, S. L. Toward multimodal support for interpreted telephone dialogues. In M. M. Taylor, F. Neel, and D. G. Bouwhuis, eds., *Structure of Multimodal Dialogue*. Elsevier Science Publishers B.V., Amsterdam, Netherlands, in press.
- Oviatt, S. L., and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 5(4):297-326, 1991a.
- Oviatt, S. L., and P. R. Cohen. The contributing influence of speech and interaction on human discourse patterns. In J. W. Sullivan and S. W. Tyler, eds., *Intelligent User Interfaces*, pp. 69-83. ACM Press Frontier Series. Addison-Wesley Publishing Co., New York, 1991b.
- Oviatt, S. L., P. R. Cohen, M. W. Fong, and M. P. Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In J.

- Ohala, ed., Proceedings of the 1992 International Conference on Spoken Language Processing , pp. 1351-1354, University of Alberta, October 1992.
- Oviatt, S. L., P. R. Cohen, M. Wang, and J. Gaston. A simulation-based research strategy for designing complex NL systems. In ARPA Human Language Technology Workshop, Princeton, N.J., March 1993.
- Pallett, D. S., J. G. Fiscus, W. M. Fisher, and J. S. Garofolo. Benchmark tests for the DARPA spoken language program. In Proceedings of the ARPA Workshop on Human Language Technology, San Mateo, Calif., Morgan Kaufmann Publishers, Inc., 1993.
- Pavan, S., and B. Pelletti. An experimental approach to the design of an oral cooperative dialogue. In Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218). Commission of the European Communities, 1990.
- Peckham, J. Speech understanding and dialogue over the telephone: An overview of the ESPRIT SUNDIAL project. In Proceedings of the Speech and Natural Language Workshop, pp. 14-28, San Mateo, Calif., February. Morgan Kaufmann Publishers, Inc., 1991.
- Perrault, C.R., and J. F. Allen. A plan-based analysis of indirect speech acts. American Journal of Computational Linguistics, 6(3):167-182, 1980.
- Petajan, E., B. Bradford, D. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. In Proceedings of Human Factors in Computing Systems (CHI'88), pp. 19-25, New York. Association for Computing Machinery Press, 1988.
- Polanyi, R., and R. Scha. A syntactic approach to discourse semantics. In Proceedings of the 10th International Conference on Computational Linguistics, pp. 413-419, Stanford, Calif., 1984.
- Pollack, A. Computer translator phones try to compensate for Babel. New York Times, January 29, 1993.
- Price, P. J., Evaluation of spoken language systems: The ATIS domain. In Proceedings of the 3rd DARPA Workshop on Speech and Natural Language, pp. 91-95, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1990.
- Price, P., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. In Proceedings of the Speech and Natural Language Workshop, pp. 372-377, San Mateo, Calif., October. Morgan Kaufmann Publishers, Inc., 1991.
- Proceedings of the Speech and Natural Language Workshop, San Mateo, Calif., October, 1991, Morgan Kaufmann Publishers, Inc.
- Rabiner, L. R., J. G. Wilpon, and A. E. Rosenberg. A voice-controlled, repertory-dialer system. Bell System Technical Journal, 59(7):1153-1163, September 1980.
- Rheingold, H. Virtual Reality. Summit Books, 1991.
- Roe, D. B., F. Pereira, R. W. Sproat, and M. D. Riley. Toward a spoken language translator for restricted-domain context-free languages. In Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology, pp. 1063-1066, Genova, Italy. European Speech Communication Association, 1991.
- Rosenhoover, F. A., J. S. Eckel, F. A. Gorg, and S. W. Rabeler. AFTI/F-16 voice interactive avionics evaluation. In Proceedings of the National Aerospace and Electronics Conference (NAECON'87). IEEE, 1987.
- Rubin-Spitze, J., and D. Yashchin. Effects of dialogue design on customer responses in automated operator services. In Proceedings of Speech Tech'89, 1989.
- Rudnick, A. I. Mode preference in a simple data-retrieval task. In ARPA Human Language Technology Workshop, Princeton, N.J., March 1993.
- Searle, J. R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

- Shneiderman, B. Natural vs. precise concise languages for human operation of computers: Research issues and experimental approaches. In Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics, pp. 139-141, Philadelphia, Pa., June 1980a.
- Shneiderman, B. Software Psychology: Human Factors in Computer and Information systems. Winthrop Publishers, Inc., Cambridge, Mass., 1980b.
- Shneiderman, B. Direct manipulation: A step beyond programming languages. IEEE Computer, 16 (8):57-69, 1983.
- Shriberg, E., E. Wade, and P. Price. Human-machine problem-solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In Proceedings of Speech and Natural Language Workshop, pp. 49-54, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1992.
- Sidner, C., and D. Israel. Recognizing intended meaning and speaker's plans. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence, pp. 203-208, Vancouver, B.C., 1981.
- Simpson, C. A., and T. N. Navarro. Intelligibility of computer generated speech as a function of multiple factors. In Proceedings of the National Aerospace and Electronics Conference (NAECON), pp. 932-940, New York, May. IEEE, 1984.
- Simpson, C. A., C. R. Coler, and E. M. Huff. Human factors of voice I/O for aircraft cockpit controls and displays. In Proceedings of the Workshop on Standardization for Speech I/O Technology, pp. 159-166, Gaithersburg, Md., March. National Bureau of Standards, 1982.
- Simpson, C. A., M. E. McCauley, E. F. Roland, J. C. Ruth, and B. H. Williges. System design for speech recognition and generation. Human Factors, 27(2):115-141, 1985.
- Small, D., and L. Weldon. An experimental comparison of natural and structured query languages. Human Factors, 25:253-263, 1983.
- Spitz, J. Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In Proceedings of the 4th DARPA Workshop on Speech and Natural Language, Asilomar, Calif., February. Defense Advanced Research Projects Agency, 1991.
- Stallard, D., and R. Bobrow. Fragment processing in the DELPHI system. In Proceedings of the Speech and Natural Language Workshop, pp. 305-310, San Mateo, Calif., February. Morgan Kaufmann Publishers, Inc., 1992.
- Street, R. L., Jr., R. M. Brady, and W. B. Putman. The influence of speech rate stereotypes and rate similarity on listeners' evaluations of speakers. Journal of Language and Social Psychology, 2(1):37-56, 1983.
- Streeter, L. A., D. Vitello, and S. A. Wonsiewicz. How to tell people where to go: Comparing navigational aids. International Journal of Man-Machine Studies, 22:549-562, 1985.
- Swider, R. F. Operational evaluation of voice command/response in an Army helicopter. In Proceedings of Military Speech Tech 1987, pp. 143-146, Arlington, Va. Media Dimensions, 1987.
- Tanaka, S., D. K. Wild, P. J. Seligman, W. E. Halperin, V. Behrens, and V. Putz-Anderson. Prevalence and Work-Relatedness of Self-Reported Carpal Tunnel Syndrome Among U.S. Workers—Analysis of the Occupational Health Supplement Data of the 1988 National Health Interview Survey. National Institute of Occupational Safety and Health, and Centers for Disease Control and Prevention (Cincinnati), in submission.
- Tennant, H. R., K. M. Ross, R. M. Saenz, C. W. Thompson, and J. R. Miller. Menu-based natural language understanding. In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, pp. 151-158, Cambridge, Mass., June 1983.

- Thomas, J. C., M. B. Rosson, and M. Chodorow. Human factors and synthetic speech. In B. Shackel, ed., *Proceedings of INTERACT'84*, Amsterdam. Elsevier Science Publishers B.V. (North Holland), 1984.
- Turner, J. A., M. Jarke, E. A. Stohr, Y. Vassiliou, and N. White. Using restricted natural language for data retrieval: A plan for field evaluation. In Y. Vassiliou, ed., *Human Factors and Interactive Computer systems*, pp. 163-190. Ablex Publishing Corp., Norwood, N.J., 1984.
- VanKatwijk, A. F., F. L. VanNes, H. C. Bunt, H. F. Muller, and F. F. Leopold. Naïve subjects interacting with a conversing information system. *IPO Annual Progress Report*, 14:105-112, 1979.
- Visick, D., P. Johnson, and J. Long. The use of simple speech recognisers in industrial applications. In *Proceedings of INTERACT'84 First IFIP Conference on Human-Computer Interaction*, London, U.K., 1984.
- Voorhees, J. W., N. M. Bucher, E. M. Huff, C. A. Simpson, and D. H. Williams. Voice interactive electronic warning system (views). In *Proceedings of the IEEE/AIAA 5th Digital Avionics Systems Conference*, pp. 3.5.1-3.5.8, New York. IEEE, 1983.
- Wahlster, W. User and discourse models for multimodal communication. In J. W. Sullivan and S. W. Tyler, eds., *Intelligent User Interfaces*, pp. 45-68. ACM Press Frontier Series. Addison-Wesley Publishing Co., New York. 1991.
- Weinstein, C. Opportunities for advanced speech processing in military computer-based systems. *Proceedings of the IEEE*, 79(11):1626-1641, November 1991.
- Welkowitz, J., S. Feldstein, M. Finkelstein, and L. Aylesworth. Changes in vocal intensity as a function of interspeaker influence. *Perceptual and Motor Skills*, 10:715718, 1972.
- Williamson, J. T. Flight test results of the AFTI/F-16 voice interactive avionics program. In *Proceedings of the American Voice I/O Society (AVIOS) 87 Voice I/O Systems Applications Conference*, pp. 335-345, Alexandria, Va., 1987.
- Winograd, T. *Understanding Natural Language*. Academic Press, New York, 1972.
- Winograd, T., and F. Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Ablex Publishing Co., Norwood, N.J., 1986.
- Yamaoka, T., and H. Iida. Dialogue interpretation model and its application to next utterance prediction for spoken language processing. In *Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology*, pp. 849-852, Genova, Italy. European Speech Communication Association, 1991.
- Yato, F., T. Takezawa, S. Sagayama, J. Takami, H. Singer, N. Uratani, T. Morimoto, and A. Kurematsu. International Joint Experiment Toward Interpreting Telephony (in Japanese). Technical Report, The Institute of Electronics, Information, and Communication Engineers, 1992.
- Young, S. R., A. G. Hauptmann, W. H. Ward, E. T. Smith, and P. Werner. High level knowledge sources in usable speech recognition systems. *Communications of the ACM*, 32(2), February 1989.
- Zoltan-Ford, E. *Language Shaping and Modeling in Natural Language Interactions with Computers*. PhD thesis, Psychology Department, Johns Hopkins University, Baltimore, Md., 1983.
- Zoltan-Ford, E. Reducing variability in natural-language interactions with computers. In M. J. Alluisi, S. de Groot, and E. A. Alluisi, eds., *Proceedings of the Human Factors Society-28th Annual Meeting*, vol. 2, pp. 768-772, San Antonio, Tex., 1984.
- Zoltan-Ford, E. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527-547, 1991.
- Zue, V., J. Glass, D. Goddeau, D. Goodine, L. Hirschman, M. Phillips, J. Polifroni, and S. Seneff. The MIT ATIS system: February 1992 progress report. In *Fifth DARPA Workshop on Speech and Natural Language*, San Mateo, Calif. Morgan Kaufmann Publishers, Inc., 1992.

# Speech Communication—An Overview

*James L. Flanagan*

## SUMMARY

Advances in digital speech processing are now supporting application and deployment of a variety of speech technologies for human/machine communication. In fact, new businesses are rapidly forming about these technologies. But these capabilities are of little use unless society can afford them. Happily, explosive advances in microelectronics over the past two decades have assured affordable access to this sophistication as well as to the underlying computing technology.

The research challenges in speech processing remain in the traditionally identified areas of *recognition*, *synthesis*, and *coding*. These three areas have typically been addressed individually, often with significant isolation among the efforts. But they are all facets of the same fundamental issue—how to represent and quantify the information in the speech signal. This implies deeper understanding of the physics of speech production, the constraints that the conventions of language impose, and the mechanism for information processing in the auditory system. In ongoing research, therefore, we seek more accurate models of speech generation, better computational formulations of language, and realistic perceptual guides for speech processing—along with ways to coalesce the fundamental issues of recognition, synthesis, and coding. Successful solution will yield the

long-sought dictation machine, high-quality synthesis from text, and the ultimate in low bit-rate transmission of speech. It will also open the door to language-translating telephony, where the synthetic foreign translation can be in the voice of the originating talker.

## INTRODUCTION

Speech is a preferred means for communication among humans. It is beginning to be a preferred means for communication between machines and humans. Increasingly, for well-delimited tasks, machines are able to emulate many of the capabilities of conversational exchange. The power of complex computers can therefore be harnessed to societal needs without burdening the user beyond knowledge of natural spoken language.

Because humans are designed to live in an air atmosphere, it was inevitable that they learn to convey information in the form of longitudinal waves supported by displacement of air molecules. But of the myriad types of acoustic information signals, speech is a very special kind. It is constrained in three important ways:

- by the physics of sound generation in the vocal system,
- by the properties of human hearing and perception, and
- by the conventions of language.

These constraints have been central to research in speech and remain of paramount importance today.

This paper proposes to comment on the field of speech communication in three veins:

- first, in drawing a brief *perspective on the science*;
- second, in suggesting *critical directions of research*; and
- third, in hazarding some *technology projections*.

## FOUNDATIONS OF SPEECH TECHNOLOGY

Speech processing, as a science, might be considered to have been born from the evolution of electrical communication. Invention of the telephone, and the beginning of telecommunications as a business to serve society, stimulated work in network theory, transducer research, filter design, spectral analysis, psychoacoustics, modulation methods, and radio and cable transmission techniques. Early on, the acoustics and physiology of speech generation were identified as critical issues for understanding. They remain so today, even though much knowledge has been acquired. Alexander Graham Bell was among those

who probed the principles of speech generation in experiments with mechanical speaking machines. (He even attempted to teach his Skye terrier to articulate while sustaining a growl!) Also, it was recognized early that properties of audition and perception needed to be quantified, in that human hearing typically provides the fidelity criterion for receiving speech information. Psychoacoustic behavior for thresholds of hearing, dynamic range, loudness, pitch, and spectral distribution of speech were quantified and used in the design of early telecommunication systems. But only recently, with advances in computing power, have efforts been made to incorporate other subtleties of hearing—such as masking in time and frequency—into speech-processing algorithms. Also, only recently has adequate attention been turned to analytical modeling of language, and this has become increasingly important as the techniques for text-to-speech synthesis and automatic recognition of continuous speech have advanced.

About the middle of this century, sampled-data theory and digital computation simultaneously emerged, opening new vistas for high-quality long-distance communication and for simulating the engineering design of complex systems rapidly and economically. But computing technology soon grew beyond data sorting for business and algorithm simulation for science. Inexpensive arithmetic and economical storage, along with expanding knowledge of information signals, permitted computers to take on functions more related to decision making—understanding subtle intents of the user and initiating ways to meet user needs. Speech processing—which gives machines conversational capability—has been central to this development. Image processing and, more recently, tactile interaction have received similar emphases. But all these capabilities are of little use unless society can afford them. Explosive advances in microelectronics over the past two decades have assured affordable access to this sophistication as well as to the underlying computing technology. All indications are that computing advances will continue and that economical computation to support speech technology will be in place when it is needed.

## INCENTIVES IN SPEECH RESEARCH

Ancient experimentation with speech was often fueled by the desire to amaze, amuse, or awe. Talking statues and gods were favored by early Greeks and Romans. But sometimes fundamental curiosity was the drive (the Czar awarded Kratzenstein a prize for his design of acoustic resonators which when excited from a vibrating reed, simulated vowel timbres). And sometimes the efforts were not given scientific credence (von Kempelen's talking machine was largely ig



FIGURE 1 Ancients used talking statues to amaze, amuse, and awe.

nored because of his chess-playing "automaton" that contained a concealed human! (Dudley and Tarnoczy, 1950).

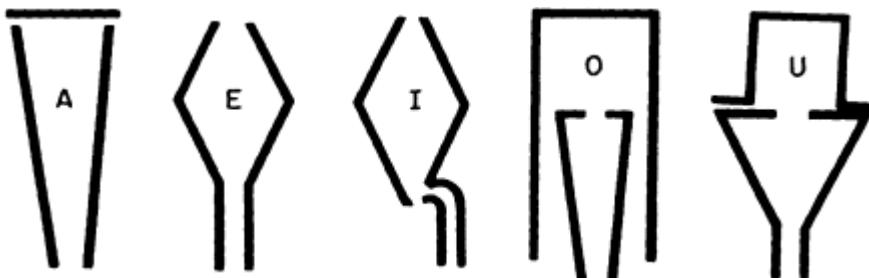


FIGURE 2 Kratzenstein's prize-winning implementation of resonators to simulate human vowel sounds (1779). The resonators were activated by vibrating reeds analogous to the vocal cords. The disparity with natural articulatory shapes points up the nonuniqueness between sound spectrum and resonator shape (i.e., job security for the ventriloquist).

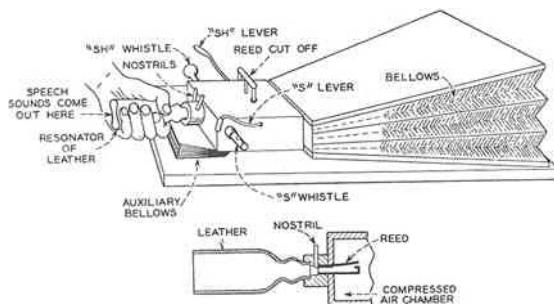


FIGURE 3 Reconstruction of von Kempelen's talking machine (1791), attributed to Sir Charles Wheatstone (1879). Typically, one arm and hand laid across the main bellows and output resonator to produce voiced sounds, while the other hand operated the auxiliary bellows and ports for voiceless sounds.

Acoustic waves spread spherically and do not propagate well over distances. But communication over distances has long been a need in human society. As understanding of electrical phenomena progressed, the electrical telegraph emerged in the mid-nineteenth century. Following this success with dots and dashes, much attention

turned to the prospect of sending voice signals over electrical wires. Invention of the telephone is history.

In the early part of the twentieth century, the incentive remained voice communication over still greater distances. Amplification of analog signals, which attenuate with distance and accumulate noise, was needed. In 1915 transcontinental telephone was achieved with marginal fidelity by electromechanical "repeaters." Transatlantic telegraph cables could not support the bandwidth needed for voice, and research efforts turned to "vocoders" for bandwidth compression. In 1927, as electronics technology emerged, transatlantic radio telephone became a reality. Understanding of bandwidth compression was then applied to privacy and encryption. Transatlantic voice on wire cable had to await the development of reliable submersible amplifiers in 1956. With these expensive high-quality voice circuits, the interest in bandwidth conservation again arose and stimulated new developments, such as Time Assignment Speech Interpolation, which provided nearly a three-fold increase in cable capacity.

From the mid-twentieth century, understanding emerged in sampled-data techniques, digital computing, and microelectronics. Stimulated by these advances, a strong interest developed in human/machine communication and interaction. The desire for ease of use in complex machines that serve human needs focused interest on spoken language communication (Flanagan et al., 1970; Rabiner et al., 1989). Significant advances in speech recognition and synthesis resulted. Bandwidth conservation and low bit-rate coding received emphasis as much for economy of storage (in applications such as voice mail) as for savings in transmission capacity. The more recent developments of mobile cellular, personal, and cordless telecommunications have brought renewed interest in bandwidth conservation and, concomitantly, a heightened incentive for privacy and encryption.

As we approach the threshold of the twenty-first century, fledgling systems are being demonstrated for translating telephony. These systems require automatic recognition of large fluent vocabularies in one language by a great variety of talkers; transmission of the inherent speech information; and natural-quality synthesis in a foreign language—preferably with the exact voice quality of the original talker. At the present time, only "phrase book" type of translation is accomplished, with limited grammars and modest vocabularies, and the synthesized voice does not duplicate the quality of individual talkers. Translating telephony and dictation machines require major advances in computational models of language that can accommodate natural conversational grammars and large vocabularies. Recognition systems using models for subword units of speech are envi

sioned, with linguistic rules forming (a) acceptable word candidates from the estimated strings of Phonetic units, (b) sentence candidate from the word strings, and (c) semantic candidates from the sentences. Casual informal conversational speech, with all its vagaries and nongrammatical structure, poses special challenges in devising tractable models of grammar, syntax, and semantics.



FIGURE 4a Concept demonstration of translating telephony by NEC Corporation at Telecom 1983, Geneva. The application scenario was conversation between a railway stationmaster in Japan and a British tourist who had lost her luggage. Real-time, connected speech, translated between Japanese and English, used a delimited vocabulary and "phrase book" grammar.

## TECHNOLOGY STATUS

A fundamental challenge in speech processing is how to represent, quantify, and interpret information in the speech signal. Traditionally, research focuses on the sectors of coding, speech and speaker recognition, and synthesis.

### Coding.

High-quality digital speech coding has been used for many years in the form of Pulse Code Modulation (PCM), using a typical transmission rate of 64k bits/second. In recent years, capacity-expanding Adaptive Differential PCM (ADPCM) at 32k bits/second has served in the telephone plant, particularly for

private lines. Economical systems for voice mail have derived from compression algorithms for 16k bits/second Sub-Band Coding and low-delay Code Excited Linear Prediction (CELP), and this technology—implemented for 8k bits/second—is currently being tested in digital mobile cellular telephones.



FIGURE 4b An international joint experiment on interpreting telephony was held in January 1993, linking ATR Laboratories (Japan), Carnegie-Mellon University (United States), Siemens A. G. (Germany), and Karlsruhe University (Germany). Spoken sentences were first recognized and translated by a computer into written text, which was sent by modem over a telephone line. A voice synthesizer at the receiving end then spoke the translated words. The system demonstrated was restricted to the task of registering participants for an international conference. (Photograph courtesy of ATR Laboratories, Japan.)

Signal quality typically diminishes with coding rate, with a notable "knee" at about 8k bits/second. Nevertheless, vocoder rates of 4k and 2k bits/second are finding use for digital encryption over voice bandwidth channels. The challenge in coding is to elevate quality at low transmission rates. Progress is being made through incorporation of perceptual factors and through improved representation of spectral and excitation parameters (Jayant et al., 1990).

There are experimental reasons to believe that high quality can be achieved at rates down to the range of 2000 bits/second. Improve

ments at these rates may come from two directions: (i) dynamic adaptation of perceptual criteria in coding, and (ii) articulatory modeling of the speech signal.

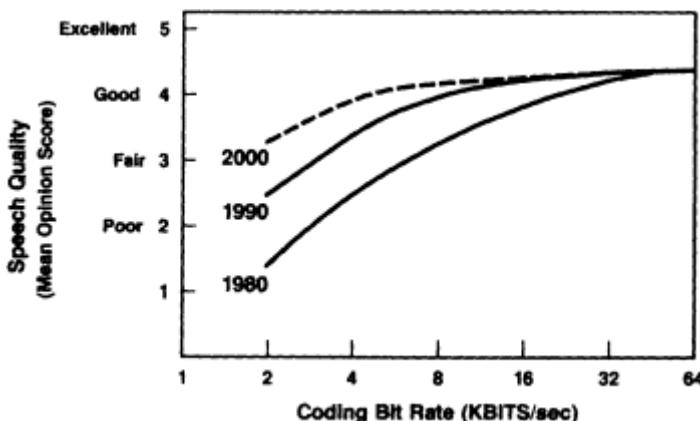


FIGURE 5 Influence of coding rate on the quality of telephone bandwidth speech. Increasingly complex algorithms are used as coding rate diminishes. The research effort focuses on improving quality and immunity to interference at coding rates of 8 kbps and lower.

In coding wideband audio signals the overt use of auditory perception factors within the coding algorithm ("hearing-specific" coders) has been remarkably successful, allowing wideband signal representation with an average of less than two bits per sample. The implication of this is that FM stereo broadcast quality can be transmitted over the public switched digital telephone channels provided by the basic-rate ISDN (Integrated Services Digital Network). Alternatively, one can store up to eight times more signal on a high-fidelity compact disc recording than is conventionally done.

For stereo coding, the left-plus-right and left-minus-right signals are transform-coded separately (typically by 2048-point FFTs). For each spectrum at each moment, a masking threshold is computed, based on the distribution of spectral energy and on critical-band masking in the ear. Any signal components having spectral amplitudes less than this threshold will not be heard at that moment in the presence of stronger neighbors; hence, these components need not be allocated any bits for transmission. Similarly, if bits are assigned to the stronger components so that the quantizing noise spectrum is maintained below this masking threshold, the quantizing noise will not be au

dible. The computation to accomplish the coding, while substantial, is not inordinate in terms of presently available DSP chips.

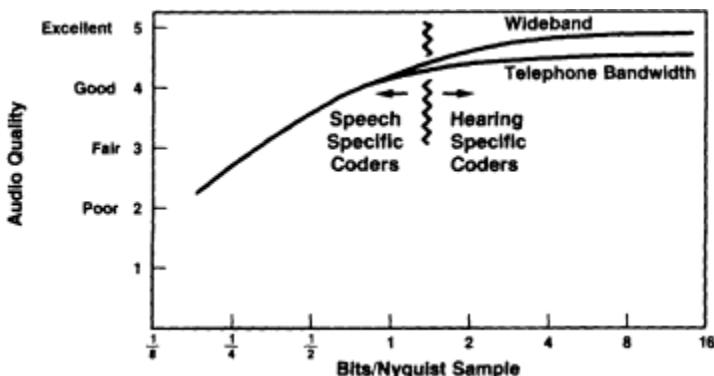


FIGURE 6 Influence of digital representation on audio signal quality. Increasingly complex algorithms are used as representation bits per sample diminish. Hearing-specific coders incorporate human perceptual factors, such as masking in frequency.

This and related techniques are strongly influencing international standards for speech and music coding. And it appears that continued economies can be won through perceptual factors such as masking in the time dimension. (See subsequent discussion of temporal masking.)

### **Recognition and synthesis.**

Unhappily, advances in recognition and in synthesis, particularly in text-to-speech synthesis, have not been strongly coupled and have not significantly cross-nurtured one another. This seems to be largely because recognition has taken a pattern-matching direction, with the immensely successful hidden Markov models (HMMs), while synthesis has relied heavily on acoustic phonetics, with formant models and fractional-syllable libraries contributing to the success. Nevertheless, the techniques are destined to be used hand in hand in voice-interactive systems. Both can benefit from improved computational models of language.

Present capabilities for machine dialogue permit intelligent fluent interaction by a wide variety of talkers provided the vocabulary is limited and the application domain is rigorously constrained (Flanagan, 1992). Typically, a finite-state grammar is used to provide enough coverage for useful conversational exchange. Vocabularies of a couple hundred words and a grammar that permits billions of sentences about a specific task—say, obtaining airline flight information—are

typical. Word recognition accuracy is above 90 percent for vocabularies of several hundred words spoken in connected form by a wide variety of talkers. For smaller vocabularies, such as the digits, recognition accuracies are also in the high 90s for digit strings (e.g., seven-digit telephone numbers) spoken in connected form. With currently available signal processor chips the hardware to support connected-digit recognition is relatively modest.

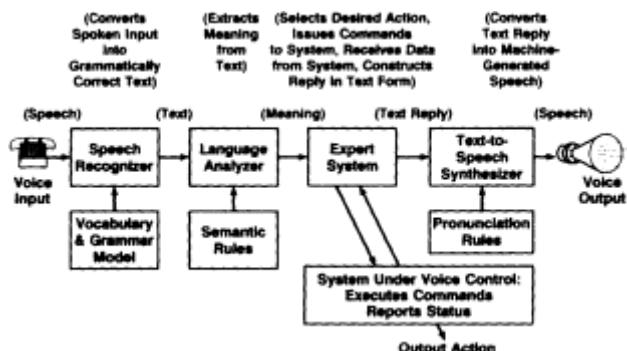


FIGURE 7 Recognition and synthesis systems permit task-specific conversational interaction. Expansions of vocabulary size, talker independence, and language models that more nearly approach natural spoken language, together with high-quality synthesis, are research targets (Flanagan, 1992).

Again, a significant frontier is in developing computational models of language that span more natural language and permit unfettered interaction. Computational linguistics can make strong contributions in this sector.

### **Talker verification.**

Using cepstrum, delta cepstrum, and HMM techniques, the ability to authenticate "enrolled" talkers over clean channels is relatively well established (Soong and Rosenberg, 1988). The computation needed is easily supported, but not much commercial deployment has yet been seen. This results not so much from any lack of desire to have and use the capability but to an apparently low willingness to pay for it. Because speech recognition and talker verification share common processes, combining the features in an interface is natural. The investment in recognition can thereby provide verification for a minimal increment in cost. New applications of this type are emerging in the banking sector where personal verification is needed for services such as cash-dispensing automatic teller machines.

### **Autodirective microphone arrays.**

In many speech communication environments, particularly in teleconferencing and in the use of voice-

interactive terminals, it is more natural to communicate without handheld or body-worn microphones. The freedom to move about the work place, without tether or encumbrance, and to speak as in face-to-face conversation is frequently an advantage. Autodirective microphone arrays, especially beam-forming systems, permit good-quality sound pickup and mitigate the effects of room reverberation and interfering acoustic noise (Flanagan et al., 1991).

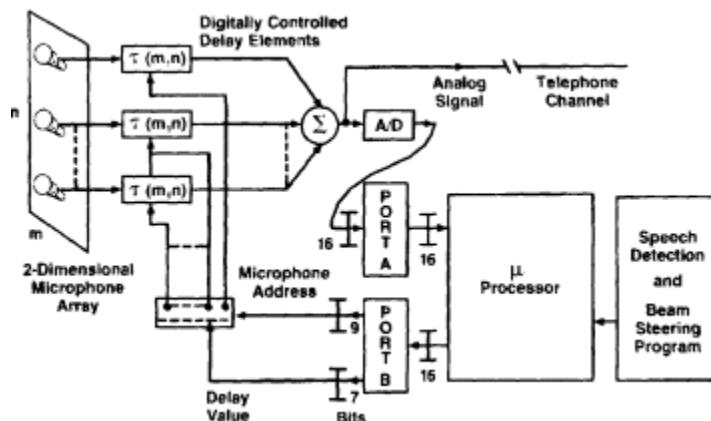


FIGURE 8a Beam-forming, signal-seeking microphone arrays permit natural communication without hand-held or body-worn microphones.

High-performance, low-cost electret microphones, in combination with economical distributed signal processors, make large speech-seeking arrays practical. Each sensor can have a dedicated processor to implement beam forming and steering. A host controller issues appropriate beam-forming and beam-pointing values to each sensor while supporting algorithms for sound source location and speech/ nonspeech identification. The array is typically used with multiple beams in a "track-while-scan" mode. New research on three-dimensional arrays and multiple beam forming is leading to high-quality signal capture from designated spatial volumes.

## CRITICAL DIRECTIONS IN SPEECH RESEARCH

### Physics of Speech Generation; Fluid-Dynamic Principles

The aforementioned lack of naturalness in speech generated from compact specifications stems possibly from two sources. One is the synthesizer's crude approximation to the acoustic properties of the

vocal system. The other is the shortcomings in control data that do not adequately reflect natural articulation and prosody. Both of these aspects affect speech quality and certainly affect the ability to duplicate individual voice characteristics.

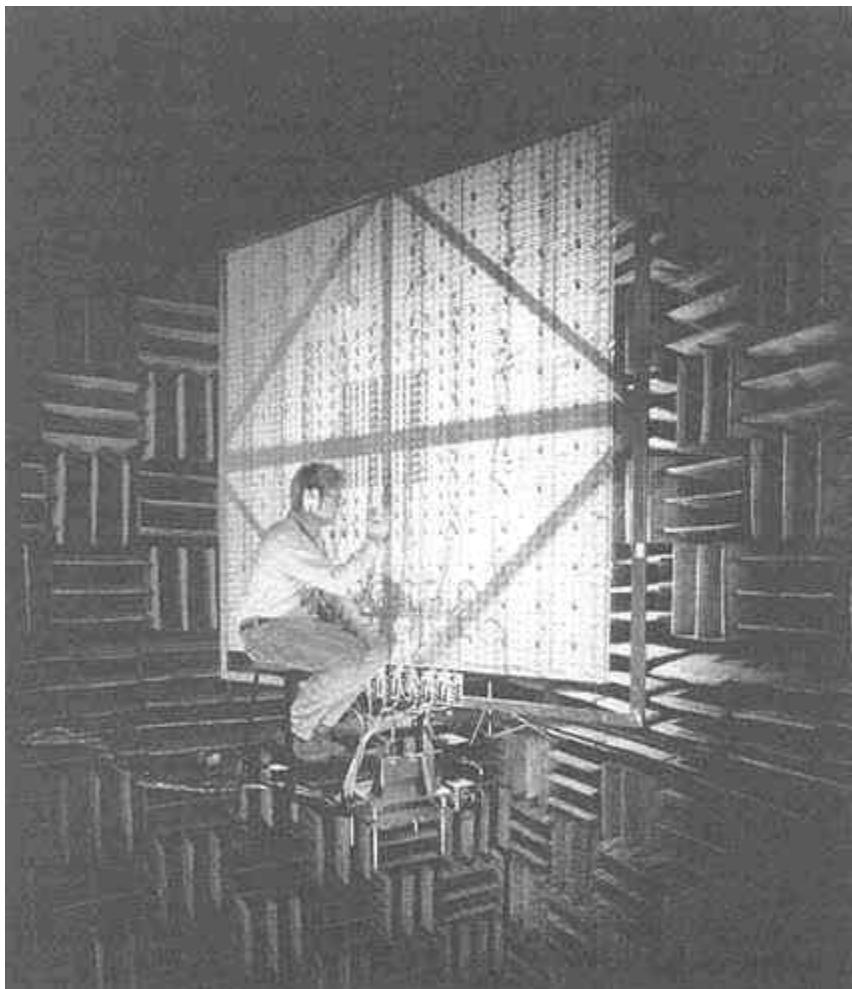


FIGURE 8b Large two-dimensional array of 408 electret microphones. Each microphone has a dedicated chip for beam forming.

Traditional synthesis takes as its point of departure a source-filter approximation to the vocal system, where in source and filter do not interact. Typically, the filter function is approximated in terms of

a hard-walled tube, supporting only linear one-dimensional wave propagation. Neither is realistic.

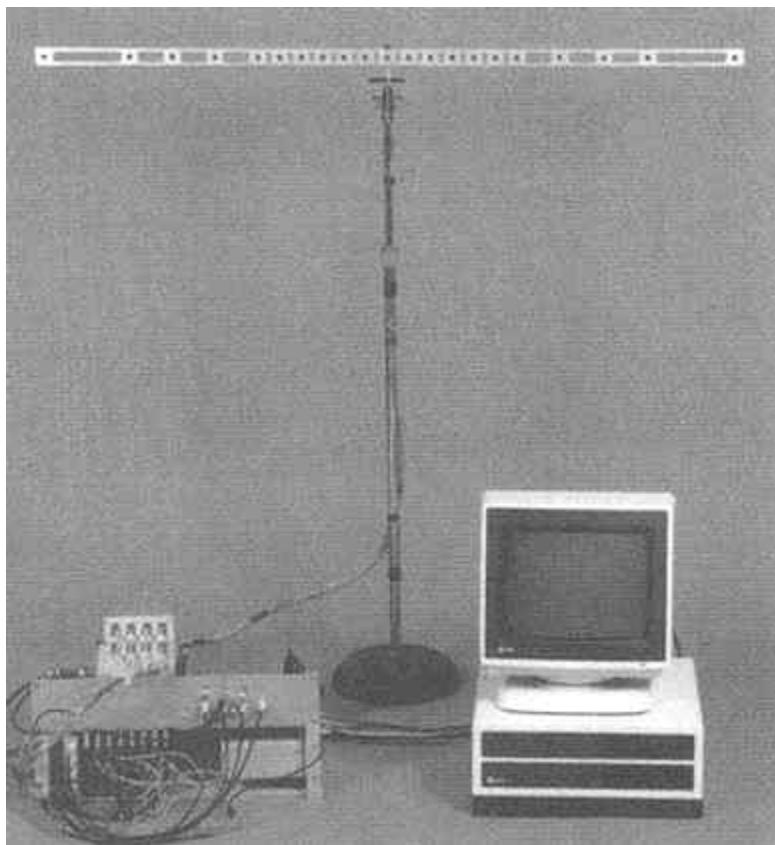


FIGURE 8c One-dimensional track-while-scan beam former for small conference rooms.

Advances in parallel computation open the possibility for implementing speech synthesis from first principles of fluid dynamics. Given the three-dimensional, time-varying, soft-walled vocal tract, excited by periodically valved flow at the vocal cords and by turbulent flow at constrictions, the Navier-Stokes equation can be solved numerically on a fine space-time grid to produce a remarkably realistic description of radiated sound pressure. Nonlinearities of excitation, generation of turbulence, cross-modes of the system, and acoustic interaction between sources and resonators are taken into account. The formula

tion requires enormous computation, but the current initiatives in high-performance computing promise the necessary capability.

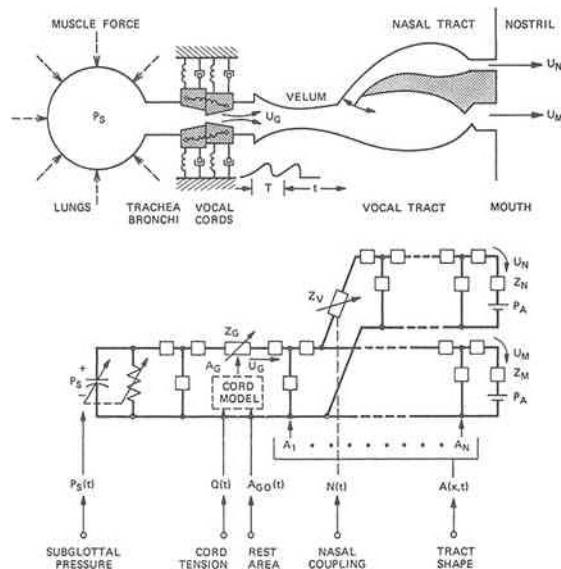


FIGURE 9a Traditional representation of sound generation and propagation in the vocal tract. One-dimensional approximation of sound propagation permits computation of pressure and velocity distributions along tract and at radiating ports. Turbulent excitation is computed from the Reynolds number at each location along the tract. Vocal cord simulation permits source-filter interaction.

### Computational Models of Language

Already mentioned is the criticality of language models for fluent, large-vocabulary speech recognition. Tractable models that account for grammatical behavior (in spoken language), syntax, and

semantics are needed for synthesis from text as urgently as for recognition. Statistical constraints in spoken language are as powerful as those in text and can be used to complement substantially the traditional approaches to parsing and determining parts of speech.

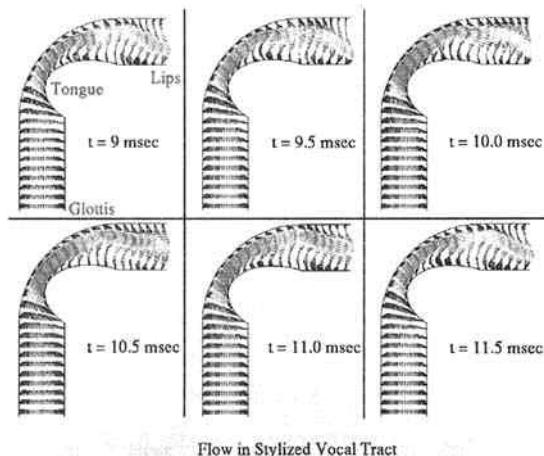


FIGURE 9b Sound generation in the vocal tract computed from fluid-dynamic principles. The magnitude and direction of the velocity vector at each point in two dimensions, in response to a step of axial velocity at the vocal cords, are calculated on a supercomputer (after Don Davis, General Dynamics). Warm color highlights regions of high-velocity amplitude. The plot shows flow separation downstream of the tongue constriction and nonplanar wavefronts.

### **Information Processing in the Auditory System; Auditory Behavior**

Mechanics and operation of the peripheral ear are relatively well understood. Psychoacoustic behavior is extensively quantified. Details of neural processing, and the mechanism for interpreting neural

<u>TRIGRAM</u>	<u>PROBABILITIES (%)</u>			
<u>TRIGRAM</u>	ITALIAN	JAPANESE	GREEK	FRENCH
igh	3	0	0	9
ett	70	0	3	22
cci	25	0	0	0
fuj	0	30	0	0
oto	0	61	14	0
mur	0	86	0	0
los	4	0	65	0
dis	3	0	74	5
kis	0	6	73	0
euv	0	0	0	9
nie	1	0	2	50
ois	10	6	0	61
geo	0	0	38	14
eil	0	0	0	50

FIGURE 10a Illustrative probabilities for selected text trigrams across several languages (10 in total). While the number of possible trigrams is on the order of 20,000, the number of trigrams that actually occur in the language is typically fewer by an order of magnitude—constituting great leverage in estimating allowed symbol sequences within a language and providing a tool for estimating etymology from the individual probabilities.

<u>COMPUTED ESTIMATES OF ETYMOLOGY</u>				
<u>NAME</u>	<u>LIKELIHOOD RATIO (R)*</u>			
ALDRIGHETTI	0.65	IT	0.24	L
ANGELETTI	1.00	IT		
BELLOTTI	1.00	IT		
IANNUCCI	1.00	IT		
ITALIANO	1.00	IT		
LOMBARDINO	0.58	IT	0.42	SP
MARCONI	0.58	IT		
OLIVETTI	1.00	IT		
ASAHARA	1.00	JA		
ENOMOTO	1.00	JA		
FUJIMAKI	1.00	JA		
FUJIMOTO	1.00	JA		
FUJIMURA	1.00	JA		
FUNASAKA	1.00	JA		
TOYOTA	1.00	JA		
UMEDA	0.96	JA		
AGNOSTOPoulos	1.00	GK		
DEMETRIADIS	1.00	GK		
DUKAKIS	0.99	RU		
ANNETTE	0.95	FR		
DENEUVE	0.92	FR	0.14	OF
BANTEGNIE	0.82	FR	0.34	ME
GRANGEois	0.93	FR	0.06	OF
BAGUENARD	0.24	MF	0.52	L
MIREILLE	0.94	FR	0.14	ME

\*CANDIDATES WITH  $R \geq 0.05$

FIGURE 10b Examples of etymology estimates for proper names. The estimate is based on the likelihood ratio (ratio of the probability that the name string belongs to language j, to the average probability of the name string across all languages). The languages included are English, French, German, Japanese, Greek, Russian, Swedish, Spanish, Italian, and Latin. (Data from K. Church, AT&T Bell Laboratories.)

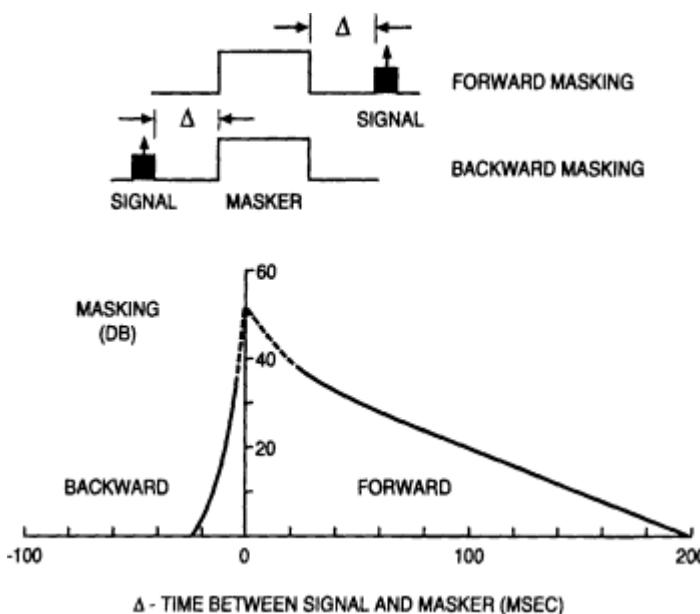


FIGURE 11a Masking in time. A loud sound either before or after a weaker one can raise the threshold of detectability of the latter.

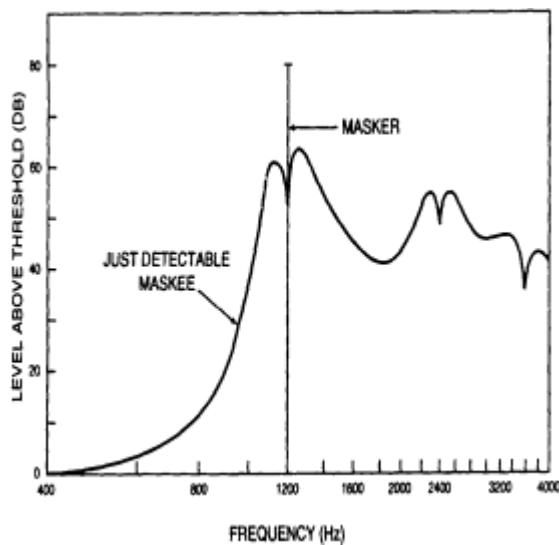


FIGURE 11b Masking in frequency. A loud tone (at 1200 Hz here) can elevate the threshold of detectability of an adjacent tone, particularly one higher in frequency.

information, are not well established. But this does not preclude beneficially utilizing behavioral factors in speech processing. Over the past, telecommunications and audio technology have exploited major aspects of human hearing such as ranges of frequency, amplitude, and signal-to-noise ratio. But now, with inexpensive computation, additional subtleties can be incorporated into the representation of audio signals. Already high-fidelity audio coding incorporates some constraints of simultaneous masking in frequency. Masking in time is an obvious target of opportunity. Relatively untouched, so far, is the esoteric behavior of binaural release from masking, where interaural phase markedly controls perceptibility.

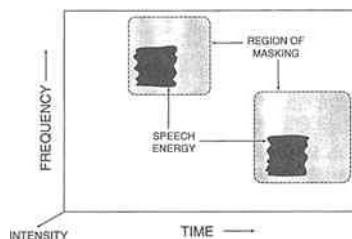


FIGURE 11c Illustration of the time-frequency region surrounding intense, punctuate signals where masking in both time and frequency is effective.

### Coalescing Speech Coding, Synthesis, and Recognition

The issues of coding, recognition, and synthesis are not disjoint—they are facets of the same underlying process of speech and hearing. We might strive therefore for research that unifies the issues from the different sectors. Better still, we might seek an approach that *coalesces* the problems into a common understanding. One such effort is the "voice mimic."

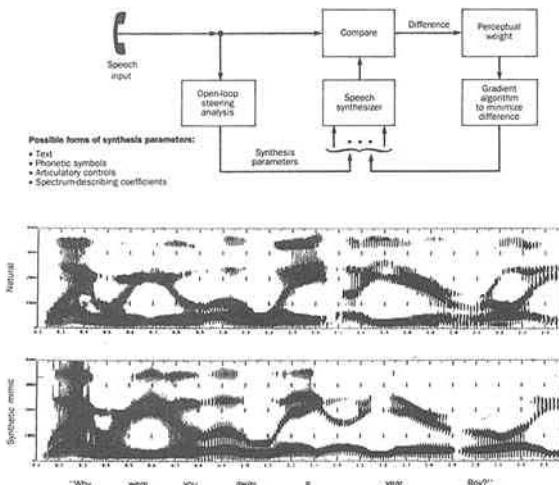


FIGURE 12 Computer voice mimic system. Natural continuous input speech is approximated by a computed synthetic estimate. Spectral differences between real and synthetic signals are perceptually weighted and used in a closed *loop* to adjust iteratively the parameters of the synthesis, driving the difference to a minimum.

The voice mimic attempts to generate a synthetic speech signal that, within perceptual accuracy, duplicates an input of arbitrary natural speech. Central to the effort is a computer model of the vocal cords and vocal tract (to provide the acoustic synthesis), a dynamic model of articulation described by nearly orthogonal vocal tract shape parameters (to generate the cross-sectional area function), and, ideally, a discrete phonetic symbol-to-shape mapping. A perceptually weighted error, measured in the spectral domain for natural and synthetic signals, drives the synthesis parameters so as to minimize the mimicking error, moment by moment. Open-loop analysis of the input natural speech is useful in steering the closed-loop optimization.

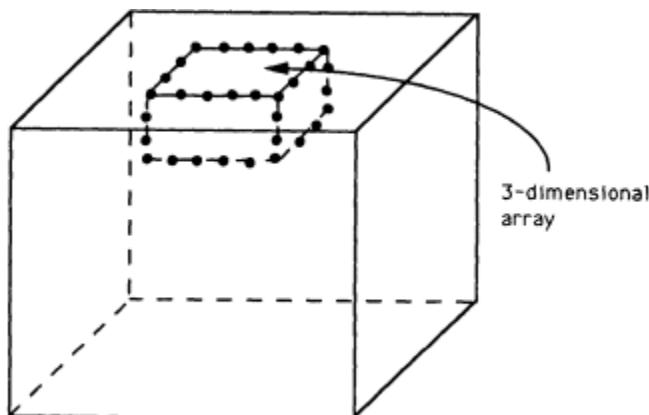


FIGURE 13a Three-dimensional microphone array arranged as a "chandelier" in a reverberant room. Multiple beams are formed and directed to the sound source and its significant images.

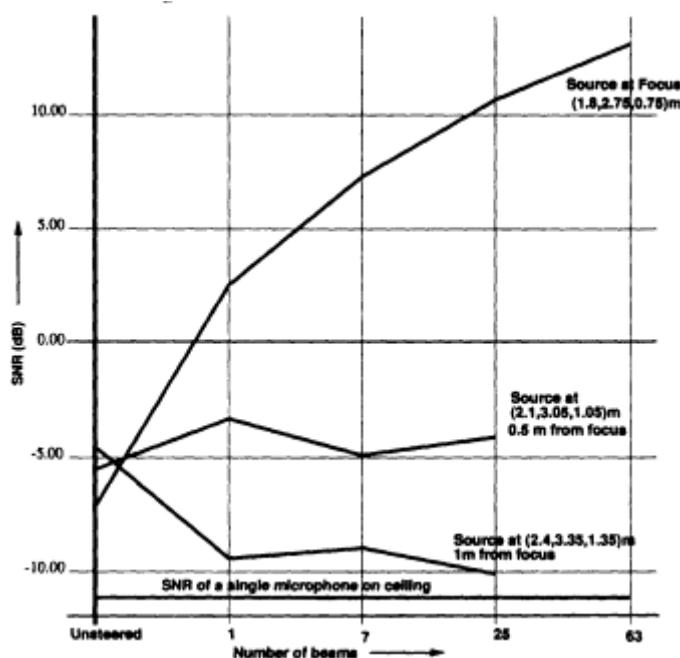


FIGURE 13b Signal-to-noise ratios measured on two octaves of speech for a  $7 \times 7$   $\times 7$  rectangular microphone array positioned at the ceiling center in a computer-simulated hard-walled room of dimensions  $7 \times 5 \times 3$  meters. Source images through third order are computed, and multiple beams are steered to the source and its images.

Ideally, one would like to close the loop at the text level, in which case the problems of recognition, coding, and synthesis coalesce and are simultaneously solved—the result producing as one, a voice typewriter, the ultimate low bit-rate coder, and high-quality text synthesis. Present realities are removed from this, but good success is being achieved on connected input speech at the level of articulatory parameter adjustment.

Lest enthusiasm run too high, it should be quickly mentioned that the required computation is enormous—about 1000 times real-time on a parallel computer. Or, for real-time operation, about 100 billion floating-point operations are required per second (100 Gflops). This amount of computation is not as intimidating or deterring as it once was. Through highly parallel architectures, one can now foresee teraflop capability (though it is less clear how to organize algorithms and software to utilize this power).

### **"Robust" Techniques for Speech Analysis**

Most algorithms for coding and recognition can be made to perform well with "clean" input; that is, with high-quality signal having negligible interference or distortion. Performance diminishes significantly with degraded input. And machine performance diminishes more precipitously than human performance. For example, given a specific level of recognition accuracy, the human listener can typically achieve this level with input signal-to-noise ratios that are 10 to 15 dB lower than that required by typical automatic systems.

A part of this problem appears to be the linear analysis used for most processing. Linear predictive coding, to estimate short-time spectra, is representative. Sizeable durations of the signal contribute to computation of covariance values, so that extensive amounts of noise-contaminated samples are averaged into the analysis. One alternate procedure of interest at present is to eliminate the worst noise-contaminated samples and reconstitute the discarded samples by a nonlinear interpolation algorithm. Another is the use of auditory models of basilar membrane filtering and neural transduction for characterizing signal features.

### **Three Dimensional Sound Capture and Projection**

High-quality, low-cost electret microphones and economical digital signal processors permit the use of large microphone arrays for hands-free sound capture in hostile acoustic environments. Moreover, three-dimensional arrays with beam steering to the sound source and

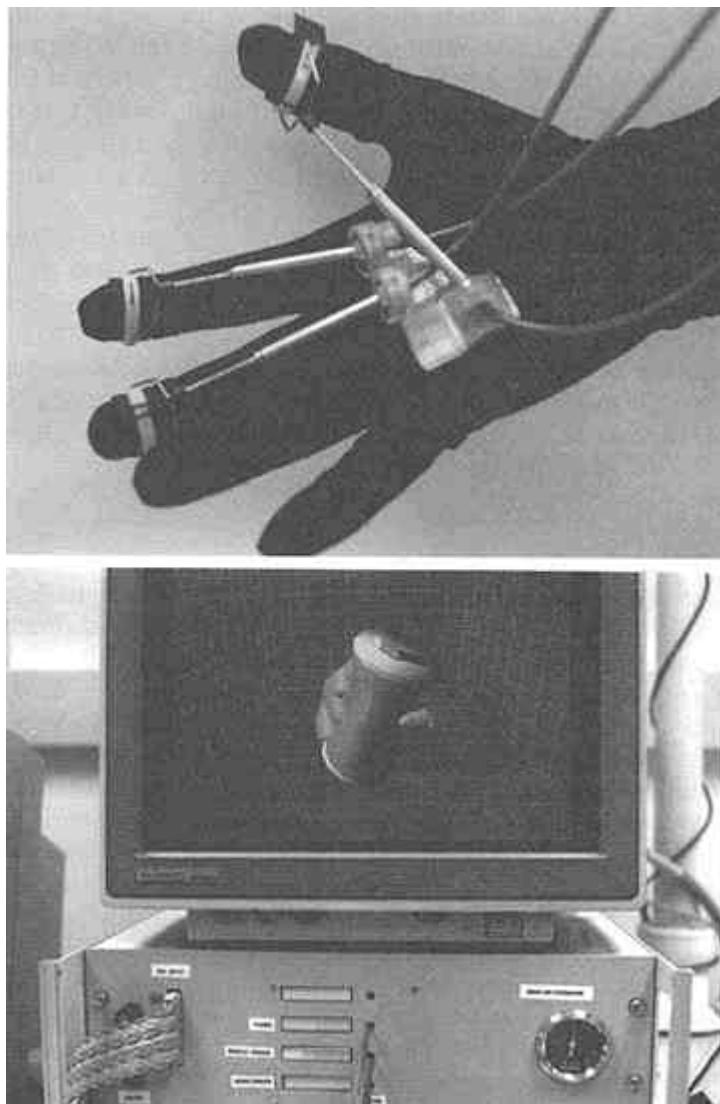


FIGURE 14 (Top) Force feedback applique for a VPL data glove at the CAIP Center. Using the force feedback glove, the wearer can compute a virtual object, and sense tactily the relative position of the object and its programmed compliance. Alternatively, the force feedback device can be programmed for force output sequences for medical rehabilitation and exercise of injured hands. (Bottom) Through the force feedback glove, a user creates and senses plastic deformation of a virtual soft-drink can. (Photograph courtesy of the CAIP Center, Human/Machine Interface Laboratory.)

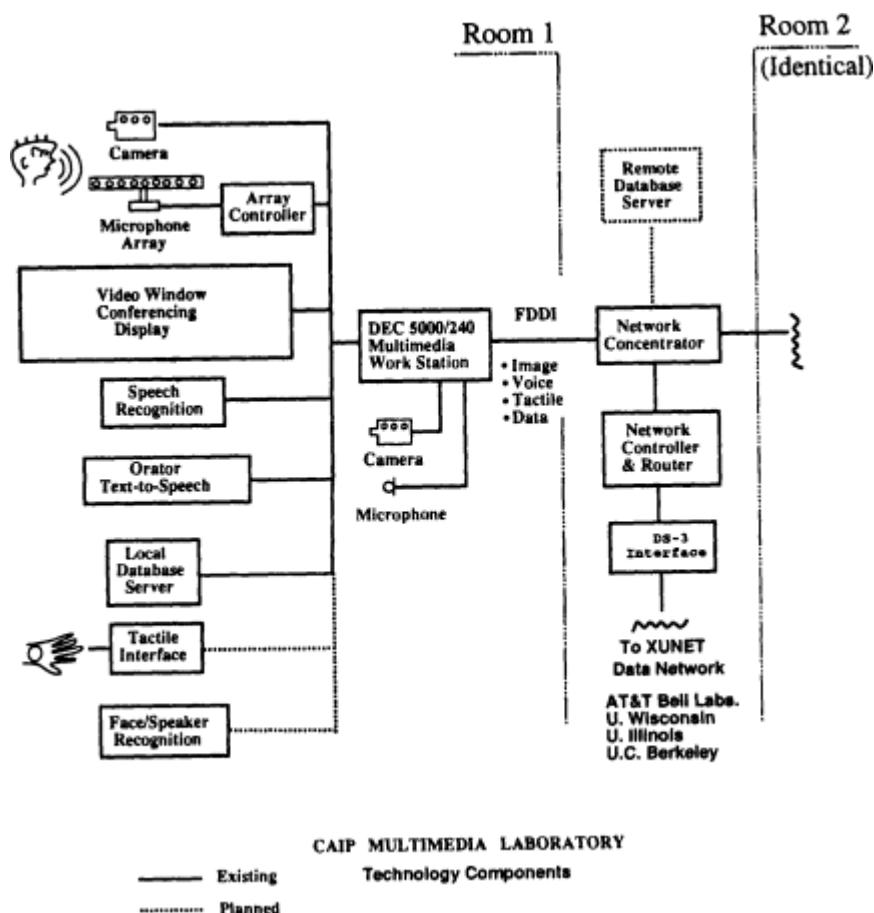


FIGURE 15a Experimental video/audio conferencing system at the CAIP Center, Rutgers University. The system incorporates a number of as-yet imperfect technologies for image, voice, and tactile interaction. The system includes an autodirective beam-steering microphone array, speech recognizer control of call setup and video conferencing display, text-to-speech voice response, image compression for digital transmission, and an interface to the AT&T Bell Laboratories experimental high-speed packet data network, XUNET (Fraser et al., 1992).

multiple significant images in a reverberant enclosure provide significant improvements in pickup quality. Spatial selectivity in three dimensions is a by-product. Computer simulations are providing designs that are being digitally implemented and tested in real environments.



FIGURE 15b Large-screen video projection lends presence for group conferencing and instruction. Auto-directive microphone arrays permit hands-free sound pickup. System features are controlled by automatic recognition of spoken commands. Access to privileged data can be controlled by face and voice recognition for authorized individuals.

Design of receiving arrays is similar to that for transmitting (or projecting) sound—though the costs of transducers for receiving and transmitting differ greatly. Increased spatial realism in sound projection will result from this new understanding.

### Integration of Sensory Modalities for Sight, Sound, and Touch

The human's ability to assimilate information, perceive it, and react is typically more limited in rate than the transmission capacities that convey information to the user terminal. The evolution of global end-to-end digital transport will heighten this disparity and will emphasize the need to seek optimal ways to match information displays to human processing capacity.

Simultaneous displays for multiple sensory modalities provide benefits if they can be appropriately orchestrated. The sensory modalities of immediate interest are *sight*, *sound*, and *touch*. Our understanding of the first two is more advanced than for the latter, but new methods for force feedback transducers on data gloves and "smart skin" implementations aspire to advance tactile technology (Flanagan, *in press*).

Ease of use is directly correlated with successful *integration* of multiple sensory channels. On the speech technology side, this means integration into the information system of the piece parts for speech recognition, synthesis, verification, low bit-rate coding, and hands-free sound pickup. Initial efforts in this direction are designed for conferencing over digital telephone channels (Berkley and Flanagan, 1990). The speech features allow call setup, information retrieval, speaker verification, and conferencing—all totally under voice control. Additionally, low bit-rate coding of color images enables high-quality video transmission over modest capacity.

## SPEECH TECHNOLOGY PROJECTIONS—2000

How good are we at forecasting technology? In my experience, not so good. But not so bad either. I recently got out a set of vugraphs on coding, synthesis, recognition, and audio conferencing that I prepared in 1980. These were made for 5-year and 10-year forecasts as part of a planning exercise. To my surprise about half of the projections were accurate. Notable were subband coding for initial voicemail products (called AUDIX) and 32-kbps ADPCM for transmission economies on private line. But there were some stellar oversights. My 1980 vugraphs of course did not predict CELP, though I was in intimate contact with the fundamental work that led to it.

Despite the intense hazard in anticipating events, several advances seem likely by the year 2000:

- *Signal representation of good perceptual quality at < 0.5 bits/sample.* This will depend on continued advances in microelectronics, especially the incorporation of psychoacoustic factors into coding algorithms.
- *Multilingual text-to-speech synthesis with generic voice qualities.* Multilingual systems are emerging now. The outlook for duplication of individual voice characteristics by rule is not yet supported by fundamental understanding. But generic qualities, such as voice characteristics for man, woman, and child, will be possible.
- *Large-vocabulary (100K-word) conversational interaction with ma*

*chines, with task-specific models of language.* Recognition of unrestricted vocabulary, by any talker on any subject, will still be on the far horizon. But task-specific systems will function reliably and be deployed broadly. A strong emphasis will continue on computational models that approximate natural language.

- *Expanded task-specific language translation.* Systems that go substantially beyond the "phrase-book" category are possible, but still with the task-specific limitation and generic qualities of voice synthesis.
- *Automated signal enhancement, approaching perceptual acuity.* This is among the more problematic estimates, but improved models of hearing and nonlinear signal processing for automatic recognition will narrow the gap between human and machine performance on noisy signals. Comparable recognition performance by human and machine seems achievable for limited vocabularies and noisy inputs. Interference-susceptible communications, such as air-to-ground and personal cellular radio, will benefit.
- *Three-dimensional sound capture and projection.* Inexpensive high-quality electret transducers, along with economical single-chip processors, open possibilities for combatting multipath distortion (room reverberation) to obtain high-quality sound capture from designated spatial volumes. Spatial realism in projection and natural hands-free communication are added benefits. Current research suggests that these advances are supportable.
- *Synergistic integration of image, voice, and tactile modalities.* Although the constituent technologies for sight, sound, and touch will have imperfect aspects for the foreseeable time, proper design of application scenarios will enable productive use of these modalities in interactive workstations. Human factors engineering is central to success. Expanded utility of tactile displays depends on new transducer developments—for example, the design of transducer arrays capable of representing texture in its many subtleties.
- *Requisite economical computing.* Indications are that microelectronic advances will continue. Presently deployed on a wide basis is 0.9- $\mu\text{m}$  technology that provides computations on the order of 50 Mflops on a single chip and costs less than a dollar per Mflop. By 2000, the expectation is for wide deployment of 0.35- $\mu\text{m}$  (and smaller) technology, with commensurate gate densities. Computation on the order of 1 Gflop will be available on a single chip. This availability of computing will continually challenge speech researchers to devise algorithms of enormous sophistication. If the challenge is in fact met, the year 2001 may actually see a HAL-like conversational machine.

## ACKNOWLEDGMENTS

In addition to current university research, this paper draws liberally from material familiar to me over a number of years while at AT&T Bell Laboratories, for whom I continue as a consultant. I am indebted to Bell Labs for use of the material and for kind assistance in preparing this paper. I am further indebted to the Eighteenth Marconi International Fellowship for generous support of this and related technical writings.

## REFERENCES

- Berkley, D. A., and J. L. Flanagan, "HuMaNet: An experimental human/machine communication network based on ISDN," *AT&T Tech. J.*, 69, 87-98 (Sept./Oct. 1990).
- Dudley, H. O., and T. H. Tarnoczy, "The speaking machine of Wolfgang von Kempelen," *J. Acoust. Soc. Am.*, 22, 151-166 (1950).
- Flanagan, J. L., "Speech technology and computing: A unique partnership," *IEEE Commun.*, 30(5), 84-89 (May 1992).
- Flanagan, J. L., "Technologies for multimedia communications," *Proc. IEEE, Special Issue* (in press).
- Flanagan, J. L., C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, 22-45 (Oct. 1970).
- Flanagan, J. L., D. A. Berkley, G. W. Elko, J. E. West, and M. M. Sondhi, "Autodirective microphone systems," *Acustica*, 73, 58-71 (Feb. 1991).
- Fraser, A. G., C. R. Kalmanek, A. E. Kaplan, W. T. Marshall, and R. C. Restrick, "XUNET 2: A nationwide testbed in high-speed networking," *Proc. INFOCOM '92*, Florence, Italy, May 1992.
- Jayant, N. S., V. B. Lawrence, and D. P. Prezas, "Coding of speech and wideband audio," *AT&T Tech. J.*, 69(5), 25-41 (Sept./Oct. 1990).
- Rabiner, L. R., B. S. Atal, and J. L. Flanagan, "Current methods for digital speech processing," pp. 112-132 in *Selected Topics in Signal Processing*, S. Haykin (ed.), Prentice-Hall, New York (1989).
- Soong, F. K., and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-36, 871-879 (June 1988).

## BIBLIOGRAPHY

- Fant, G., *Acoustic Theory of Speech Production*, Mouton and Co., s'Gravenhage, Netherlands, 1960.
- Flanagan, J. L., *Speech Analysis, Synthesis and Perception*, Springer Verlag, New York, 1972.
- Furui, S., and Sondhi, M., eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, 1992.
- Furui, S., *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York, 1989.
- Ince, A. N., ed., *Digital Speech Processing*, Kluwer Academic Publishers, Boston, 1992.

- Jayant, N. S., and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1984.
- Lee, E. A., and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, Boston, 1988.
- Olive, J. P., A. Greenwood, and J. Coleman, *Acoustics of American English Speech—A Dynamic Approach*, Springer Verlag, New York, 1993.
- O'Shaughnessy, D., *Speech Communication; Human and Machine*, Addison-Wesley Publishing Co., New York, 1987.
- Rabiner, L. R., and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993.
- Rabiner, L. R., and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, N.J., 1978.

## **SPEECH SYNTHESIS TECHNOLOGY**



# Computer Speech Synthesis: Its Status and Prospects

*Mark Liberman*

## SUMMARY

Computer speech synthesis has reached a high level of performance, with increasingly sophisticated models of linguistic structure, low error rates in text analysis, and high intelligibility in synthesis from phonemic input. Mass market applications are beginning to appear. However, the results are still not good enough for the ubiquitous application that such technology will eventually have. A number of alternative directions of current research aim at the ultimate goal of fully natural synthetic speech. One especially promising trend is the systematic optimization of large synthesis systems with respect to formal criteria of evaluation. Speech recognition has progressed rapidly in the past decade through such approaches, and it seems likely that their application in synthesis will produce similar improvements.

Many years ago at Bell Laboratories, Joseph Olive and I sat late one evening at a computer console. We were listening with considerable satisfaction to the synthetic speech produced by a new program. A member of the custodial staff, who had been mopping the floor in the hall outside, stuck his head in the door, furrowed his brow, and asked: "They got dogs in here?"

Things have changed considerably since then. For one thing, computer

consoles no longer exist. Also, speech synthesis has improved a great deal. The best systems—of which the current Bell Labs system is surely an example—are entirely intelligible, not only to their creators but also to the general population, and sometimes they even sound rather natural. Here I will give a personal view of where this technology stands today and where it seems to be headed. This assessment distills contributions from participants in the colloquium presentations and discussion, who deserve the credit for any useful insights. Omissions, mistakes, and false predictions are of course my own.

The ongoing microelectronics revolution has created a striking opportunity for speech synthesis technology. The computer whose console was mentioned earlier could not do real-time synthesis, even though it filled most of a room and cost hundreds of thousands of dollars. Almost every personal computer now made is big and powerful enough to run high-quality speech synthesis in real-time, and an increasing number of such machines now come with built-in audio output. Small battery-powered devices can offer the same facility, and multiple channels can be added cheaply to telecommunications equipment.

Why then is the market still so small? Partly, of course, because the software infrastructure has not yet caught up with the hardware. Just as widespread use of graphical user interfaces in applications software had to wait for the proliferation of machines with appropriate system-level support, so widespread use of speech synthesis by applications will depend on common availability of platforms offering synthesis as a standard feature. However, we have to recognize that there are also remaining problems of quality. Today's synthetic speech is good enough to support a wide range of applications, but it is still not enough like natural human speech for the truly universal usage that it ought to have.

If there are also real prospects for significant improvement in synthesis quality, we should consider a redoubled research effort, especially in the United States, where the current level of research in synthesis is low compared to Europe and Japan. Although there is excellent synthesis research in several United States industrial labs, there has been essentially no government-supported synthesis research in the United States for some time. This is in sharp contrast to the situation in speech recognition, where the ARPA's Human Language Technology Program has made great strides, and also in contrast to the situation in Europe and Japan. In Europe there have been several national and European Community-level programs with a focus on synthesis, and in Japan the ATR Interpreting Telephony Laboratory has made significant investments in synthesis as well as recognition.

technology. There are a number of new ideas at all levels of the problem and also a more general sense that a methodology similar to the one that has worked so well in speech recognition research will also raise speech synthesis quality to a new level.

Before considering in more detail what this might mean, we should consider some of the ways in which speech synthesis research has developed differently from speech recognition research. We will start by exploring what is meant by the term *computer speech synthesis*.

Obviously, this term refers to the creation by computer of human-like speech, but that only tells us what the output of the process is. Synthesized speech output may come from a wide range of processes that differ enormously in the nature of their inputs and the nature of their internal structures and calculations.

The input may be

1. an uninterpreted reference to a previously recorded utterance;
2. a message drawn from a small finite class of texts, such as telephone numbers;
3. a message drawn from a larger or even infinite, but still restricted, class of texts, such as names and addresses;
4. a message drawn from unrestricted digital text, including anything from electronic mail to on-line newspapers to patent or legal texts, novels, or cookbooks;
5. a message composed automatically from nontextual computer data structures (which we might think of as analogous to "concepts" or "meanings"); or
6. a specification of the phonological content of a message, which for most applications must be produced from one of the types of input given previously.

Most commercial applications so far have been of type 1 or 2. Classical "text-to-speech" systems are of type 4 and/or 6. Ultimate human-computer interaction systems are likely to be of type 5, with a bit of 4. Many of the people closely involved in applying speech synthesis technology think that the most promising current opportunities are of type 3. Note that choosing such restricted-domain applications has been crucial to the success of computer speech recognition.

The system-internal structures and processes of "speech synthesis" may involve

1. reproduction of digitally stored human voice, perhaps with compression / expansion;

2. construction of messages by concatenation of digitally stored voice fragments;
3. construction of messages by concatenation of digitally stored voice fragments with modifications of the original timing and pitch;
4. construction of messages by concatenation of digitally stored voice fragments with rule-generated synthetic pitch contours and rule-generated segmental timing values;
5. construction of messages using rule-generated synthetic time functions of acoustic parameters;
6. construction of messages using rule-generated synthetic controls for the kinematics of simplified analogs of human vocal tract; and
7. construction of messages by realistic modeling of the physiological and physical processes of human speech production, including dynamic control of articulation and models of the airflow dynamics in the vocal tract.

The largest scale of commercial activity has been of types 1 and 2, which might be called stored voice. This includes telecommunication intercepts, Texas Instruments' Speak 'N Spell toy, voice mail prompts, and so forth. Much classical speech synthesis research was of type 5 or 6. Several of the best current systems, and what some consider to be the most promising areas of research, are of types 3 and 4, techniques that are sometimes called *concatenative synthesis*.

These alternative types of computer-spoken messages, and alternative techniques for producing them, seem so different that people often feel that it is unreasonable to use the same phrase to describe them. Despite many efforts to clarify the terminology, however, there is a stubborn tendency to use *speech synthesis* for all of these cases. This tendency is understandable, since there is indeed a kind of continuum of techniques and applications, and as the range of data-intensive synthesis techniques increases, the category boundaries become increasingly blurred. However, it creates considerable confusion, and so we will adopt a more carefully defined terminology.

The present discussion is focused on inputs of types 3 through 6 (i.e., restricted or unrestricted text, or nontextual computer data structures) and synthesis techniques of types 3 through 6 (which involve producing spoken messages from a phonological specification). The process of transforming text into a suitable phonological specification is generally known as *text analysis*, and the process of creating sound from this specification has (confusingly) no common name other than *speech synthesis*, which as we have seen is used for many other things as well. We will refer to it as *speech synthesis proper*, sometimes abbreviated as *speech synthesis* or *synthesis* if the context is clear.

To put the present research situation in perspective, it is useful to present a bit of history. Lawrence Rabiner did his Ph.D. research on a speech synthesis system almost 30 years ago. In using this work as a point of reference, we do not mean to exaggerate its historical role. In a sketch of the intellectual history of speech synthesis, we would cover the early work of Delattre, Cooper, Holmes, Mattingly, Fant, Dixon, and many others, and Rabiner's dissertation would find its place primarily as an influence on the subsequent research of Dennis Klatt. However, in order to make some general points about trends in speech research over the past three decades, Rabiner's work is a particularly useful point of departure.

The results were presented in Rabiner's 1964 MIT dissertation and also described in a 1968 *Bell System Technical Journal* article. This system used a technique of type 5 (rule-generated time functions of acoustic parameters), with a tinge of type 6 (rule-generated articulatory kinematics) in the control of fundamental frequency, based on a concept of subglottal pressure as the crucial variable. Its input was of type 6, consisting of a string of phonemic symbols with stress indications and marks for word boundaries and pauses; thus, it accomplished "speech synthesis proper," with no text analysis component.

The underlying conception for this system is admirably simple: each phoneme is characterized by a single invariant acoustic target, and the observed contextually varied time functions are generated by a smoothing process. In addition to its specified control parameter values, each phoneme defines a specified frequency region around each of the formant values in that vector, indicating tolerance for coarticulatory modification.

The method for creating actual time functions from these tables is general but somewhat subtle. The parameter time functions are generated by critically damped second-degree differential equations whose time constants depend on the parameter and the pair of phonemes involved. The phonemic goals change discretely in time, but the timing of these changes depends on a nonlinear interaction of the phoneme sequence with the computed durations and the specified formant tolerances. A new set of formant targets is not introduced until the formants have reached the tolerance region of the current phoneme, and a durational criterion (only defined for stressed vowels) is also satisfied. Thus, the method could be informally summarized as "move each parameter under the control of phoneme  $i$  until all parameters are close enough to their target; then continue for a specified time if the phoneme is a stressed vowel; then switch to the target for phoneme  $i + 1$ ." There are some additional complexities, such as the provision for delaying by a specified amount the change in formant targets for certain formants in a few specified phoneme sequences.

This system was state of the art in 1964, but it was more than a decade earlier than the Bell Labs system that the janitor mistook for a dog, and if we played it alongside one of today's systems, it would be quite clear how far we have come in 30 years.

Enormous progress has been made in the area of text analysis (which of course was outside the scope of Rabiner's dissertation). In the 1960s methods for translating English text into phonological strings did not have very good performance. A high proportion of words were mispronounced, and the assignment of phrasing, phrasal stress, and phrasal melody was ineffective. Today's best text analysis algorithms have mispronunciation rates that are best measured in errors per 10,000 input words and do a reasonable (and improving) job of phrasing and accent assignment. There are a number of factors behind the improvement, but the most important reason is that today's programs simply contain much more information about text than their predecessors. This information may be explicit (e.g., lists of words with their pronunciations) or implicit (statistical rules summarizing the behavior of large bodies of training material).

A similar process has characterized the improvements in speech synthesis proper, the production of sound from a given phonological string. Today's systems are still based on the same general strategy of phonological units sequenced in time. However, the inventory of units is much larger, each unit typically involving two, three, or more phonetic segments, either as distinguishing context for the unit or as part of the unit itself. Often, the internal structure of each unit is much more elaborate, sometimes including an entire stretch of fully specified speech. The timing rules distinguish many more cases, and the procedures for selecting units, combining them, and establishing their time patterns are often quite complex. Between larger tables of units and more complex combination rules, today's systems simply incorporate much more information than Rabiner's system did. Measured in terms of the size in bits of the programs and tables, today's systems are probably two to three orders of magnitude larger.

Although this additional complexity seems essential to improved quality, it is a mixed blessing. It may be argued that most of the recent progress in speech recognition research has been due to two factors:

1. simple architectures that permit program parameters to be optimized with respect to large bodies of actual speech and
2. easily calculated objective evaluation metrics that permit alternative designs to be compared quantitatively.

A similar methodology began to be applied in text analysis more than a decade ago, and it has now become the norm in such work. It is the main reason that text analysis has made such rapid progress, to the point that the real quality bottleneck appears to be in speech synthesis proper, the sound production end of the system.

With some notable exceptions, this methodology has been absent from research in speech synthesis proper until recently. Consider Rabiner's system in light of the two success factors just mentioned. His table of phonemic targets would certainly be amenable to corpus-based optimization; indeed, one can optimize arbitrarily large tables of acoustic targets, as long as enough data are brought to bear. However, Rabiner's method for time-function generation has some properties that would make optimization of its constants somewhat tricky and would hinder optimization of the table of phonemic targets as well. It seems clear that the system was not designed with corpus-based optimization in mind; if it had, Rabiner would no doubt have made certain choices somewhat differently.

Rabiner's 1964 work also does not contain any definition of an evaluation metric that would permit alternative architectures to be compared in a quantitative way. For instance, it is now generally accepted that a single acoustic target for each (surface) phonemic segment is not adequate. Rabiner's 1964 work does not specify a framework in terms of which alternative approaches to the question of subphonemic variation could be compared objectively.

Of course, it is entirely unfair to criticize Rabiner's 1964 work in these terms. His design decisions were not made with these aims in mind. His approach instead seems to be based on a different assumption, which it shared with most other synthesis work of the past 30 years—namely, that success would come from the introduction of a modest amount of fairly high level scientific knowledge in the form of human-coded programs. From this point of view, the most important goal is not to design a system that can easily be subjected to systematic formal optimization, but rather a system that will permit the introduction of certain scientific models in a convenient and appropriate form.

This was an appropriate point of view in the context of a system as compact and conceptually simple as Rabiner's was. However, the post-Rabiner direction of research in segmental synthesis was (by necessity) toward expanded tables of values, increased algorithmic complexity, and proliferation of special cases in the time-function generation process. These moves made objective optimization even harder to contemplate; at the same time, they brought systems to a level of complexity that taxed the researchers' ability to manage their

development and modification. Researchers like Klatt certainly paid close attention to speech data in setting their parameters, and they often engaged in informal interactive "copy synthesis" as a method for tuning up parameters and algorithms. However, the resulting systems were certainly not designed to facilitate overall optimization of parameters, or objective comparison of alternative algorithms against a large speech database. As the systems grew larger and larger, and their internal interactions grew more and more complex, interactive experimentation by human developers became a less and less viable method for managing the development process.

One might argue that concatenative synthesis methods caught on earlier to the benefits of explicit grounding in large amounts of speech data. Certainly one general lesson of the past decade has been that systems based on minimal manipulation of large bodies of natural speech data often sound better than systems that do deeper and more sophisticated calculations, with a more complex model of how their primitive elements interact. The high quality achieved by some implementations of methods such as PSOLA (pitch-synchronous overlap-add approach) even suggests to some that the apotheosis of superficiality might extend to time domain over frequency domain methods of signal manipulation. However, even very data-intensive concatenative approaches have usually not been quantitatively optimized in the way that speech recognition algorithms routinely are. Instead, someone simply picks an inventory design, a segmentation scheme, and a set of rules for choosing and combining elements and then sets to work building an inventory by manual accumulation of individual elements.

Only within the past few years have we seen a general use of systematic optimization techniques for purposes of inventory design, unit segmentation, unit selection, and unit combination algorithms. The general approach is to define a perceptually reasonable acoustic distortion metric and use it in a global comparison of alternatives (in allophonic clustering, in segmentation points, in unit selection, or whatever). To make this method work effectively, one must usually design the overall system specifically with such a process in view. Psychological tests would be the optimal basis of such an effort, but objective (if psychologically motivated) distortion metrics have the advantage of being quicker and cheaper. Although such objective distortion metrics are far from a perfect image of the human judgments that provide the ultimate evaluation of any synthesis system, they usually provide the only feasible way to perform the massive and systematic comparison of alternatives that is needed. Testing with

human subjects can then be used to provide validation at strategically chosen points.

Researchers at NTT and ATR in Japan have been especially prominent in these explorations, and their initial results look very promising. As such methods gain wider application, and especially as we see general availability of the large-scale single-speaker databases that will be required to support them, we can hope to see an increased rate of improvement in segmental speech synthesis quality. Thus, increased investment in speech synthesis research is warranted, both because there is an opportunity created by advances in microelectronics and because there are significant new ideas and new methods waiting to be applied.

As this research goes forward, it faces some pointed questions. What will it take to make synthetic speech that sounds entirely natural, or at least better than word concatenation voice response systems for restricted phrase types such as name and address sequences? Will progress come by a scientific route, through better modeling of human speech production, or by an engineering route, through larger inventories of prerecorded elements with optimal automatic selection and combination methods? How far can we push current ideas about text analysis algorithms? How can we produce more natural-sounding modulation of pitch, amplitude, and timing, and how important are such prosodic improvements relative to segmental improvements?

What will it take to put speech synthesis into true mass market applications? What will those applications be? Will the key development be cheaper hardware, a particular "killer" application, or better-quality synthesis? Will there be a gradual spread of the existing niche markets or a single breakthrough?

How should we quantify progress in synthesis quality? What is the proper place for subjective testing relative to objective distortion metrics?

The papers by Carlson and Allen in this volume present a solid foundation of fact for evaluating these questions, and a wide variety of opinions were aired in the symposium discussion, from which an individual point of view has been distilled in this introduction. The next decade will be a lively and interesting time in the field of speech synthesis research, and there is little doubt that the situation will look very different 10 years from now.

# Models of Speech Synthesis

*Rolf Carlson*

## SUMMARY

The term "speech synthesis" has been used for diverse technical approaches. In this paper, some of the approaches used to generate synthetic speech in a text-to-speech system are reviewed, and some of the basic motivations for choosing one method over another are discussed. It is important to keep in mind, however, that speech synthesis models are needed not just for speech generation but to help us understand how speech is created, or even how articulation can explain language structure. General issues such as the synthesis of different voices, accents, and multiple languages are discussed as special challenges facing the speech synthesis community.

## INTRODUCTION

The term "speech synthesis" has been used for diverse technical approaches. Unfortunately, any speech output from computers has been claimed to be speech synthesis, perhaps with the exception of playback of recorded speech.<sup>1</sup> Some of the approaches used to gen

---

<sup>1</sup> The foundations for speech synthesis based on acoustical or articulatory modeling can be found in Fant (1960), Holmes et al. (1964), Flanagan (1972), Klatt (1976), and Allen et al. (1987). The paper by Klatt (1987) gives an extensive review of the developments in speech synthesis technology.

erate true synthetic speech as well as high-quality waveform concatenation methods are presented below.

### **Knowledge About Natural Speech**

Synthesis development can be grouped into three main categories: acoustic models, articulatory models, and models based on the coding of natural speech. The last group includes both predictive coding and concatenative synthesis using speech waveforms. Acoustic and articulatory models have had a long history of development, while natural speech models represent a somewhat newer field. The first commercial systems were based on the acoustic terminal analog synthesizer. However, at that time, the voice quality was not good enough for general use, and approaches based on coding attracted increased interest. Articulatory models have been under continuous development, but so far this field has not been exposed to commercial applications due to incomplete models and high processing costs.

We can position the different synthesis methods along a "knowledge about speech" scale. Obviously, articulatory synthesis needs considerable understanding of the speech act itself, while models based on coding use such knowledge only to a limited extent. All synthesis methods have to model something that is partly unknown. Unfortunately, artificial obstacles due to simplifications or lack of coverage will also be introduced. A trend in current speech technology, both in speech understanding and speech production, is to avoid explicit formulation of knowledge and to use automatic methods to aid the development of the system. Since such analysis methods lack the human ability to generalize, the generalization has to be present in the data itself. Thus, these methods need large amounts of speech data. Models working close to the waveform are now typically making use of increased unit sizes while still modeling prosody by rule. In the middle of the scale, "formant synthesis" is moving toward the articulatory models by looking for "higher-level parameters" or to larger prestored units. Articulatory synthesis, hampered by lack of data, still has some way to go but is yielding improved quality, due mostly to advanced analysis-synthesis techniques.

### **Flexibility and Technical Dimensions**

The synthesis field can be viewed from many different angles. We can group the models along a "flexibility" scale. Multilingual systems demand flexibility. Individual voices, speaking styles, and accents also need a flexible system in which explicit transformations

can be modeled. Most of these variations are continuous rather than discrete. The importance of separating the modeling of speech knowledge from acoustic realization must be emphasized in this context.

In the overview by Furui (1989), synthesis techniques are divided into three main classes: waveform coding, analysis-synthesis, and synthesis by rule. The analysis-synthesis method is defined as a method in which human speech is transformed into parameter sequences, which are stored. The output is created by a synthesis based on concatenation of the prestored parameters. In a synthesis-by-rule system the output is generated with the help of transformation rules that control the synthesis model such as a vocal tract model, a terminal analog, or some kind of coding.

It is not an easy task to place different synthesis methods into unique classes. Some of the common "labels" are often used to characterize a complete system rather than the model it stands for. A rule-based system using waveform coding is a perfectly possible combination, as is speech coding using a terminal analog or a rule-based diphone system using an articulatory model. In the following pages, synthesis models will be described from two different perspectives: the sound-generating part and the control part of the system.

### The Sound-Generating Part

The sound-generating part of the synthesis system can be divided into two subclasses, depending on the dimensions in which the model is controlled. A vocal tract model can be controlled by spectral parameters such as frequency and bandwidth or shape parameters such as size and length. The source model that excites the vocal tract usually has parameters to control the shape of the source waveform. The combination of time-based and frequency-based controls is powerful in the sense that each part of the system is expressed in its most explanatory dimensions. A drawback of the combined approach can be that it makes interaction between the source and the filter difficult. However, the merits seem to outweigh the drawbacks.

#### Simple Waveform Concatenation

The most radical solution to the synthesizer problem is simply to have a set of prerecorded messages stored for reproduction. Simple coding of the speech wave might be performed in order to reduce the amount of memory needed. The quality is high, but the usage is limited to applications with few messages. If units smaller than sentences are used, the quality degenerates because of the problem of

connecting the pieces without distortion and overcoming prosodic inconsistencies. One important and often forgotten aspect in this context is that a vocabulary change can be an expensive and time consuming process, since the same speaker and recording facility have to be used as with the original material. The whole system might have to be completely rebuilt in order to maintain equal quality of the speech segments.

### Analysis-Synthesis Systems

Synthesis systems based on coding have as long a history as the vocoder. The underlying philosophy is that natural speech is analyzed and stored in such a way that it can be assembled into new utterances. Synthesizers such as the systems from AT&T Bell Labs (Olive, 1977, 1990; Olive and Liberman, 1985), Nippon Telephone & Telegraph (NTT) (Hakoda et al., 1990; Nakajima and Hamada, 1988) and ATR Interpreting Telephone Research Laboratories (ATR) (Sagisaka, 1988; Sagisaka et al., 1992) are based on the source-filter technique where the filter is represented in terms of linear predictive coding (LPC) or equivalent parameters. This filter is excited by a source model that can be of the same kind as the one used in terminal analog systems. The source must be able to handle all types of sounds: voiced and unvoiced vowels and consonants.

Considerable success has been achieved by systems that base sound generation on concatenation of natural speech units (Moulines et al., 1990). Sophisticated techniques have been developed to manipulate these units, especially with respect to duration and fundamental frequency. The most important aspects of prosody can be imposed on synthetic speech without considerable loss of quality. The pitch-synchronous overlap-add approach (PSOLA) (Charpentier and Moulines, 1990) methods are based on concatenation of waveform pieces. The frequency domain approach (FD-PSOLA) is used to modify the spectral characteristics of the signal; the time domain approach (TD-PSOLA) provides efficient solutions for real-time implementation of synthesis systems. Earlier systems like SOLA (Roucos and Wilgus, 1985) and systems for divers' speech restoration also did direct processing of the waveform (Liljencrants, 1974).

[Figure 1](#) shows the basic function of a PSOLA-type system. A database of carefully selected utterances is recorded, and each pitch period is marked. The speech signal is split into a sequence of windowed samples of the speech wave. At resynthesis time the waveforms are added according to the desired pitch, amplitude, and duration.

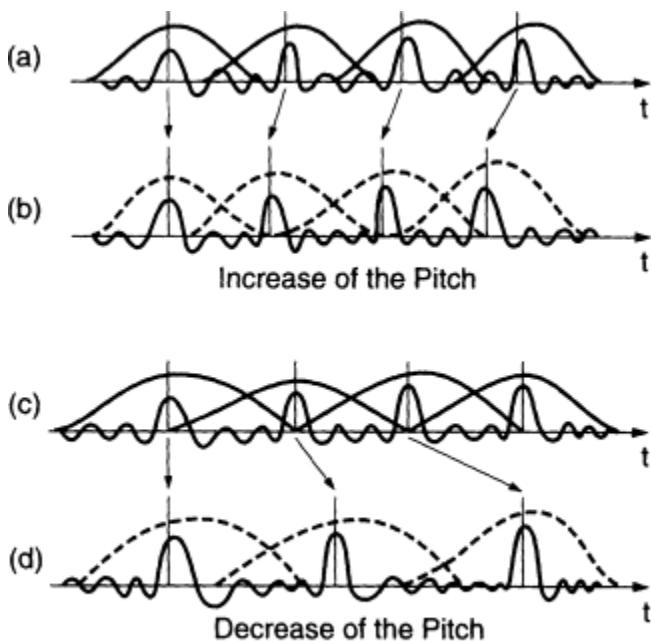


FIGURE 1 Example of the PSOLA method (from Sagisaka, 1990).

### Source Models

The traditional source model for the voiced segments has been a simple or double impulse. This is one reason why text-to-speech systems from the 1980s have had serious problems, especially when different voices are modeled. While the male voice sometimes has been regarded to be generally acceptable, an improved glottal source will open the way to more realistic synthesis of child and female voices and also to more naturalness and variation in male voices.

Most source models work in the time domain with different controls to manipulate the pulse shape (Ananthapadmanabha, 1984; Hedelin, 1984; Holmes, 1973; Klatt and Klatt, 1990; Rosenberg, 1971). One version of such a voice source is the LF-model (Fant et al., 1985). It has a truncated exponential sinusoid followed by a variable cut-off-6dB/octave low-pass filter modeling the effect of the return phase, that is, the time from maximum excitation of the vocal tract to complete closure of the vocal folds. [Figure 2](#) explains the function of the control parameters. In addition to the amplitude and fundamental frequency control, two parameters influence the amplitudes of the two

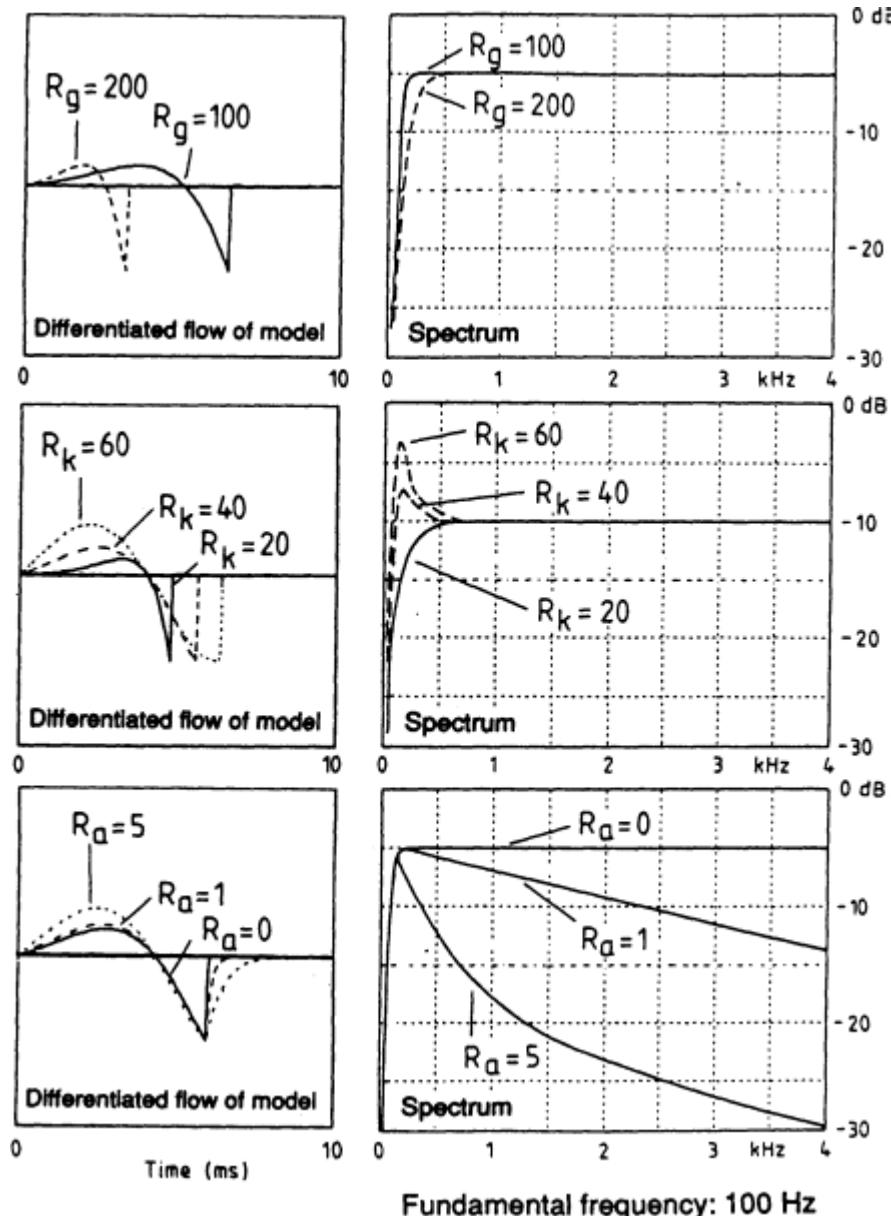


FIGURE 2 Influence of the parameters  $R^g$ ,  $R^k$ , and  $R^a$  on the differentiated glottal flow pulse shape and spectrum (from Gobl and Karlsson, 1991). The spectra are preemphasized by 6 dB/octave.

to three lowest harmonics, and one parameter influences the high-frequency content of the spectrum. Another vocal source parameter is the diplophonia parameter (Klatt and Klatt, 1990) with which creak, laryngalization, or diplophonia can be simulated. This parameter influences the function of the voiced source in such a way that every second pulse is lowered in amplitude and shifted in time.

The next generation of source models has to include adequate modeling of noise excitation in order to synthesize a natural change between voiced and unvoiced segments. The work of Rothenberg (1981) can serve as a guide for future implementations. In some earlier work at the Royal Institute of Technology (KTH), we were able to use a model that included a noise source (Rothenberg et al., 1975). High-quality synthesis of extralinguistic sounds such as laughter could be produced with this model in addition to reasonable voiced-unvoiced transitions.

The acoustic interactions between the glottal source and the vocal tract also must be considered (Bickley and Stevens, 1986). One of the major factors in this respect is the varying bandwidth of the formants. This is especially true for the first formant, which can be heavily damped during the open phase of the glottal source. However, it is not clear that such a variation can be perceived by a listener (Ananthapadmanabha et al., 1982). Listeners tend to be rather insensitive to bandwidth variation (Flanagan, 1972). In more complex models the output is glottal opening rather than glottal flow. The subglottal cavities can then be included in an articulatory model.

Noise sources have attracted much less research effort than the voiced source. However, some aspects have been discussed by Stevens (1971), Shadle (1985), and Badin and Fant (1989). Today, simple white noise typically is filtered by resonances that are stationary within each parameter frame. The new synthesizers do have some interaction between the voice source and the noise source, but the interaction is rather primitive. Transient sounds and aspiration dependent on vocal cord opening are still under development.

### **Formant-Based Terminal Analog**

The traditional text-to-speech systems uses a terminal analog based on formant filters. The vocal tract is simulated by a sequence of second-order filters in cascade while a parallel structure is used mostly for the synthesis of consonants. One important advantage of a cascade synthesizer is the automatic setting of formant amplitudes. The disadvantage is that it sometimes can be hard to do detailed spectral matching between natural and synthesized spectra because of the

simplified model. Parallel synthesizers such as that of Holmes (1983) do not have this limitation.

The Klatt model is widely used in research for both general synthesis purposes and perceptual experiments. A simplified version of this system is used in all commercial products that stem from synthesis work at the Massachusetts Institute of Technology (MIT): MITalk (Allen et al., 1987), DECTalk, and the system at Speech Technology Laboratory (Javkin et al., 1989). An improved version of the system has been commercialized as a research vehicle by Sensimetrics Corporation (Williams et al., 1992). Similar configurations were used in the ESPRIT/Polyglot project (Boves, 1991).

A formant terminal analog GLOVE (Carlson et al., 1991a), based on the OVE synthesizer (Liljencrants, 1968), has been developed at KTH and is used in current text-to-speech modeling (Carlson et al., 1982, 1991b). The main difference between these and the Klatt model is the manner in which consonants are modeled. In the OVE a fricative is filtered by a zero-pole-pole configuration rather than by a parallel system. The same is true for the nasal branch of the synthesizer.

New parameters have been added to the terminal analog model so that it is now possible to simulate most human voices and to replicate an utterance without noticeable quality reduction. However, it is interesting to note that some voices are easier to model than others. Despite the progress, speech quality is not natural enough in all applications of text to speech. The main reasons for the limited success in formant-based synthesis can be explained by incomplete phonetic knowledge. It should be noted that the transfer of knowledge from phonetics to speech technology has not been an easy process. Another reason is that the efforts using formant synthesis have not explored control methods other than the explicit rule-based description.

### **Higher-Level Parameters**

Since the control of a formant synthesizer can be a very complex task, some efforts have been made to help the developer. The "higher-level parameters" (Stevens and Bickley, 1991; Williams et al., 1992) explore an intermediate level that is more understandable from the developer's point of view compared to the detailed synthesizer specifications. The goal of this approach is to find a synthesis framework to simplify the process and to incorporate the constraints that are known to exist within the process. A formant frequency should not have to be adjusted specifically by the rule developer depending on nasality or glottal opening. This type of adjustment might be better

handled automatically according to a well-specified model. The same process should occur with other parameters such as bandwidths and glottal settings. The approach requires a detailed understanding of the relationship between acoustic and articulatory phonetics.

### Articulatory Models

An articulatory model will ultimately be the most interesting and flexible solution for the sound-generating part of text-to-speech systems. Development is also advancing in this area, but the lack of reliable articulatory data and appropriate control strategies still presents challenges. One possible solution that has attracted interest is to automatically train neural networks to control such a synthesizer. Rahim et al. (1993) and Bailly et al. (1991) have explored such methods.

Articulatory models, now under improvement, stem from basic work carried out at such laboratories as AT&T Bell Labs, MIT, and KTH. At each time interval, an approximation of the vocal tract is used either to calculate the corresponding transfer function or to directly filter a source waveform. Different vocal tract models have been used based on varying assumptions and simplifications. The models by Flanagan et al. (1975), Coker (1976), and Mermelstein (1973) have been studied by many researchers in their development of current articulatory synthesis.

The term "articulatory modeling" is often used rather loosely. Only part of the synthesis model is usually described in physical terms, while the remaining part is described in a simplified manner. Compare, for example, the difference between a tube model that models a static shape of the vocal tract with a dynamic physical model that actually describes how the articulators move. Thus, a complete articulatory model for speech synthesis has to include several transformations. The relationship between an articulatory gesture and a sequence of vocal tract shapes must be modeled. Each shape must be transformed into some kind of tube model with its acoustic characteristics. The acoustics of the vocal tract can then be modeled in terms of an electronic network. At this point, the developer can choose to use the network as such to filter the source signal. Alternatively, the acoustics of the network can be expressed in terms of resonances that can control a formant-based synthesizer. The main difference is the domain, time, or frequency in which the acoustics is simulated.

The developer has to choose at which level the controlling part of the synthesis system should connect to the synthesis model. All levels are possible, and many have been used. One of the pioneering efforts using articulatory synthesis as part of a text-to-speech system

was done by AT&T Bell Labs (Coker, 1976). Lip, jaw, and tongue positions were controlled by rule. The final synthesis step was done by a formant-based terminal analog. Current efforts at KTH by Lin and Fant (1992) use a parallel synthesizer with parameters derived from an articulatory model. In the development of articulatory modeling for text to speech, we can take advantage of parallel work on speech coding based on articulatory modeling (Sondhi and Schroeter, 1987). This work focuses not only on synthesizing speech but also on how to extract appropriate vocal tract configurations. Thus, it will also help us to get articulatory data through an analysis-synthesis procedure. This section has not dealt with the important work carried out to describe speech production in terms of physical models. The inclusion of such models still lies in the future, beyond the next generation of text to speech systems, but the results of these experiments will improve the current articulatory and terminal analog models.

## THE CONTROL PART

Models of segmental coarticulation and other phonetic factors are an important part of a text-to-speech system. The control part of a synthesis system calculates the parameter values at each time frame. Two main types of approaches can be distinguished: rule-based methods that use an explicit formulation of existing knowledge and library-based methods that replace rules by a collection of segment combinations. Clearly, each approach has its advantages. If the data are coded in terms of targets and slopes, we need methods to calculate the parameter tracks. The efforts of Holmes et al. (1964) and the filtered square wave approach by Liljencrants (1969) provide some classical examples in this context.

To illustrate the problem, I have chosen some recent work by Slater and Hawkins (1992). The work was motivated by the need to improve the rule system in a text-to-speech system for British English. Data for the second formant frequency at the onset of a vowel after a velar stop and at the midpoint in the vowel were analyzed, and, as expected, a clear correlation between the frequencies at these positions could be noted. The data could be described by one, two, or three regression lines, depending on the need for accuracy. This could then be modeled by a set of rules. As an alternative, all data points can be listed. Unfortunately, the regression lines change their coefficients depending on a number of factors such as position and stress. To increase the coverage, we need to expand the analysis window and include more dimensions or increase the number of units. Eventually, we will reach a point where the rules become too complex or

the data collection becomes too huge. This is the point where new dimensions such as articulatory parameters might be the ultimate solution.

### Concatenation of Units

One of the major problems in concatenative synthesis is to make the best selection of units and describe how to combine them. Two major factors create problems: distortion because of spectral discontinuity at the connecting points and distortion because of the limited size of the unit set. Systems using elements of different lengths depending on the target phoneme and its function have been explored by several research groups. In a paper by Olive (1990), a new method for concatenating "acoustic inventory elements" of different sizes is described. The system, developed at ATR, is also based on nonuniform units (Sagisaka et al., 1992).

Special methods to generate a unit inventory have been proposed by the research group at NTT in Japan (Hakoda et al., 1990; Nakajima and Hamada, 1988). The synthesis allophones are selected with the help of the context-oriented clustering (COC) method. The COC searches for the phoneme sequences of different sizes that best describe the phoneme realization.

The context-oriented clustering approach is a good illustration of a current trend in speech synthesis: automatic methods based on databases. The studies are concerned with much wider phonetic contexts than before. (It might be appropriate to remind the reader of similar trends in speech recognition.) One cannot take into account all possible coarticulation effects by simply increasing the number of units. At some point, the total number might be too high or some units might be based on very few observations. In this case a normalization of data might be a good solution before the actual unit is chosen. The system will become a rule-based system. However, the rules can be automatically trained from data in the same way as speech recognition (Philips et al., 1991).

### Rules and Notations

Development tools for text-to-speech systems have attracted considerable efforts. The publication of *The Sound Pattern of English* by Chomsky and Halle (1968) impelled a new kind of synthesis system based on rewrite rules. Their ideas inspired researchers to create special rule compilers for text-to-speech developments in the early 1970s. New software is still being developed according to this basic prin

ciple, but the implementations vary depending on the developer's tastes. It is important to note that crucial decisions often are hidden in the systems. The rules might operate rule by rule or segment by segment. Other important decisions are based on the following questions: How is the backtrack organized? Can nonlinear phonology be used (Pierrehumbert, 1987), as in the systems described by Hertz (Hertz, 1991; Hertz et al., 1985) and the Institute for Perception Research (Van Leeuwen and te Lindert, 1991, 1993)? Are the default values in the phoneme library primarily referred to by labels or features? These questions might seem trivial, but we see many examples of how the explicit design of a system influences the thinking of the researcher.

### Automatic Learning

Synthesis has traditionally been based on very labor-intensive optimization work. Until recently, the notion of analysis by synthesis had been explored mainly by manual comparisons between hand-tuned spectral slices and a reference spectrum. The work of Holmes and Pearce (1990) is a good example of how to speed up this process. With the help of a synthesis model, spectra are automatically matched against analyzed speech. Automatic techniques, such as this, will probably also play an important role in making speaker-dependent adjustments. One advantage of these methods is that the optimization is done in the same framework as that to be used in the production. The synthesizer constraints are thus already imposed in the initial state.

Methods for pitch-synchronous analysis will be of major importance in this context. Experiments such as the one presented by Talkin and Rowley (1990) will lead to better estimates of pitch and vocal tract shape. These automatic procedures will, in the future, make it possible to gather a large amount of data. Lack of glottal source data currently is a major obstacle for the development of speech synthesis with improved naturalness.

Given that we have a collection of parameter data from analyzed speech corpora, we are in a good position to look for coarticulation rules and context-dependent variations. The collection of speech corpora also facilitates the possibilities of testing duration and intonation models (Carlson and Granstrom, 1986; Kaiki et al., 1990; Riley, 1990; Van Santen and Olive, 1990).

## SPEAKING CHARACTERISTICS AND SPEAKING STYLES

Currently available text-to-speech systems are not characterized by a great amount of flexibility, especially not when it comes to variations in voice or speaking style. On the contrary, the emphasis has been on a neutral way of reading, modeled after the reading of nonrelated sentences. There is, however, a very practical need for different speaking styles in text-to-speech systems. Such systems are now used in a variety of applications, and many more are projected as the quality is improved. The range of applications demands a variation close to that found in human speakers. General use in reading stock quotations, weather reports, electronic mail, or warning messages are examples in which humans would choose rather different ways of reading. Apart from these practical needs in text-to-speech systems, there is the scientific interest in formulating our understanding of human speech variability in explicit models.

The current ambition in speech synthesis research is to model natural speech at a global level, allowing for changes of speaker characteristics and speaking style. One obvious reason is the limited success in enhancing the general speech quality by only improving the segmental models. The speaker-specific aspects are regarded as playing a very important role in the acceptability of synthetic speech. This is especially true when the systems are used to signal semantic and pragmatic knowledge.

One interesting effort to include speaker characteristics in a complex system has been reported by the ATR group in Japan. The basic concept is to preserve speaker characteristics in interpreting systems (Abe et al., 1990). The proposed voice conversion technique consists of two steps: mapping code book generation of LPC parameters and a conversion synthesis using the mapping code book. The effort has stimulated much discussion, especially considering the application as such. The method has been extended from a frame-by-frame transformation to a segment-by-segment transformation (Abe, 1991).

One concern with this type of effort is that the speaker characteristics are specified through training without a specific higher-level model of the speaker. It would be helpful if the speaker characteristics could be modeled by a limited number of parameters. Only a small number of sentences might in this case be needed to adjust the synthesis to one specific speaker. The needs in both speech synthesis and speech recognition are very similar in this respect.

A voice conversion system that combines the PSOLA technique for modifying prosody with a source-filter decomposition that enables spectral transformations has been proposed (Valbret et al., 1992).

Duration-dependent vowel reduction has been another topic of research in this area. It seems that vowel reduction as a function of speech tempo is a speaker-dependent factor (Van Son and Pols, 1989). Duration and intonation structures and pause insertion strategies reflecting variability in the dynamic speaking style are other important speaker-dependent factors. Parameters such as consonant-vowel ratio and source dynamics are typical parameters that must be considered in addition to basic physiological variations.

The differences between male and female speech have been studied by a few researchers (Karlsson, 1992; Klatt and Klatt, 1990). A few systems, such as that of Syrdal (1992), use a female voice as a reference speaker. The male voice differs from the female voice in many respects in addition to the physiological aspects. To a great extent, speaking habits are formed by the social environment, dialect region, sex, education, and by a communicative situation that may require formal or informal speech. A speaker's characteristics must be viewed as a complete description of the speaker in which all aspects are linked to each other in a unique framework (Cohen, 1989; Eskenazi and Lacheret-Dujour, 1991).

The ultimate test of our descriptions is our ability to successfully synthesize not only different voices and accents but also different speaking styles (Bladon et al., 1987). Appropriate modeling of these factors will increase both the naturalness and intelligibility of synthetic speech.

## MULTILINGUAL SYNTHESIS

Many societies in the world are increasingly multilingual. The situation in Europe is an especially striking example of this. Most of the population is in touch with more than one language. This is natural in multilingual societies such as Switzerland and Belgium. Most schools in Europe have foreign languages on their mandatory curriculum. With the opening of the borders in Europe, more and more people will be in direct contact with several languages on an almost daily basis. For this reason, text-to-speech devices, whether they are used professionally or not, ought to have a multilingual capability.

Based on this understanding, many synthesis efforts are multilingual in nature. The Polyglot project, supported by the European ESPRIT program, was a joint effort by several laboratories in several countries. The common software in this project was, to a great extent, language independent, and the language-specific features were specified by rules, lexica, and definitions rather than by the software itself. This is also the key to the multilingual effort at KTH. About one-

third of the systems delivered by the company INFOVOX are multilingual. The synthesis work pursued at companies such as ATR, CNET, DEC, and AT&T Bell Labs is also multilingual. It is interesting to see that the world's research community is rather small. Several of the efforts are joint ventures such as the CNET and CSTR British synthesis and the cooperation between Japanese (ATR) and U.S. partners. The Japanese company Matsushita even has a U.S. branch (STL) for its English effort, originally based on MITalk.

### **Speech Quality**

The ultimate goal for synthesis research, with few exceptions, is to produce the highest speech quality possible. The quality and the intelligibility of speech are usually very difficult to measure. No single test is able to pinpoint where the problems lie. The Department of Psychology at the University of Indiana started a new wave of innovation in evaluation of synthesis systems to which a number of groups have made subsequent substantial contributions. But we are still looking for a simple way to measure progress quickly and reliably as we continue development of speech synthesis systems. The recent work that has been done in the ESPRIT/SAM projects, the COCOSDA group, and special workshops will set new standards for the future.

### **CONCLUDING REMARKS**

In this paper a number of different synthesis methods and research goals to improve current text-to-speech systems have been touched on. It might be appropriate to remind the reader that nearly all methods are based on historical developments, where new knowledge has been added piece by piece to old knowledge rather than by a sudden change of approach. Perhaps the most dramatic change is in the field of synthesis tools rather than in the understanding of the "speech code." However, considerable progress can be seen in terms of improved speech synthesis quality. Today, speech synthesis is an appreciated facility even outside the research world, especially as applied to speaking aids for persons with disabilities. New synthesis techniques under development in speech research laboratories will play a key role in future man-machine interaction.

## ACKNOWLEDGMENTS

I would like to thank Bjorn Granstrom for valuable discussions during the preparation of this paper. This work has been supported by grants from the Swedish National Board for Technical Development.

## REFERENCES

- Abe, M. (1991), "A segment-based approach to voice conversion," Proc. ICASSP-91.
- Abe, M., K. Shikano, and H. Kuwabara (1990), "Voice conversion for an interpreting telephone," Proc. Speaker characterisation in speech technology, Edinburgh, UK.
- Allen, J., M. S. Hunnicutt, and D. Klatt (1987), "From Text to Speech." The MITalk System. Cambridge University Press, Cambridge, England.
- Ananthapadmanabha, T. V. (1984), "Acoustic analysis of voice source dynamics," STLQPSR 2-3/1984, pp. 1-24.
- Ananthapadmanabha, T. V., L. Nord, and G. Fant (1982), "Perceptual discriminability of nonexponential/exponential damping of the first formant of vowel sounds," in Proceedings of the Representation of Speech in the Peripheral Auditory System, Elsevier Biomedical Press, Amsterdam, pp. 217-222.
- Bardin, P., and G. Fant (1989), "Fricative modeling: Some essentials," Proceedings of the European Conference on Speech Technology.
- Bailly, G., R. Laboissiere, and J. L. Schwartz (1991), "Formant trajectories as audible gestures: An alternative for speech synthesis," J. Phon., 19(1).
- Bickley, C., and K. Stevens (1986), "Effects of the vocal tract constriction on the glottal source: Experimental and modeling studies," J. Phon., 14:373-382.
- Bladon, A., R. Carlson, B. Granstrom, S. Hunnicutt, and I. Karlsson (1987), "A text-to-speech system for British English, and issues of dialect and style," in Proceedings of the European Conference on Speech Technology, Edinburgh, Sept. 1987, vol. 1, ed. by J. Laver and M. A. Jack, pp. 55-58.
- Boves, L. (1991), "Considerations in the design of a multi-lingual text-to-speech system," J. Phone., 19:(1).
- Carlson, R., and B. Granstrom (1986), "A search for durational rules in a real-speech data base," Phonetica, 43:140-154.
- Carlson, R., B. Granstrom, and S. Hunnicutt (1982), "A multi-language text-to-speech module," Proc. ICASSP-82, vol. 3, Paris, pp. 1604-1607.
- Carlson, R., B. Granstrom, and I. Karlsson (1991) "Experiments with voice modeling in speech synthesis," Speech Commun., 10:481-489.
- Carlson, R., B. Granstrom, and S. Hunnicutt (1991b), "Multilingual text-to-speech development and applications," A. W. Ainsworth (ed.), in Advances in Speech, Hearing and Language Processing, JAI Press, London.
- Charpentier, F., and E. Moulines (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., 9(5/6):453-467.
- Chomsky, N., and M. Halle (1968), The Sound Pattern of English, Harper & Row, New York.
- Cohen, H. M. (1989), "Phonological Structures for Speech Recognition," Ph.D. thesis, Computer Science Division, University of California, Berkeley.
- Coker, C. H. (1976), "A model for articulatory dynamics and control," Proc. IEEE, 64:452-460.

- Eskenazi, M., and A. Lacheret-Dujour (1991), "Exploration of individual strategies in continuous speech," *Speech Commun.*, 10:249-264.
- Fant, G. (1960), *Acoustic Theory of Speech Production*, Mouton, The Hague.
- Fant, G., J. Liljencrants, and Q. Lin (1985), "A four parameter model of glottal flow," *Speech Transmission Laboratory Quarterly and Status Report*, No. 4.
- Flanagan, J. L. (1972), *Speech Analysis, Synthesis and Perception*, Springer Verlag, Berlin.
- Flanagan, J. L., K. Ishizaka, and K. L. Shipley (1975), "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Syst. Tech. J.*, 54:485-506.
- Furui, S. (1989), *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York.
- Gobl, C., and I. Karlsson (1991), "Male and Female Voice Source Dynamics," *Proceedings of the Vocal Fold Physiology Conference*, Gauffin and Hammarberg, eds. Singular Publishing Group, San Diego.
- Hakoda, K., S. Nakajima, T. Hirokawa, and H. Mizuno (1990), "A new Japanese text-to-speech synthesizer based on COC synthesis method," *Proc. ICSLP90*, Kobe, Japan.
- Hedelin, P. (1984), "A glottal LPC-vocoder," *Proc. IEEE*, San Diego, pp. 1.6.1-1.6.4.
- Hertz, S. R. (1991), "Streams, phones, and transitions: Toward a new phonological and phonetic model of formant timing," *J. Phon.*, 19(1).
- Hertz, S. R., J. Kadin, and K. J. Karplus (1985), "The Delta rule development system for speech synthesis from text." *Proc. IEEE*, 73(11).
- Holmes, J. N. (1973), "Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, AU-21:298305.
- Holmes, J. (1983), "Formant synthesizers, cascade or parallel," *Speech Commun.*, 2:251273.
- Holmes, J., I. G. Mattingly, and J. N. Shearne (1964), "Speech synthesis by rule," *Lang. Speech*, 7:127-143.
- Holmes, W. J., and D. J. B. Pearce (1990), "Automatic derivation of segment models for synthesis-by-rule." *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France.
- Javkin, H., et al. (1989), "A multi-lingual text-to-speech system," *Proc. ICASSP-89*.
- Kaiki, N., K. Takeda, and Y. Sagisaka (1990), "Statistical analysis for segmental duration rules in Japanese speech synthesis," *Proceedings of the International Conference on Spoken Language Processing*, Kobe, Japan.
- Karlsson, I. (1992), "Modeling speaking styles in female speech synthesis," *Speech Commun.*, 11:491-497.
- Klatt, D. K. (1976), "Structure of a phonological rule component for a synthesis-by-rule program," *IEEE Trans. ASSP-24*.
- Klatt, D. (1980), "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, 67:971-995.
- Klatt, D. K. (1987) "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, 82 (3):737-793.
- Klatt, D., and L. Klatt (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, 87:820-857.
- Liljencrants, J. (1968), "The OVE III speech synthesizer," *IEEE Trans. Audio Electroacoust.* 16 (1):137-140.
- Liljencrants, J. (1969), "Speech synthesizer control by smoothed step functions," *STLQPSR* 4/1969, pp. 43-50.
- Liljencrants, J. (1974), "Metoder fvr proportionell frekvenstransponering av en signal." Swedish patent number 362975.

- Lin, Q., and G. Fant (1992), "An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract," *Proc. ICASSP-92*.
- Mermelstein, P. (1973), "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, 53:1070-1082.
- Moulines, E., et al. (1990), "A real-time French text-to-speech system generating high quality synthetic speech," *Proc. ICASSP-90*.
- Nakajima, S., and H. Hamada (1988), "Automatic generation of synthesis units based on context oriented clustering," *Proc. ICASSP-88*.
- Olive, J. P. (1977), "Rule synthesis of speech from dyadic units," *Proc. ICASSP-77*, pp. 568-570.
- Olive, J. P. (1990), "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds," *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France.
- Olive, J. P., and M. Y. Liberman (1985), "Text-to-speech—an overview." *J. Acoust. Soc. Am. Suppl. 1*, 78(Fall):S6.
- Philips, M., J. Glass, and V. Zue (1991), "Automatic learning of lexical representations for sub-word unit based speech recognition systems," *Proceedings of the European Conference on Speech Communication and Technology*.
- Pierrehumbert, J. B. (1987), "The phonetics of English intonation," Bloomington, IULC.
- Rahim, M., C. Coodeyear, B. Kleijn, J. Schroeter, and M. Sondi (1993), "On the use of neural networks in articulatory speech synthesis," *J. Acoust. Soc. Am.*, 93(2):11091121.
- Riley, M. (1990), "Tree-based modeling for speech synthesis," *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France.
- Rosenberg, A. E. (1971), "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, 53:1632-1645.
- Rothenberg, M. (1981), "Acoustic interactions between the glottal source and the vocal tract," in *Vocal Fold Physiology*, ed. by K. M. Stevens and M. Hirano, University of Tokyo Press, Tokyo, pp. 303-323.
- Rothenberg, M., R. Carlson, B. Granstrom, and J. Lindqvist-Gauffin (1975), "A three-parameter voice source for speech synthesis," *Proceedings of the Speech Communication Seminar*, Stockholm, 1974, in *Speech Communication*, vol. 2, Almqvist and Wiksell, Stockholm, pp. 235-243.
- Roucos, S., and A. Wilgus (1985), "High quality time-scale modification for speech," *Proc. ICASSP-85*, pp. 493-496.
- Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," *Proc. ICASSP-88*.
- Sagisaka, Y. (1990), "Speech synthesis from text," *IEEE Commun. Mag.*, January.
- Sagisaka, Y., N. Kaikin, N. Iwahashi, and K. Mimura (1992), "ATR v-TALK speech synthesis system," *Proc. ICSLP-92*, Banff, Canada.
- Shadle, C. H. (1985), "The acoustics of fricative consonants," Ph.D. thesis, MIT, Cambridge, Mass.
- Slater, A., and S. Hawkins (1992), "Effects of stress and vowel context on velar stops in British English," *Proc. ICSLP-92*, Banff, Canada.
- Sondhi, M. M., and J. Schroeter (1987), "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. ASSP*, 35(7).
- Stevens, K. N. (1971), "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *J. Acoust. Soc. Am.*, 50(4):1180-1192.
- Stevens, K., and C. Bickley (1991), "Constraints among parameters simplify control of Klatt formant synthesizer," *J. Phonet.*, 19(1).

- Syrdal, A. K. (1992), "Development of a female voice for a concatenative synthesis text-to-speech system." *J. Acoust. Soc. Am.*, 92(Fall).
- Talkin, D., and M. Rowley (1990), "Pitch-synchronous analysis and synthesis for TTS systems," *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France.
- Valbret, H., E. Moulines, and J. P. Tubach (1992), "Voice transformation using PSOLA technique," *Proc. ICASSP-92*, San Francisco, pp. 1-145-1-148.
- Van Leeuwen, H. C., and E. te Lindert (1991), "Speechmaker, text-to-speech synthesis based on a multilevel, synchronized data structure," *Proc. ICASSP-91*.
- Van Leeuwen, H. C., and E. te Lindert (1993), "Speech maker: A flexible and general framework for text-to-speech synthesis, and its application to Dutch," *Comput. Speech Lang.*, 7 (2):149-168.
- Van Santen, J., and J. P. Olive (1990), "The analysis of segmental effect on segmental duration," *Comput. Speech Lang.*, No. 4.
- Van Son, R. J. J. H., and L. Pols (1989), "Comparing formant movements in fast and normal rate speech," *Proceedings of the European Conference on Speech Communication and Technology 89*.
- Williams, D., C. Bickley, and K. Stevens (1992), "Inventory of phonetic contrasts generated by high level control of a formant synthesizer," *Proc. ICSLP-92*, Banff, Canada, pp. 571-574.

# Linguistic Aspects of Speech Synthesis

*Jonathan Allen*

## SUMMARY

The conversion of text to speech is seen as an analysis of the input text to obtain a common underlying linguistic description, followed by a synthesis of the output speech waveform from this fundamental specification. Hence, the comprehensive linguistic structure serving as the substrate for an utterance must be discovered by analysis from the text. The pronunciation of individual words in unrestricted text is determined by morphological analysis or letter-to-sound conversion, followed by specification of the word-level stress contour. In addition, many text character strings, such as titles, numbers, and acronyms, are abbreviations for normal words, which must be derived. To further refine these pronunciations and to discover the prosodic structure of the utterance, word part of speech must be computed, followed by a phrase-level parsing. From this structure the prosodic structure of the utterance can be determined, which is needed in order to specify the durational framework and fundamental frequency contour of the utterance. In discourse contexts, several factors such as the specification of new and old information, contrast, and pronominal reference can be used to further modify the prosodic specification. When the prosodic correlates have been computed and the segmental sequence is assembled, a complete input suitable for speech synthesis has been determined. Lastly, multilingual systems utilizing rule frameworks are mentioned, and future directions are characterized.

## INTRODUCTION

To facilitate human-machine communication, there is an increasing need for computers to adapt to human users. This means of interaction should be pleasant, easy to learn, and reliable. Since some computer users cannot type or read, the fact that speech is universal in all cultures and is the common basis for linguistic expression means that it is especially well suited as the fabric for communication between humans and computer-based applications. Moreover, in an increasingly computerized society, speech provides a welcome humanizing influence. Dialogues between humans and computers require both the ability to recognize and understand utterances and the means to generate synthetic speech that is intelligible and natural to human listeners. In this paper the process of converting text to speech is considered as the means for converting text-based messages in computer-readable form to synthetic speech. Both text and speech are physically observable surface realizations of language, and many attempts have been made to perform text-to-speech conversion by simply recognizing letter strings that could then be mapped onto intervals of speech. Unfortunately, due to the distributed way in which linguistic information is encoded in speech, it has not been possible to establish a comprehensive system utilizing these correspondences. Instead, it has been necessary to first analyze the text into an underlying abstract linguistic structure that is common to both text and speech surface realizations. Once this structure is obtained, it can be used to drive the speech synthesis process in order to produce the desired output acoustic signal. Thus, text-to-speech conversion is an *analysis-synthesis system*. The analysis phase must detect and describe language patterns that are implicit in the input text and that are built from a set of abstract linguistic objects and a relational system among them. It is inherently linguistic in nature and provides the abstract basis for computing a speech waveform consistent with the constraints of the human vocal apparatus. The nature of this linguistic processing is the focus of this paper, together with its interface to the signal processing composition process that produces the desired speech waveform.

As in many systems, the complexity of the relationship between the text input and the speech output forces levels of intermediate representation. Thus, the overall conversion process is broken up through the utilization of two distinct hierarchies. One of these is *structural* in nature and is concerned with the means to compose bigger constructs from smaller ones (e.g., sentences are composed of words). The second hierarchy is an arrangement of different abstractions that pro

vide *qualitatively differing* constraint domains that interact to characterize all linguistic forms. These abstract domains include phonetics, phonology, the lexicon, morphology, syntax, semantics, acoustics, anatomy, physiology, and computation. In computing the overall text-to-speech process, these hierarchies are exploited to provide the environment for encoding relationships between linguistic entities. In this way, as the linguistic framework is built up, algorithms are utilized to produce additional facts, thus further extending the total characterization of the desired utterance. Thus, words can be "parsed" to discover their constituent morphemes, each of which corresponds to a lexical entry that provides both the phonological and the syntactic nature of the morpheme. The goal of the conversion process is to produce a comprehensive framework sufficient to allow the computation of the output speech waveform. Furthermore, we take as a working hypothesis the proposition that *every aspect of linguistic structure manifests itself in the acoustic waveform*. If this is true, the analysis part of the conversion process must provide a *completely specified* framework in order to ensure that the output speech waveform will be responsive to all linguistic aspects of the utterance.

Given the need to derive this structural framework, we can seek to understand the nature of the framework, how it is represented, how it can be discovered, and how it can be interpreted to produce synthetic speech. Answers to these questions are found from study of the various constraints on speech and language production, to which we turn now.

## CONSTRAINTS ON SPEECH PRODUCTION

For any text-to-speech system, the process by which the speech signal is generated is constrained by several factors. The *task* in which the system is used will constrain the number and kind of speech voices required (e.g., male, female, or child voices), the size and nature of the vocabulary and syntax to be used, and the message length needed. Thus, for restricted systems such as those that provide announcements of arrivals and departures at a railroad station, the messages are very short and require only limited vocabulary, syntax, and range of speaking style, so a relatively simple utterance composition system will suffice. In this paper, however, it is assumed that the vocabulary, syntax, and utterance length are *unrestricted* and that the system must strive to imitate a native speaker of the language reading aloud. For the *language* being used, the linguistic structure provides many constraining relationships on the speech signal. The phonetic repertoire of sounds; the structure of syllables, morphemes, words,

phrases, and sentences; the intended meaning and emphasis; and the interactive dialogue pattern restrict the class of possible linguistic structures. While many of the techniques described here are applicable to several languages, most of the results cited are for English. (Multilingual systems are described in a later section.) Of course, for all speakers, the *human vocal apparatus* limits the class of signals that can emanate from the lips and nose. The oral and nasal passages serve as a time-varying filter to acoustic disturbances that are excited either by the vocal cords or frication generated at some constriction in the vocal tract. All of these constraints, acting together, drive the speech generation process, and hence the text-to-speech process must algorithmically discover the overall ensemble of constraints to produce the synthetic speech waveform.

### WORD-LEVEL ANALYSIS

Early attempts to build text-to-speech systems sought to discover direct *letter-to-sound* relationships between letter strings and phoneme sequences (Venezky, 1970). Unfortunately, as noted above, a linguistic analysis is needed, and there is a consequent need for a constraining lexicon. This dictionary is used in several ways. Borrowed foreign words, such as "parfait" and "tortilla" retain their original pronunciation and do not follow the letter-to-sound rules of the language that imports them. Also, closed-class (function) words differ in pronunciation from open-class words. Thus, the letter "f" in "of" is pronounced with vocal cord voicing, whereas the "f" in open-class words such as "rooft" is unvoiced. Similarly, the "a" in "have" is pronounced differently than the "a" in "behave" and other open-class words. Hence, it makes sense to place these frequently occurring function words in a lexicon, since otherwise they will needlessly complicate any set of pronunciation rules. If the dictionary contains *morphs* (the surface textual realizations of abstract morphemes) rather than words, then algorithms can be introduced (Allen et al., 1987; Allen, 1992) to discover morph boundaries within words that delimit text letter strings that can be used to determine corresponding phoneme sequences. Thus, there are many pronunciations of the letter string "ea" as found in "reach," "tear," "steak," and "leather," but the "ea" in "changeable" is broken up by the internal morph boundary, and hence the "ea" is not functioning as a vowel digraph for purposes of pronunciation. Similarly, the "th" in "dither" is functioning as a consonant cluster, but in "hothouse" there is a morph boundary between the "t" and the "h," thus breaking up the cluster. For all of these reasons, a morph lexicon is both necessary and essential to

algorithms that determine the pronunciation of *any* English word. Furthermore, contemporary lexicons "cover" over 99 percent of all words and provide much more accurate pronunciations than letter-to-sound rules, which will be discussed later.

Given a morph lexicon, word-level linguistic analysis consists of finding the constituent morphemes (and morphs) of each word. These units have a number of valuable properties, in addition to those already noted above. Morphemes are the *atomic* minimal syntactic units of a language, and they are very *stable* in the language in that new morphemes are rarely introduced, and existing ones are rarely dropped from the language. These morphemes have large *generative power* to make words, so that a morph lexicon of given size can easily cover at least an order-of-magnitude larger number of words. Furthermore, as we have seen above, many language phenomena extend only within morph boundaries, and regularly inflected words (e.g., "entitled") and regular compound words (e.g., "snowplow") are readily *covered* by lexical morphemes.

The parsing of words to reveal their constituent morphemes (Allen, 1992; Allen et al., 1987) is an interesting recursive process that must recognize the mutating effects of vocalic suffixes. There are several such changes, such as consonant doubling to produce "fitted" from "fit + ed," the change of "y" to "i" in "cities" from "city + es," and restoration of the final silent "e" as in "choking" from "choke + ing." Note that in each of these cases the vocalic nature of the first letter of the suffix triggers the mutation that takes place during the morph composition process, and it is this change that must be undone in order to recognize the constituent lexical morphs in a word. In addition to the difficulties introduced by these mutations, it turns out that the parsing of words into morphs is ambiguous, so that, for example, "scarcity" can be covered by "scar + city," "scarce + ity," or "scar + cite + y." Fortunately, a simple test that prefers affixed forms over compounds can accurately pick the correct parse. It is interesting that these tests apply to the abstract morphemic structure of the parse, rather than the surface morph covering. For example, "teething" can be parsed into both "teethe + ing" and "teeth + ing," but in the latter analysis "teeth" is already an inflected form ("tooth" + PLURAL), and the parsing tests will prefer the simpler earlier analysis that contains only one inflection. Comprehensive experience with morphemic analysis, together with the systematic construction of large morph lexicons, have provided a robust basis for computing the pronunciation of individual words, and these techniques are now used in all high-performance text-to-speech systems.

## LETTER-TO-SOUND RULES

We have already noted the ability of morph covering analyses to cover over 99 percent of all words. Consequently, letter-to-sound analysis is attempted only when a morph covering is unavailable, since experience shows that phoneme strings obtained by letter-to-sound analysis are inferior to those found through morph analysis. Since letter-to-sound correspondences do not apply across morph boundaries, any word subjected to letter-to-sound analysis must have any detectable affixes stripped off, leaving a presumed root word for further analysis. Thus, the word "theatricality" is analyzed to "theatr + ic + al + ity." The string of three suffixes is tested for correctness by a compact categorical grammar. Thus, the terminal suffix "ity" produces nouns from adjectives, the medial suffix "al" produces adjectives from nouns or adjectives, and the initial suffix "ic" produces adjectives from nouns. In this way the suffixes are seen to be compatible in terms of their parts-of-speech properties, and hence the string of suffixes is accepted.

Once affixes have been stripped, the residual root is searched for recognizable letter strings that are present in known letter-string-to-phoneme-string correspondences. Consonant clusters are searched for first, since their pronunciation is more stable than vowel clusters, longest string first. Hence, the string "chr" will be found first in "Christmas," while the shorter string "ch" is found in "church." Vowel correspondences are least reliable and are established last in the overall process using both text and the computed phoneme environments. Vowel digraphs are the hardest strings to convert, and "ea" is subject to no fewer than 14 rule environments. Examples include "reach," "tear," "steak," "leather," and "theatricality." A complete algorithm has been described by Allen et al. (1987).

The advent of large lexicons in machine-readable form, together with modern computing platforms and searching algorithms, have led to sets of letter-to-sound rules that effectively complement the morphemic analysis procedures. Detailed informational analyses (Lucassen and Mercer, 1984) have been performed that permit the rational choice of rule correspondences, together with a quantitative assessment of the contribution of each letter or phoneme in the rule context to the accuracy of the rule. For specific applications, desired pronunciations of words, whether they would be analyzed by morph covering or letter-to-sound procedures, can be forced by the simple expedient of placing the entire word directly in the lexicon and hence treating it as an exception. A particularly difficult specific application is the pronunciation of surnames, as found in, say, the Manhat

tan telephone directory, where many names of foreign origin are found. In this case, etymology is first determined from spelling using trigram statistics (probability estimates of strings of three adjacent letters) (Church, 1986; Liberman and Church, 1992). Then specialized rules for each language can be utilized. Thus, the "ch" in "Achilles" is pronounced differently than the "ch" in "Church." It is interesting that the use of simple letter statistics, which reflect in part the phonotactics of the underlying language, can be combined with other constraints to produce good results on this exceedingly difficult task, where the frequency distribution of surnames is very different than for ordinary words.

### MORPHOPHONEMICS AND LEXICAL STRESS

When morph pronunciations are composed, adjustments take place at their boundaries. Thus, the PLURAL morpheme, normally expressed by the morph "s" or "es," takes on differing pronunciation based on the value of the voicing feature of the root word to which the suffix attaches. Hence, the plural of "horse" requires that a short neutral vowel be inserted between the end of the root and the phonemic realization of PLURAL, else the plural form would only lengthen the terminal /s/ in "horse." On the other hand, if the root word does not end in an s-like phoneme, pronunciation of the plural form has the place and fricative consonant features of /s/ but follows the voicing of the root. Since "dog" ends in a voiced stop consonant, its plural suffix is realized as a /z/, while for the root "cat," terminated by an unvoiced stop consonant, the plural is realized as the unvoiced fricative /s/. A similar analysis applies to the computation of the pronunciation of the morpheme affix PAST, as in "pasted," "bagged," and "plucked." There are additional morphophonemic rules used in text-to-speech systems (Allen et al., 1987), and they are all highly regular and productive. Without their use, the lexicon would be needlessly enlarged.

One of the major achievements of modern linguistics is the understanding of the lexical stress system of English (Chomsky and Halle, 1968). Prior to the mid-1950s, the stress contours of words were specified by long word lists of similar stress pattern, and those learning English as a second language were expected to assimilate these lists. Over the past 40 years, however, comprehensive rules have been derived whose application computes the surface stress contour of words from an underlying phonological specification. These rules are complex in nature, and apply not only to monomorphemic roots, but also to affixed words and compounds. This elegant theory is

remarkably robust and has been extensively tested over large lexicons. The trio of words "system, systematic, systematize" illustrates the substantial stress shifts that can take place not only in the location of stress (first syllable in "system," third syllable in "systematic") but also in stress value (the vowel corresponding to the letter "a" is fully realized and stressed in "systematic" but reduced to a neutral vowel in "systematize"). Furthermore, it becomes clear that the lexicon must contain the nonreduced forms of vowels, since otherwise the correct pronunciation of affixed words may not be computable. Thus, the second vowel of "human" is reduced, but its underlying nonreduced form must be known in the lexicon, since in "humanity" the vowel is not reduced. Some suffixes are never reduced, as "eer" in "engineer," and always receive main stress. The rules for compounds, such as "snowplow" are simple, placing stress on the first morph, but the detection of long multiword compounds is difficult (Sproat, 1990; Sproat and Liberman, 1987) (see the section titled "Prosodic Marking") and typically depends on heuristic principles. As an example, in "white house" the construction is attributive, and stress is placed on the head of the phrase, "house." But in the textually similar phrase "White House," the capital letters serve to denote a specific house, namely the residence of the President of the United States, and hence the phrase denotes a compound noun, with stress on the first word.

After morphophonemic and lexical stress rules are applied, a number of phonological adjustments are made, based on articulatory smoothing. Alveolar (dental ridge) flapped consonants are produced as rapid stops in words such as "butter," and two voiceless stop consonants can assimilate, as in "Pat came home," where the "t" is assimilated into the initial consonant of the word "came." Sometimes the effect of articulatory smoothing must be resisted in the interest of intelligibility, as in the insertion of a glottal stop between the two words "he eats." Without this hiatus mechanism, the two words would run on into one, possibly producing "heats" instead of the desired sequence.

## ORTHOGRAPHIC CONVENTIONS

Abbreviations and symbols must be converted to normal words in order for a text-to-speech system to properly pronounce all of the letter strings in a sample of text. While the pronunciation of these is largely a matter of convention, and hence not of immediate linguistic interest, some linguistic analysis is often necessary to pick the appropriate pronunciation when there are ambiguous interpretations. The symbol "I" can be used to designate a pronoun, a letter name, or a

Roman numeral, and "Dr." can stand for "Doctor" or "Drive." The string "2/3" can indicate the fraction "two-thirds," "February third," or the phrasal string "two slash three." Numbers and currency pose several problems of interpretation, so that "3.45" can be read "three point four five" or "three dollars and forty-five cents." While these ambiguities can often be resolved by heuristic contextual analysis (including syntactic constraints), some conventions are applied inconsistently. In the analysis of one corpus, the string "I.R.S." appeared (with periods) 22 times, whereas "IRS" appeared 428 times. Lest this pattern seem to predict a rule, "N.Y." was found 209 times, whereas "NY" occurred only 14 times! Recently, comprehensive statistical analyses of large corpora have been completed (Liberman and Church, 1992), and decision trees (Brieman et al., 1984) have been constructed automatically from a body of classified examples, once a set of features has been specified. As a result, the quality of conversions of abbreviations to phonemic representation has improved markedly, demonstrating the power of statistical classification and regression analysis.

### PART-OF-SPEECH ASSIGNMENT

Much of the linguistic analysis used by text-to-speech systems is done at the word level, as discussed above. But there are many important phonological processes that span multiple word phrases and sentences and even paragraph level or discourse domains. The simplest of these constraints is due to syntactic part of speech. Many words vary with their functioning part of speech, such as "wind, read, use, invalid, and survey." Thus, among these, "use" can be a noun or verb and changes its pronunciation accordingly, and "invalid" can be either a noun or an adjective, where the location of main stress indicates the part of speech. At the single-word level, suffixes have considerable constraining power to predict part of speech, so that "dom" produces nouns, as in "kingdom," and "ness" produces nouns, as in "kindness." But in English, a final "s," functioning as an affix, can form a plural noun or a third-person present-tense singular verb, and every common noun can be used as a verb. To disambiguate these situations and reliably compute the functioning part of speech, a dynamic programming algorithm has been devised (Church, 1988; DeRose, 1988; Jelinek, 1990; Kupiec, 1992) that assigns parts of speech with very high-accuracy. Once again, this algorithm relies on a statistical study of a tagged (marked for part-of-speech) corpus and demonstrates the remarkable capabilities of modern statistical techniques.

## PARSING

In addition to determining the functioning part of speech for individual words, modern text-to-speech systems also perform some form of limited syntactic analysis, or parsing. These analyses can be used in many ways. As has already been demonstrated, individual word pronunciations can vary with part of speech. In addition, a parsing analysis can provide the structural basis for the marking of prosodic (or suprasegmental) features such as prominence, juncture, and sentence type (declarative, question, or imperative). The accurate calculation of segment durations and pitch contours requires such prosodic marking based on at least minimal syntactic information, or else the resulting speech will be flat, hard to listen to, and even lacking in intelligibility.

Since the justification for parsing is to help provide the structural basis for intelligible and natural-sounding synthetic speech, it has long been assumed that there is a direct relationship between syntactic structure and prosodic structure (the way in which speakers naturally group words). Over the past decade, however, this view has been increasingly challenged, and many phonologists now believe that there are substantial differences between the two structures (Selkirk, 1984). Nevertheless, the local phrase-level parsing used by contemporary systems provides an initial structure that is very useful, even though it may later be modified to provide the substrate for prosodic marking (next section). An even stronger view would claim that what is desired is the relationship between pragmatic and semantic structure and sound and that any correspondence between syntax and intonation is largely the by-product of the relations between syntax and intonation, on the one hand, and the higher-level constraints of semantics and pragmatics, on the other hand (Monaghan, 1989). Nevertheless, despite this possibility, phrase-level parsing must for the present provide the needed structural basis given the lack of such higher-level constraints when the system input consists of text alone. When the input is obtained from a message-producing system, where semantic and pragmatic considerations guide the message composition process, alternative prosodic structures may be available for the determination of prosodic correlates (Young and Fallside, 1979).

Unfortunately, full clause-level parsing of unrestricted text is an unsolved problem. Nevertheless, phrase-level parsing is fast and reliable and avoids the substantial complexities of clause-level analysis. The main reason for the success of phrase-level analysis is the high syntactic constraining power of determiner sequences in noun phrases and auxiliary sequences in verb phrases. Many text-to-speech sys

tems provide rapid and accurate phrase-level parsing (Allen et al., 1987; Church, 1988) that provides a sufficient base for the instantiation of prosodic cues. Thus, in the classic sentence "He saw the man in the park with the telescope," where determination of the attachment of the prepositional phrases is several-ways ambiguous, the pronunciation (including prosodics) can be derived from the unambiguous phrasal analysis, without the need for resolving the clause-level ambiguity. Of course, clause-level parsing can often be exploited when available, so that a parser for such structures would be useful, providing it failed gracefully to the phrase-level when an unambiguous clause-level analysis could not be obtained. Even if such a comprehensive parser were available, however, many researchers do not believe that its benefits outweigh its cost in terms of both computational expense and necessity as input for prosodic algorithms (Bachenko and Fitzpatrick, 1990), so there is little motivation to extend the scope of syntactic analysis to the clause- level.

### PROSODIC MARKING

Once a syntactic analysis is determined, it remains to mark the text for prosodic features. These include mainly intonation, prominence, juncture, and sentence type. Speech synthesis procedures can then interpret the segmental phonetic content of the utterance, along with these prosodic markers, to produce the timing and pitch framework of the utterance, together with the detailed segmental synthesis. Many linguistic effects contribute to the determination of these prosodic features. At the lexical level, some words are inherently stressed. For example, in "Hillary might not make cookies for me," the past tense modal auxiliary "might" and the negative "not" express doubt and negation and are reliably stressed, so they are marked for prominence (O'Shaughnessy and Allen, 1983). Pronominal reference can also be designated prosodically. Thus, in "She slapped him in the face and then she hit the man," if "him" and "the man" are coreferential, then "hit" receives prominence and "the man" is reduced. But if "him" and "the man" refer to distinct individuals, then "man" is prominent and "hit" is reduced. Correct determination of pronominal reference is not available from simple phrase-level syntactic analysis and must rely on larger scope discourse analysis, which is beginning to be used in text-to-speech systems, but the existence of these phenomena shows the need for, and utility of, such structural information.

As noted in the previous section on parsing, there has been an increasing emphasis during the past decade on prosodic structure

(the natural grouping of words in an utterance) as distinct from syntactic structure. The relationship between these two structures was examined linguistically in Selkirk (1984), and an emphasis on "performance structures" as natural groupings was presented in Gee and Grosjean (1983). These studies emphasized that performance structures have relatively small basic units, a natural hierarchy, and that the resulting overall structure was more balanced than that provided by syntactic constituent analysis. The "performance" aspect of these analyses utilized subjective appraisal of junctural breaks and discovered that word length and the syntactic label of structural nodes played an important role. These natural groupings were found to provide a flatter hierarchical structure than that provided by a syntactic analysis, so that a long verb phrase, "has been avidly reading about the latest rumors in Argentina," which would result in a hierarchy of seven levels in a typical syntactic analysis, would be grouped into three performance chunks—"has been avidly reading" "about the latest rumors" "in Argentina"—which utilize only four levels of hierarchy. The smallest chunks encode what is probably the smallest bundle of coherent semantic information, as suggested by a "case" type of analysis, which is usually associated with more inflected languages than English. That is, the elements of these chunks form a tightly bound package of conceptual information that might be lexicalized into a single lexical item in another language. These chunks are centered on noun or verb heads, such as "in the house" and "will have been reading it," where there are negligible prosodic breaks between the words.

Although the performance structure analysis presented by Gee and Grosjean (1983) presupposed a complete syntactic analysis in order to determine the performance structure, Bachenko and Fitzpatrick (1990) rejected the need for clausal structure and predicate-argument relations. Furthermore, "readjustment rules" that had been proposed to convert the syntactic structure to that needed for prosodics, were abandoned, and an algorithm was provided to generate prosodic phrases that was claimed to be "discourse neutral," with only 14 percent of the phrases studied discourse determined. In this approach there was no need to recognize verb phrase and sentential constituents: only noun phrases, prepositional phrases, and adjectival phrases "count" in the derivation of the prosodic chunks. This was an extremely encouraging result, since a phrase-level parser, of the type described in the previous section, can provide the needed units. Constituency, adjacency, and length were found to be the main factors determining (discourse-neutral) prosodic phrasing. Of course, these prosodic boundaries can be shifted by discourse phrasing, occasioned by emphasis,

contrast, parallelism, and coreference, but the phrasing required for a neutral reading can be directly obtained using these phrasal analyses.

Following the analysis introduced by Bachenko and Fitzpatrick (1990), there has been much interest in automatically computing prosodic phrase boundaries. Rather than formulating these techniques in terms of rules, statistical techniques have been exploited. Wang and Hirschberg (1992) used classification and regression tree (CART) techniques (Brieman et al., 1984) to combine many factors that can affect the determination of these boundaries. In Ostendorf and Veilleux (1993) a hierarchical stochastic model of prosodic phrases is introduced and trained using "break index" data. For predicting prosodic phrase breaks from text, a dynamic programming algorithm is provided for finding the maximum probability prosodic parse. These recent studies are very encouraging, as they provide promising techniques for obtaining the prosodic phrasing of sentences based on input text. The evolution of these techniques, from initial linguistic investigations, through psycholinguistic experiments, to the present computational linguistics studies, is extremely interesting, and the interested reader can gain much insight and understanding from the trajectory of references cited. A useful summary is also provided by Wightman et al. (1992).

Multiword compounds are often hard to analyze, but their perception is highly facilitated with proper prosodic marking. "Government tobacco price support system" and "power generating station control room complex" are two examples of long compounds in need of prosodic cues to reveal their structure. Many of these examples appear to require semantic analysis that is not available, but surprising improvements have been obtained through careful study of many examples (Sproat, 1990; Sproat and Liberman, 1987). In addition to use of the compound stress rule, which places stress for a two-constituent compound (e.g., "sticky bun") on the left, words can be lexically typed in a way that facilitates prediction of stress. Thus measure words (e.g., "pint," "dollar") can combine with a phrase on their right to form a larger phrase that normally takes stress on the right element, as in "dollar bill" and "pint jug." While these rules are useful, for large compound nominals, further heuristics must be applied in addition to the recursive use of the simple compound rule. A rhythm rule can be used to prevent clashes between strong stresses. In this way the stress on "Hall" in "City Hall parking lot" is reduced and that of "City" is raised, so that while "parking" retains the main stress, the next largest stress is two words away, "City." An interesting example of the power of statistics is the use of mutual informa

tion (Sproat, 1990) to resolve possible ambiguous parsings. For example, "Wall Street Journal" could be parsed as either ([Wall Street] Journal), or (Wall [Street Journal]), where the latter parse would incorrectly imply main stress on "Street." But in a corpus derived from the Associated Press Newswire for 1988, "Wall Street" occurs 636 times outside the context of "Wall Street Journal," whereas "Street Journal" occurs only five times outside this context, and hence the mutual information measure will favor the first (correct) parse and corresponding main stress on "Journal," with "Wall Street" treated as a normal two-word compound.

Of course, virtually any word in a sentence can be emphasized, and if this is marked in the text by underlining or italics, then prominence can be provided for that word.

Lastly, junctural cues are an important aid to perception. In "The dog Bill bought bit him," the reduced relative clause is not explicitly marked, so that a junctural pause after "bought" can indicate to the listener the end of this embedded clause. It has recently been shown (Price et al., 1991) that, for a variety of syntactic classes, naive listeners can reliably separate meanings on the basis of differences in prosodic information. These results were obtained from listener judgments of read speech where the ambiguous material was embedded in a larger context. For example, the sentence "They rose early in May." can be used in the following two ways:

- "In spring there was always more work to do on the farm. May was the hardest month. *They rose early in May.*"
- "Bears sleep all winter long, usually coming out of hibernation in late April, but this year they were a little slow. *They rose early in May.*"

The fact that listeners can often successfully disambiguate sentences from prosodic cues can be used to build algorithms to pick one of several possible parses based on these cues (Ostendorf et al., 1993). Using the notion of "break index" (a measure of the junctural separation between two neighboring words) introduced by Price et al. (1991) and statistical training procedures, the candidate parsed text versions are analyzed in terms of these break indices automatically in order to synthesize predicted prosodic structures, which are then compared with the analyzed prosodic structure obtained from the spoken utterance. This is a good example of analysis-by-synthesis processing, where the correct structural version of an utterance is found by synthesizing all possible versions prosodically (at an abstract level of prosodic structure using break indices) and then comparing them with the

prosodically analyzed spoken version. While prosodic correlates (e.g., pitch and durations) can rarely be used directly in a bottom-up manner to infer structure, analysis-by-synthesis techniques utilize a scoring of top-down-generated structures to determine by verification the most likely parse.

Standards for the prosodic marking of speech are currently being developed, together with text-based algorithms to create this encoding. Once a large corpus of text is analyzed in this way, and compared with manually provided markings, a rich new enhancement to the overall framework of linguistic analysis will be available, contributing greatly to increased naturalness and intelligibility of synthetic speech.

In this section, emphasis has been placed on determination of prosodic structure, including prosodic boundaries. Once this structure is available, prosodic correlates must be specified. These include durations and the overall timing framework, and the fundamental frequency contour reflecting the overall intonation contour and local pitch accents to mark stress. Not surprisingly, statistical techniques have been developed for the fitting of segmental durations within syllables and higher-level units (Campbell, 1992; Riley, 1992; Van Santen, 1992). The placement of pitch accent has also been determined by use of classification and regression tree analysis, basing the result on factors such as part of speech of the word and its adjacent words and its position in a larger prosodic constituent (Ross et al., 1992). These techniques are also able to introduce shifting of accent to avoid rhythmic clash with a stressed syllable in the next word. Once the overall intonational contour and pitch accent determination is made, the corresponding specification can be used as input to an algorithm (Pierrehumbert, 1981), which will generate the needed fundamental frequency contour.

## DISCOURSE-LEVEL EFFECTS

Beyond the sentence level, there are numerous attributes of the overall discourse (Grosz et al., 1989; Grosz and Sidner, 1986) that influence the prosodic structure of the extended utterance. Since a topic is usually established in a discourse and then comments are made about the topic, facts that were previously established are given reduced prominence when later repeated in the discourse. In the question-answer sequence "What did Joe buy at the mall? Joe bought a boom box at the mall," only "boom box" will receive prominence in the second sentence, since all other syntactic arguments for the verb "buy" have been established in the previous sentence. This phenomenon is

called "new/old information," constraining new information to receive prominence, and old information to be reduced. Determination of what is new and what is old is by no means simple, but a variety of counting techniques have been introduced to heuristically estimate the occurrence of old information, all other terms assumed to be new.

There are a variety of focus-shifting transformations available in English to lead the listener to the intended focus of the sentence or perhaps to distract the listener from the normal focus of the sentence. The passive transformation is probably the most frequently occurring example of this effect. Thus, "John bought the books" can be passivized to "The books were bought by John," which can optionally have the agent (John) deleted to form "The books were bought." In this way the initial focus on "John" is first shifted to "the books," and then "John" disappears altogether. In "The likelihood of a tax on the middle class is small, the President thinks," the agent (the President) has been moved to the end of the sentence, with reduced prominence, hence removing the focus of the sentence from him. Such transformations are frequently used to achieve the desired focus, and it is important to mark the sentential prominences accordingly.

Pragmatic knowledge of the world can provide a bias that sometimes overwhelms other constraints. Thus, "He hit the man with the book" is ambiguous. Either "He" hit "the man with the book," or "He hit the man" "with the book." The plausibility of each interpretation can often be inferred from the discourse context and the prosodic structure appropriately marked. It is probably of some help to text-to-speech systems that the reading with the largest pragmatic bias is likely to be perceived, even if the prosodic correlates mark the alternate reading (Wales and Toner, 1979). This effect also indicates that a pragmatically rare interpretation (and hence one with substantial new information) must be strongly marked prosodically in order for the intended reading to be perceived.

Discourse-level context may also facilitate the prosodic marking of complex nominals, designate prepositional phrase attachment, and disambiguate conjoined sentences. Thus, in "The bright students and their mentors . . .," "bright" can modify just the "students" or both the "students and their mentors." While there can be no doubt that marking of intended syntactic structure is useful for perception, many of these constructions are inherently ambiguous, and no general techniques are available for the exploitation of discourse structure for purposes of disambiguation. Indeed, relatively little is known concerning discourse structure and how it can be discovered, given only text as input. On the other hand, when synthesis is performed from

an abstract message concept (Young and Fallside, 1979), rather than only the resultant surface text, discourse structure may be readily available at the abstract level, and hence directly utilized by prosodic marking procedures.

As the development of discourse theory grows, a number of investigators are creating algorithms for the control of prosodics based on discourse constraints. Within a computer-aided instruction application, discourse effects on phrasing, pitch range, accent location, and tune have been demonstrated by Hirschberg and Pierrehumbert (1986). In Hirschberg (1992), limited discourse-level information, including given/new distinctions, and some information on focus, topic, and contrast, together with refined parts-of-speech distinctions, have been used to assign intonational features for unrestricted text. Discourse connectivity is often signaled with cue phrases, such as "but," "now," "by the way," and "in any case," and their relation to intonation has been described by Hirschberg and Litman (1987). For applications where a message is composed using a discourse formalism, new algorithms are likely that will provide more natural synthetic speech, but when unrestricted text is the input and the discourse domain is similarly unrestricted, the determination of contrast, coreference, and new/old information is very difficult, making incorporation of corresponding intonational effects unlikely. Nevertheless, as discourse theory evolves, its relation to prosody will be established, even if it remains difficult to determine the discourse structure from the input provided. Furthermore, a well-developed theory will facilitate input analysis, and the design of applications that can take advantage of the discourse/prosody mappings that become understood. Although much remains to be discovered, the relation of meaning to intonational contours in discourse (Hobbs, 1990) is of great importance, and the prospect of a system where specific facets of discourse meaning can be manipulated prosodically is indeed very exciting.

## MULTILINGUAL SYNTHESIS

Several groups have developed integrated rule frameworks and languages for their design and manipulation (Carlson and Granstrom, 1986; Hertz, 1990; Van Leeuwen and te Lindert, 1993). Using these structures, a flexible formalism is available for expressing rules and for utilizing these rules to "fill in" the coordinated comprehensive linguistic description of an utterance. The complex data structures provided in these systems also facilitate the alignment of constraints across several domains (or levels of representation), such as a textual character string, the names of words, their constituent morphs and

phonemes, and the overall syntactic structure. These unified procedures are a considerable improvement over isolated ad hoc rule systems that apply at only one level of linguistic representation. Furthermore, they facilitate the writing of new rules and experimentation with an overall integrated rule system. Thus, it is no surprise that these groups have built several text-to-speech systems for many different languages. Although good-quality text-to-speech systems have resulted from the exploitation of these frameworks, single-language ad hoc systems currently provide better-quality speech, but this state of affairs probably reflects mostly the amount of time spent on refinement of the rules, rather than any intrinsic limitation of the coordinated framework and rule approach.

## THE FUTURE

Contemporary text-to-speech systems are available commercially and are certainly acceptable in many applications. There is, however, both much room for improvement and the need for enhancements to increase the intelligibility, naturalness, and ease of listening for the resultant synthetic speech. In recent years much progress has followed from the massive analysis of data from large corpora. Modern classification and decision tree techniques (Brieman et al., 1984) have produced remarkable results where no linguistic theory was available as a basis for rules. In general, the use of standard algorithmic procedures, together with statistical parameter fitting, has been very successful. To further this process, large tagged databases are needed, using standard techniques that can be employed by many diverse investigators. Such databases are just beginning to be developed for prosodic phenomena (Silverman et al., 1992), but they can also be extremely useful for enhancing naturalness at the segmental level. While these statistical techniques can often extract a great deal of useful information from both texts and tagged phonetic transcriptions, the quest for appropriate linguistic models must be aggressively extended at all levels of representation. Where good models are available, such as for morphemic structure and lexical stress, the results are exceedingly robust. Linguistic descriptions of discourse are much needed, and a more detailed and principled prosodic theory that could guide both analysis and synthesis algorithms would be exceedingly useful. Of course, for some tasks, such as the conversion of abbreviations and standard symbols, there is relatively little linguistic content, and statistical techniques will have to bear the brunt of the task.

The use of articulation as a basis for phonology (Browman and

Goldstein, 1989) and synthesis may provide a fundamental solution to many problems of speech naturalness and may also introduce useful constraints for speech recognition. Many facts of speech production are best represented at the articulatory level, and a rule system focused on articulatory gestures is likely to be simpler than the current rule systems based on acoustic phonetics. Unfortunately, the acquisition of articulatory information is exceedingly difficult, since it involves careful observation of the entire set of speech articulators, many of which are either hidden from normal view or are difficult to observe without perturbing the normal speech production process. Nevertheless, improvements in the understanding and representation of articulation cannot help but improve synthesis, and it is important that the research community make a long-term commitment to the acquisition of this knowledge.

Lastly, contemporary research is benefiting from quickly evolving computational and experimental technology, which will provide the substrate for many new studies, as well as cost-effective systems for many applications. These facilities allow an attack on text and speech analysis at a level of complexity that was not hitherto possible. Future research will utilize statistical discovery procedures to suggest new linguistic formalisms and to organize observations of very large corpora. It is clear that the text-to-speech research community is now positioned to make large improvements in speech quality over extensive texts and also to contribute directly to the overall base of knowledge in linguistics, computational linguistics, phonetics, and articulation.

## REFERENCES

- Allen, J. (1992) "Overview of Text-to-Speech Systems," in S. Furui, and M. Sondhi, eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York. pp. 741-790.
- Allen, J., M. S. Hunnicutt, and D. H. Klatt (1987), *From Text to Speech: The MITalk System*, Cambridge University Press, London.
- Bachenko, J., and E. Fitzpatrick (1990), "A Computational Grammar of Discourse-Neutral Prosodic Phrasing in English," *Comput. Linguist.*, 16:155-170.
- Brieman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, Calif.
- Browman, C. P., and L. Goldstein (1989), "Articulatory Gestures as Phonological Units," *Phonology*, 6(2):201.
- Campbell, W. N. (1992), "Syllable-Based Segmental Duration," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, eds, Elsevier, New York, pp. 211-224.
- Carlson, R., and B. Granstrom (1986), "Linguistic Processing in the KTH Multilingual text-to-speech system" in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2403-2406.

- Chomsky, A. N., and M. Halle (1968), *Sound Pattern of English*, Harper & Row, New York.
- Church, K. W. (1986), "Stress Assignment in Letter to Sound Rules for Speech Synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2423-2426.
- Church, K. W. (1988), "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," in *Proceedings of the 2nd Conference on Applied Natural Language Processing*, Austin, Texas, pp. 136-143.
- Coker, C. H., K. W. Church, and M. Y. Liberman (1990), "Morphology and Rhyming: Two Powerful Alternatives to Letter-to-Sound Rules for Speech Synthesis," in *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 83-86.
- DeRose, S. (1988), "Grammatical Category Disambiguation by Statistical Optimization," *Comput. Linguist.*, 14(1).
- Gee, J. P., and F. Grosjean (1983), "Performance Structures: A Psycholinguistic and Linguistic Appraisal," *Cognit. Psychol.*, 15:411-458.
- Grosz, B. J., and C. L. Sidner (1986), "Attention, Intentions, and the Structure of Discourse," *Comput. Linguist.*, 12(3):175-204.
- Grosz, B. J., M. E. Pollack, and C. L. Sidner (1989), "Discourse," Chapter 11 in *Foundations of Cognitive Science*, M. Posner, ed., MIT Press, Cambridge, Mass.
- Hertz, S. R. (1990), "A Modular Approach to Multi-Dialect and Multi-Language Speech Synthesis Using the Delta System," in *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 225-228.
- Hirschberg, J. (1992), "Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, eds., Elsevier, New York, pp. 367-376.
- Hirschberg, J., and D. Litman (1987), "Now Let's Talk About Now: Identifying Cue Phrases Intonationally," in *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 163-171.
- Hirschberg, J., and J. B. Pierrehumbert (1986), "The Intonational Structure of Discourse," in *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 136-144.
- Hobbs, J. R. (1990), "The Pierrehumbert-Hirschberg Theory of Intonational Meaning Made Simple: Comments on Pierrehumbert and Hirschberg," in *Plans and Intentions in Communication and Discourse*, P. R. Cohen, J. Morgan, and M. E. Pollack, eds., MIT Press, Cambridge, Mass.
- Jelinek, F. (1990), "Self-Organized Language Modeling for Speech Recognition," in *Readings in Speech Recognition*, A. Waibel, and K. Lee, eds., Morgan Kaufmann, San Mateo, Calif.
- Klatt, D. (1975), "Vowel Lengthening Is Syntactically Determined in a Connected Discourse," *J. Phon.*, 3:129-140.
- Klatt, D. H. (1987), "Review of Text-to-Speech Conversion for English," *J. Acoust. Soc. Am.*, 82 (3):737.
- Kupiec, J. (1992), "Robust Part-of-Speech Tagging Using a Hidden Markov Model," *Comput. Speech Lang.*, 6:225-242.
- Liberman, M. Y., and K. W. Church (1992), "Text Analysis and Word Pronunciation in Text-to-Speech Synthesis," Chapter 24 in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi eds., Marcel Dekker, New York, pp. 791-831.
- Lucassen, J. M., and R. L. Mercer (1984), "An Information Theoretic Approach to the Automatic Determination of Phonemic Base Forms," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 42.5.1-42.5.4.
- Monaghan, A. I. C. (1989), "Using Anomalies for Intonation in Text-to-Speech: Some

- Proposals," University of Edinburgh Department of Linguistics, Work in Progress 22.
- O'Shaughnessy, D., and J. Allen (1983), "Linguistic Modality Effects on Fundamental Frequency in Speech," *J. Acoust. Soc. Am.*, 74(4):1155-1171.
- Ostendorf, M., and N. M. Veilleux (1993), "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *Comput. Linguist.*, 19.
- Ostendorf, M., C. W. Wightman, and N. M. Veilleux (1993), "Parse Scoring with Prosodic Information: An Analysis/Synthesis Approach," *Comput. Speech Lang.*, 7:193210.
- Pierrehumbert, J. B. (1981), "Synthesizing Intonation," *J. Acoust. Soc. Am.*, 70:985-995.
- Price, P. J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong (1991), "The Use of Prosody in Syntactic Disambiguation," *J. Acoust. Soc. Am.*, 90(6):2956-2970.
- Riley, M. D. (1992), "Tree-based Modeling of Segmental Durations," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, eds., Elsevier, New York, pp. 265-273.
- Ross, K., M. Ostendorf, and S. Shattuck-Hufnagel (1992), "Factors Affecting Pitch Accent Placement," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 365-368.
- Selkirk, E. O. (1984), *Phonology and Syntax: The Relation Between Sound and Structure*, MIT Press, Cambridge, Mass.
- Silverman, K., M. Beckman, J. Pittrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "TOBI: A Standard for Labeling English Prosody," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 867-870.
- Sproat, R. W. (1990), "Stress Assignment in Complex Nominals for English Text-to-Speech," in *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 129-132.
- Sproat, R. W., and M. Y. Liberman (1987), "Toward Treating English Nominals Correctly," in *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 140-146.
- Terken, J. (1993), "Synthesizing Natural-Sounding Intonation for Dutch: Rules and Perceptual Evaluation," *Comput. Speech Lang.*, 7:27-48.
- t'Hart, J., R. Collier, and A. Cohen (1990), *A Perceptual Study of Intonation*, Cambridge University Press, Cambridge, England.
- Van Leeuwen, H. C., and E. te Lindert (1993), "Speech Maker: A Flexible and General Framework for Text-to-Speech Synthesis, and its Application to Dutch," *Comput. Speech Lang.*, 2:149-167.
- Van Santen, J. P. H. (1992), "Deriving Text-to-Speech Durations from Natural Speech," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, eds., Elsevier, New York, pp. 275-285.
- Venezky, R. L. (1970). *The Structure of English Orthography*, Mouton, 's Gravenhage.
- Wales, R., and H. Toner (1979), "Intonation and Ambiguity," in *Sentence Processing*, W. C. Cooper and E. C. T. Walker, eds., Erlbaum, Hillsdale, N.J.
- Wang, M. Q., and J. Hirschberg (1992), "Automatic Classification of Intonational Phrase Boundaries," *Comput. Speech Lang.*, 6:175-196.
- Wightman, C., S. Shattuck-Hufnagel, M. Ostendorf, and P. Price (1992), "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries," *J. Acoust. Soc. Am.*, 91(3):1707-1717.
- Young, S. J., and F. Fallside (1979), "Speech Synthesis from Concept: A Method of Speech Output from Information Systems," *J. Acoust. Soc. Am.*, 66(3):685-695.



## **SPEECH RECOGNITION TECHNOLOGY**



# Speech Recognition Technology: A Critique

*Stephen E. Levinson*

## SUMMARY

This paper introduces the session on advanced speech recognition technology. The two papers comprising this session argue that current technology yields a performance that is only an order of magnitude in error rate away from human performance and that incremental improvements will bring us to that desired level. I argue that, to the contrary, present performance is far removed from human performance and a revolution in our thinking is required to achieve the goal. It is further asserted that to bring about the revolution more effort should be expended on basic research and less on trying to prematurely commercialize a deficient technology.

The title of this paper undoubtedly connotes different things to different people. The intention of the organizing committee of the Colloquium on Human-Machine Communication by Voice, however, was quite specific, namely to review the most advanced technology of the day as it is practiced in research laboratories. Thus, this paper fits rather neatly between one given by J. L. Flanagan, which discusses the fundamental science on which a speech recognition technology might rest, and those of J. G. Wilpon, H. Levitt, C. Seelbach, C. Weinstein, and J. Oberteuffer, which are devoted to real applications of speech recognition machines. While it is true that these ap

plications use derivatives of some of the advanced techniques discussed here, they are not as ambitious as the purely experimental systems.

In keeping with the theme of advanced technology, J. Makhoul and R. Schwartz report on the "State of the Art in Continuous Speech Recognition." They give a phonetic and phonological description of speech and show how that structure is captured by a mathematical object called a hidden Markov model (HMM). This discussion includes a brief account of the history of the HMM and its application in speech recognition. Also included in the paper are discussions of extracting features from the speech waveform, measuring the performance of the system and the possibility of using the newer methods based on artificial neural networks.

Makhoul and Schwartz conclude that, as a result of the advances made in model accuracy, algorithms, and the power of computers, a "paradigm shift" has occurred in the sense that high-accuracy, real-time, speaker-independent, continuous speech recognition for medium-sized vocabularies can be implemented in software running on commercially available workstations. This assertion provoked an important and lively debate that I shall recount later in this paper. The HMM methodology allows us to cast the speech recognition problem as that of searching for the best path through a weighted, directed graph. The paper by F. Jelinek addresses two central and specific technical issues arising from this representation. First, how does one estimate the parameters of the model (i.e., weights of the graph) from data? This is usually referred to as the training problem. Second, given an optimal model, how does one use it in the recognition task? This second problem can be cast as a combinatorial search problem to which Jelinek outlines several solutions with emphasis on a dynamic programming approach known as the Viterbi algorithm.

There is no need to review these papers in more detail here since they appear in their entirety in this volume. What does deserve discussion here are the scientific, technological, and commercial implications of these papers. These issues formed the core of the debate that ensued at the colloquium after these two excellent and comprehensive papers were presented.

I opened the discussion at the colloquium by asking the speakers to evaluate the state of the art of their most advanced laboratory prototype systems with respect to human performance in communication by spoken language. I raised this question because I think the ultimate goal of research in speech recognition is to provide a means whereby people can carry on spoken conversations with machines in the same effortless manner in which they speak to each other. As I

noted earlier, the purpose of this paper is to evaluate the highest expression of such research. Thus, while it may be comfortable to discuss progress in incremental terms, it is more instructive to evaluate our best efforts with respect to our ultimate goals.

My question turned out to be a provocative one on which opinion was sharply divided. Both speakers and a substantial number of participants vigorously supported the following propositions:

- The performance of today's best experimental systems is only an order of magnitude in error rate away from a level that compares favorably with human performance.
- When experimental systems do achieve human-like performance, their structure and methods will be strongly reminiscent of present systems.
- Today's advanced technology is commercially viable.

These and even more strongly optimistic sentiments have been expressed by Oberteuffer (1993).

I was supported in my diametrically opposite opinion of the first two assertions by a few members of the colloquium. The substance of our objections is the following. The current euphoria about speech recognition is based on Makhoul's characterization of our progress as a "paradigm shift." His use of the term is wholly inappropriate and misleading. The phrase was first used by Kuhn (1970) to characterize scientific revolution. Makhoul was thus casting incremental, technical progress as profound, conceptual scientific progress.

The difference is best understood by example. An important paradigm shift in astronomy was brought about by the combination of a heliocentric model of the solar system and the application of Newtonian mechanics to it. Placing the sun rather than the earth at the center of the solar system may seem like a radical idea. Although it is counterintuitive to the naive observer, it does not, by itself, constitute a paradigm shift. The revolutionary concept arises from the consideration of another aspect of the solar system besides planetary position. The Ptolemaic epicycles do predict the positions of the planets as a function of time. In fact, they do so more effectively than the crude elliptical orbits postulated by the Copernican/Newtonian theory. Indeed, by the incremental improvement of compounding epicycles upon epicycles, the incorrect theory can be made to appear more accurate than the coarse but correct one. So clearly, heliocentricity alone is not a paradigm shift.

However, if one asks what forces move the planets on these observed regular paths and how this accounts for their velocities and accelerations, the geocentric theory becomes mute while the classical

mechanical description of the heliocentric model turns eloquent. This, then, is the paradigm shift, and its consequences are enormous. Epicycles are acceptable for making ritual calendars and some navigational calculations, but Newtonian mechanics opens new vistas and, after some careful measurement, becomes highly accurate.

There is a very close analogy between early astronomy and modern speech recognition. At the present moment, we think of the problem of speech recognition as one of transcription, being able to convert the speech signal into a set of discrete symbols representing words. This decoding process corresponds to the computation of celestial position only. It ignores, however, the essence of speech, its capacity to convey important information (i.e., meaning), and is thus incomplete. The paradigm shift needed in our field is to make *meaning* rather than symbolic transcription the central issue in speech recognition, just as *force* replaced location as the central construct in celestial mechanics. If one can compute the forces acting on the planets, one can know their orbits and the positions come for free. Similarly, if one can extract the meaning from a spoken message, the lexical transcription will fall out. Some readers may object to this analogy by noting that the topic of "speech understanding" has been under study for two decades. Unfortunately, the current practice of "speech understanding" does not qualify as the needed paradigm shift because it is an inverted process that aims to use *meaning* to improve *transcription accuracy* rather than making *meaning* the primary aspect.

In short, the incremental improvements in phonetic modeling accuracy and search methods summarized by Makhoul and Jelinek in this session do not constitute a paradigm shift. The fact that these improved techniques can run in near real-time on cheap, readily available hardware is merely a result of the huge advances in microelectronics that came about nearly independent of work in speech technology.

Furthermore, we are very far away from human performance in speech communication. Some attendees have suggested that human performance on the standard ATIS (Air Travel Information Service) task is not much, if at all, better than our best computer programs. I doubt this to be so, but, even if it were, it ignores the simple and crucial fact that the ATIS task is not natural to humans. Although the ATIS scenario was not intended to be unnatural, experimental approaches to it ended up being tailored to our existing capabilities. As such, the task is highly artificial and has only vague similarity to natural human discourse.

I believe it is highly unlikely that any incremental improvements to our existing technology will effectively address the problem of communication with machines in ordinary colloquial discourse un

der ordinary ambient conditions. It seems to me that fundamental science to support a human/machine spoken communication technology is missing. We will return to the question of what that science might be in the paper by Levinson and Fallside (this volume), which deals with future research and technology.

The debate outlined above is central to the continued progress of our field. The way we resolve it will have an enormous effect on the ultimate fate of speech recognition technology. Unfortunately, the debate is not about purely scientific issues but rather reflects the very delicate balance among scientific, technological, and economic factors. As the paper by L. Rabiner based on his opening address to the colloquium makes clear, the explanation of these sometimes contradictory factors is one of the principal motivations for this volume.

Here, then, is this author's admittedly minority opinion concerning the commercial future of today's laboratory-state-of-the-art speech recognition. By definition, any such viewpoint involves technological forecasting, which is one of the main themes of the papers by Levinson and Fallside, S. Furui, B. Atal, and M. Marcus. For the purposes of this discussion, however, it suffices to examine but one feature of technological forecasting. When technocrats predict the future of a new technology, they tend to be overly optimistic for the near term and overly pessimistic for the long haul.

The history of computing provides a classic example. In the early 1950s Von Neumann and his contemporaries foresaw many of the features of modern computing that are now commonplace—for example, time sharing, large memories, and faster speeds—and predicted that they would be immediately available. They also guessed that "computing would be only a tiny part of human activity" (Goldstine, 1972). In fact, the technological advances took much longer to materialize than they had envisioned. Moreover, they completely failed to imagine the enormous growth, 40 years later, of the market for what they envisioned as large computers.

I suggest that the same phenomenon will occur with speech technology. The majority opinion holds that technical improvements will soon make large-vocabulary speech recognition commercially viable for specific applications. My prediction, based on the aforementioned general characterization of technological forecasting, is that technical improvements will appear painfully slowly but that in 40 to 50 years speech recognition at human performance levels will be ubiquitous. That is, incremental technical advances will, in the near term, result in a fragile technology of relatively small commercial value in very special markets, whereas major technological advances resulting from a true paradigm shift in the underlying science will enable machines to display human levels of competence in spoken language commu

nication. This, in turn, will result in a vast market of incalculable commercial value.

It is, of course, entirely possible that the majority opinion is correct, that a diligent effort resulting in a long sequence of rapid incremental improvements will yield the desired perfected speech recognition technology. It is, unfortunately, also possible that this strategy will run afoul of the "first step fallacy" (Dreyfus, 1972), which warns that one cannot reach the moon by climbing a tree even though such an action initially appears to be a move in the right direction. Ultimately, progress stops far short of the goal when the top of the tree is reached.

If, as I argue, the latter possibility exists, what strategy should we use to defend against its undesirable outcome? The answer should be obvious. Openly acknowledge the risks of the incremental approach and devote some effort to achieving the paradigm shift from signal transcription to message comprehension alluded to earlier.

Perhaps more important, however, is recognition of the uniqueness of our technological goal. Unlike all other technologies that are integral parts of our daily lives because they provide us with capabilities otherwise unattainable, automatic speech recognition promises to improve the usefulness of a behavior at which we are already exquisitely proficient. Such a promise cannot be realized if the technology supporting it degrades our natural expertise in spoken communication. Since the present state of the art requires a serious diminution of our abilities and since we presently do not know how to leap the performance chasm between humans and machines, perhaps we should invest more in research aimed at finding a more nearly anthropomorphic and, by implication, potent technology. This would, of course, alter the subtle balance among science, technology, and the marketplace more toward precommercial experimentation with proportionately less opportunity for immediate profit. There is good reason to believe, however, that ultimately this strategy will afford the greatest intellectual, financial, and social rewards.

## REFERENCES

- Dreyfus, H. L., *What Computers Can't Do: A Critique of Artificial Reason*, Harper, New York, 1972.  
Goldstine, H. H., *The Computer from Pascal to Von Neumann*, Princeton University Press, Princeton, N.J., 1972, p. 344.  
Kuhn, T. S., *The Structure of Scientific Revolutions*, 2nd ed., University of Chicago Press, Chicago, 1970, pp. 92 ff.  
Oberteuffer, J., "Major Progress in Speech Technology Affirmed at National Academy of Sciences Colloquium," *ASR News*, Vol. 4, No. 2, Feb. 1993, pp. 5-7.

# State of the Art in Continuous Speech Recognition

*John Makhoul and Richard Schwartz*

## SUMMARY

In the past decade, tremendous advances in the state of the art of automatic speech recognition by machine have taken place. A reduction in the word error rate by more than a factor of 5 and an increase in recognition speeds by several orders of magnitude (brought about by a combination of faster recognition search algorithms and more powerful computers), have combined to make high-accuracy, speaker-independent, continuous speech recognition for large vocabularies possible in real-time, on off-the-shelf workstations, without the aid of special hardware. These advances promise to make speech recognition technology readily available to the general public. This paper focuses on the speech recognition advances made through better speech modeling techniques, chiefly through more accurate mathematical modeling of speech sounds.

## INTRODUCTION

More and more, speech recognition technology is making its way from the laboratory to real-world applications. Recently, a qualitative change in the state of the art has emerged that promises to bring speech recognition capabilities within the reach of anyone with access to a workstation. High-accuracy, real-time, speaker-independent,

continuous speech recognition for medium-sized vocabularies (a few thousand words) is now possible in software on off-the-shelf workstations. Users will be able to tailor recognition capabilities to their own applications. Such software-based, real-time solutions usher in a whole new era in the development and utility of speech recognition technology.

As is often the case in technology, a paradigm shift occurs when several developments converge to make a new capability possible. In the case of continuous speech recognition, the following advances have converged to make the new technology possible:

- higher-accuracy continuous speech recognition, based on better speech modeling techniques;
- better recognition search strategies that reduce the time needed for high-accuracy recognition; and
- increased power of audio-capable, off-the-shelf workstations.

The paradigm shift is taking place in the way we view and use speech recognition. Rather than being mostly a laboratory endeavor, speech recognition is fast becoming a technology that is pervasive and will have a profound influence on the way humans communicate with machines and with each other.

This paper focuses on speech modeling advances in continuous speech recognition, with an exposition of hidden Markov models (HMMs), the mathematical backbone behind these advances. While knowledge of properties of the speech signal and of speech perception have always played a role, recent improvements have relied largely on solid mathematical and probabilistic modeling methods, especially the use of HMMs for modeling speech sounds. These methods are capable of modeling time and spectral variability simultaneously, and the model parameters can be estimated automatically from given training speech data. The traditional processes of segmentation and labeling of speech sounds are now merged into a single probabilistic process that can optimize recognition accuracy.

This paper describes the speech recognition process and provides typical recognition accuracy figures obtained in laboratory tests as a function of vocabulary, speaker dependence, grammar complexity, and the amount of speech used in training the system. As a result of modeling advances, recognition error rates have dropped several fold. Important to these improvements have been the availability of common speech corpora for training and testing purposes and the adoption of standard testing procedures.

This paper also reviews more recent research directions, including the use of segmental models and artificial neural networks in

improving the performance of HMM systems. The capabilities of neural networks to model highly nonlinear functions can be used to develop new features from the speech signal, and their ability to model posterior probabilities can be used to improve recognition accuracy.

We will argue that future advances in speech recognition must continue to rely on finding better ways to incorporate our speech knowledge into advanced mathematical models, with an emphasis on methods that are robust to speaker variability, noise, and other acoustic distortions.

## THE SPEECH RECOGNITION PROBLEM

Automatic speech recognition can be viewed as a mapping from a continuous-time signal, the speech signal, to a sequence of discrete entities, for example, phonemes (or speech sounds), words, and sentences. The major obstacle to high-accuracy recognition is the large variability in the speech signal characteristics. This variability has three main components: linguistic variability, speaker variability, and channel variability. Linguistic variability includes the effects of phonetics, phonology, syntax, semantics, and discourse on the speech signal. Speaker variability includes intra- and interspeaker variability, including the effects of coarticulation, that is, the effects of neighboring sounds on the acoustic realization of a particular phoneme, due to continuity and motion constraints on the human articulatory apparatus. Channel variability includes the effects of background noise and the transmission channel (e.g., microphone, telephone, reverberation). All these variabilities tend to shroud the intended message with layers of uncertainty, which must be unraveled by the recognition process.

### General Synthesis/Recognition Process

We view the recognition process as one component of a general synthesis/recognition process, as shown in [Figure 1](#). We assume that the synthesis process consists of three components: a structural model, a statistical variability model, and the synthesis of the speech signal. The input is some underlying event, such as a sequence of words, and the output is the actual speech signal. The structural model comprises many aspects of our knowledge of speech and language, and the statistical variability model accounts for the different variabilities that are encountered. The recognition process begins with analysis of the speech signal into a sequence of feature vectors. This analysis serves to reduce one aspect of signal variability due to changes in

pitch, etc. Given the sequence of feature vectors, the recognition process reduces to a search over all possible events (word sequences) for that event which has the highest probability given the sequence of feature vectors, based on the structural and statistical variability models used in the synthesis.

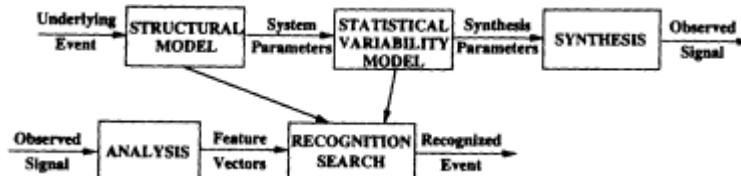


FIGURE 1 General synthesis/recognition process.

It is important to note that a significant and important amount of speech knowledge is incorporated in the structural model, including our knowledge of language structure, speech production, and speech perception. Examples of language structure include the fact that continuous speech consists of a concatenation of words and that words are a concatenation of basic speech sounds or phonemes. This knowledge of language structure is quite ancient, being at least 3000 years old. A more recent aspect of language structure that was appreciated in this century is the fact that the acoustic realization of phonemes is heavily dependent on the neighboring context. Our knowledge of speech production, in terms of manner of articulation (e.g., voiced, fricated, nasal) and place of articulation (e.g., velar, palatal, dental, labial), for example, can be used to provide parsimonious groupings of phonetic context. As for speech perception, much is known about sound analysis in the cochlea, for example, that the basilar membrane performs a form of quasi-spectral analysis on a nonlinear frequency scale, and about masking phenomena in time and frequency. All this knowledge can be incorporated beneficially in our modeling of the speech signal for recognition purposes.

### Units of Speech

To gain an appreciation of what modeling is required to perform recognition, we shall use as an example the phrase "grey whales," whose speech signal is shown at the bottom of [Figure 2](#) with the corresponding spectrogram (or voice print) shown immediately above. The spectrogram shows the result of a frequency analysis of the speech,

with the dark bands representing resonances of the vocal tract. At the top of Figure 2 are the two words "grey" and "whales," which are the desired output of the recognition system. The first thing to note is that the speech signal and the spectrogram show no separation between the two words "grey" and "whales" at all; they are in fact connected. This is typical of continuous speech; the words are connected to each other, with no apparent separation. The human perception that a speech utterance is composed of a sequence of discrete words is a purely perceptual phenomenon. The reality is that the words are not separated at all physically.

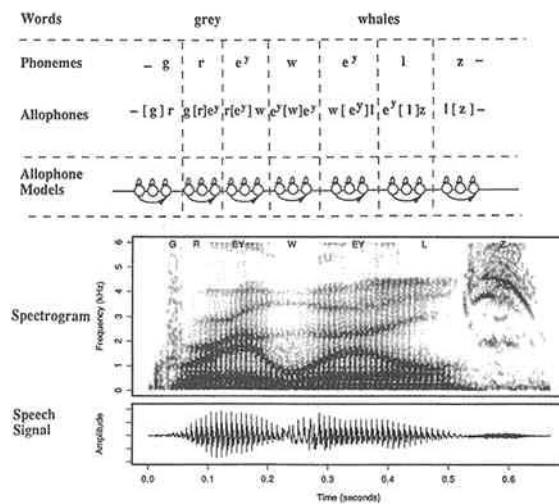


FIGURE 2 Units of speech.

Below the word level in Figure 2 is the phonetic level. Here the words are represented in terms of a phonetic alphabet that tells us what the different sounds in the two words are. In this case the phonetic transcription is given by [g r e<sup>y</sup> w e<sup>y</sup> l z ]. Again, while the sequence of phonemes is discrete, there is no physical separation

between the different sounds in the speech signal. In fact, it is not clear where one sound ends and the next begins. The dashed vertical lines shown in [Figure 2](#) give a rough segmentation of the speech signal, which shows approximately the correspondences between the phonemes and the speech.

Now, the phoneme [e<sup>y</sup>] occurs once in each of the two words. If we look at the portions of the spectrogram corresponding to the two [e<sup>y</sup>] phonemes, we notice some similarities between the two parts, but we also note some differences. The differences are mostly due to the fact that the two phonemes are in different contexts: the first [e<sup>y</sup>] phoneme is preceded by [r] and followed by [w], while the second is preceded by [w] and followed by [l]. These contextual effects are the result of what is known as coarticulation, the fact that the articulation of each sound blends into the articulation of the following sound. In many cases, contextual phonetic effects span several phonemes, but the major effects are caused by the two neighboring phonemes.

To account for the fact that the same phoneme has different acoustic realizations, depending on the context, we refer to each specific context as an allophone. Thus, in [Figure 2](#), we have two different allophones of the phoneme [e<sup>y</sup>], one for each of the two contexts in the two words. In this way, we are able to deal with the phonetic variability that is inherent in coarticulation and that is evident in the spectrogram of [Figure 2](#).

To perform the necessary mapping from the continuous speech signal to the discrete phonetic level, we insert a model—a finite-state machine in our case—for each of the allophones that are encountered. We note from [Figure 2](#) that the structures of these models are identical; the differences will be in the values given to the various model parameters. Each of these models is a hidden Markov model, which is discussed below.

## HIDDEN MARKOV MODELS

### Markov Chains

Before we explain what a hidden Markov model is, we remind the reader of what a Markov chain is. A Markov chain consists of a number of states, with transitions among the states. Associated with each transition is a probability and associated with each state is a symbol. [Figure 3](#) shows a three-state Markov chain, with transition probabilities  $a_{ij}$  between states  $i$  and  $j$ . The symbol A is associated with state 1, the symbol B with state 2, and the symbol C with state 3. As one transitions from state 1 to state 2, for example, the symbol B is

produced as output. If the next transition is from state 2 to itself, the symbol B is output again, while if the transition were to state 3, the output would be the symbol C. These symbols are called output symbols because a Markov chain is thought of as a generative model; it outputs symbols as one transitions from one state to another. Note that in a Markov chain the transitioning from one state to another is probabilistic, but the production of the output symbols is deterministic.

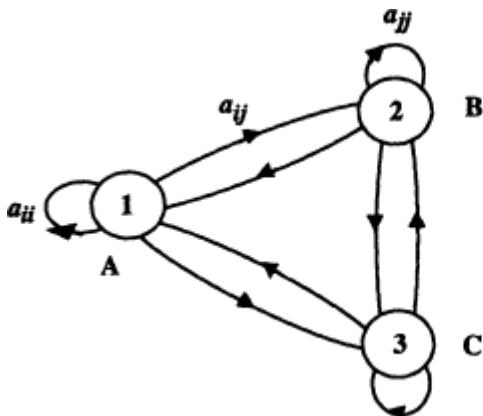


FIGURE 3 A three-state Markov chain.

Now, given a sequence of output symbols that were generated by a Markov chain, one can retrace the corresponding sequence of states completely and unambiguously (provided the output symbol for each state was unique). For example, the sample symbol sequence B A A C B B A C C C A is produced by transitioning into the following sequence of states: 2 1 1 3 2 2 1 3 3 3 1.

### **Hidden Markov Models**

A hidden Markov model (HMM) is the same as a Markov chain, except for one important difference: the output symbols in an HMM are probabilistic. Instead of associating a single output symbol per state, in an HMM all symbols are possible at each state, each with its own probability. Thus, associated with each state is a probability distribution of all the output symbols. Furthermore, the number of output symbols can be arbitrary. The different states may then have different probability distributions defined on the set of output symbols. The probabilities associated with states are known as output probabilities. (If instead of having a discrete number of output sym

bols we have a continuously valued vector, it is possible to define a probability density function over all possible values of the random output vector. For the purposes of this exposition, we shall limit our discussion to discrete output symbols.)

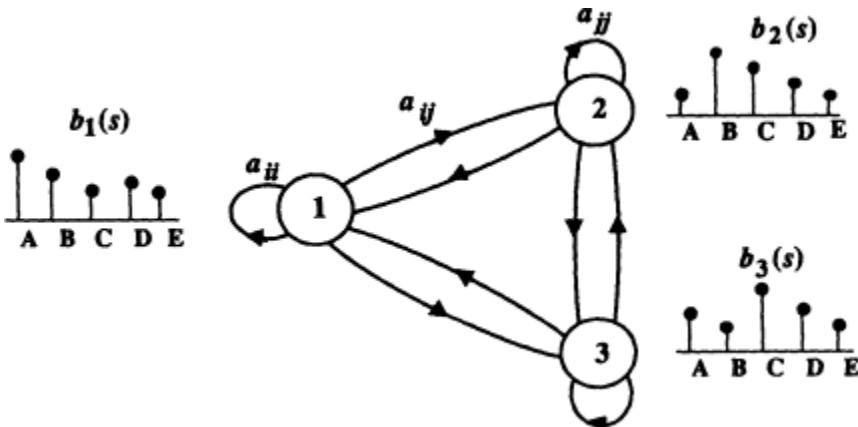


FIGURE 4 A three-state hidden Markov model.

Figure 4 shows an example of a three-state HMM. It has the same transition probabilities as the Markov chain of Figure 3. What is different is that we associate a probability distribution  $b_i(s)$  with each state  $i$ , defined over the set of output symbols  $s$ —in this case we have five output symbols—A, B, C, D, and E. Now, when we transition from one state to another, the output symbol is chosen according to the probability distribution corresponding to that state. Compared to a Markov chain, the output sequences generated by an HMM are what is known as doubly stochastic: not only is the transitioning from one state to another stochastic (probabilistic) but so is the output symbol generated at each state.

Now, given a sequence of symbols generated by a particular HMM, it is not possible to retrace the sequence of states unambiguously. Every sequence of states of the same length as the sequence of symbols is possible, each with a different probability. Given the sample output sequence—C D A A B E D B A C C—there is no way for sure to know which sequence of states produced these output symbols. We say that the sequence of states is hidden in that it is hidden from the observer if all one sees is the output sequence, and that is why these models are known as *hidden* Markov models.

Even though it is not possible to determine for sure what se

quence of states produced a particular sequence of symbols, one might be interested in the sequence of states that has the highest probability of having generated the given sequence. To find such a sequence of states requires a search procedure that, in principle, must examine all possible state sequences and compute their corresponding probabilities. The number of possible state sequences grows exponentially with the length of the sequence. However, because of the Markov nature of an HMM, namely that being in a state is dependent only on the previous state, there is an efficient search procedure called the Viterbi algorithm (Forney, 1973) that can find the sequence of states most likely to have generated the given sequence of symbols, without having to search all possible sequences. This algorithm requires computation that is proportional to the number of states in the model and to the length of the sequence.

### Phonetic HMMs

We now explain how HMMs are used to model phonetic speech events. Figure 5 shows an example of a three-state HMM for a single phoneme. The first stage in the continuous-to-discrete mapping that

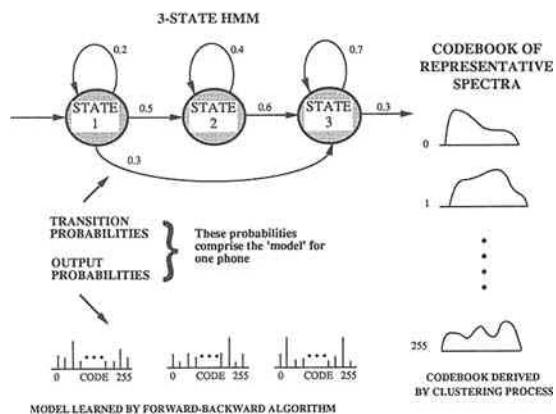


FIGURE 5 Basic structure of a phonetic HMM.

is required for recognition is performed by the analysis box in [Figure 1](#). Typically, the analysis consists of estimation of the short-term spectrum of the speech signal over a frame (window) of about 20 ms. The spectral computation is then updated about every 10 ms, which corresponds to a frame rate of 100 frames per second. This completes the initial discretization in time. However, the HMM, as depicted in this paper, also requires the definition of a discrete set of "output symbols." So, we need to discretize the spectrum into one of a finite set of spectra. [Figure 5](#) depicts a set of spectral templates (known as a codebook) that represent the space of possible speech spectra. Given a computed spectrum for a frame of speech, one can find the template in the codebook that is "closest" to that spectrum, using a process known as vector quantization (Makhoul et al., 1985). The size of the codebook in [Figure 5](#) is 256 templates. These templates, or their indices (from 0 to 255), serve as the output symbols of the HMM. We see in [Figure 5](#) that associated with each state is a probability distribution on the set of 256 symbols. The definition of a phonetic HMM is now complete. We now describe how it functions.

Let us first see how a phonetic HMM functions as a generative (synthesis) model. As we enter into state 1 in [Figure 5](#), one of the 256 output symbols is generated based on the probability distribution corresponding to state 1. Then, based on the transition probabilities out of state 1, a transition is made either back to state 1 itself, to state 2, or to state 3, and another symbol is generated based on the probability distribution corresponding to the state into which the transition is made. In this way a sequence of symbols is generated until a transition out of state 3 is made. At that point, the sequence corresponds to a single phoneme.

The same model can be used in recognition mode. In this mode each model can be used to compute the probability of having generated a sequence of spectra. Assuming we start with state 1 and given an input speech spectrum that has been quantized to one of the 256 templates, one can perform a table lookup to find the probability of that spectrum. If we now assume that a transition is made from state 1 to state 2, for example, the previous output probability is multiplied by the transition probability from state 1 to state 2 (0.5 in [Figure 5](#)). A new spectrum is now computed over the next frame of speech and quantized; the corresponding output probability is then determined from the output probability distribution corresponding to state 2. That probability is multiplied by the previous product, and the process is continued until the model is exited. The result of multiplying the sequence of output and transition probabilities gives the total probability that the input spectral sequence was "generated" by

that HMM using a specific sequence of states. For every sequence of states, a different probability value results. For recognition, the probability computation just described is performed for all possible phoneme models and all possible state sequences. The one sequence that results in the highest probability is declared to be the recognized sequence of phonemes.

We note in [Figure 5](#) that not all transitions are allowed (i.e., the transitions that do not appear have a probability of zero). This model is what is known as a "left-to-right" model, which represents the fact that, in speech, time flows in a forward direction only; that forward direction is represented in [Figure 5](#) by a general left-to-right movement. Thus, there are no transitions allowed from right to left. Transitions from any state back to itself serve to model variability in time, which is very necessary for speech since different instantiations of phonemes and words are uttered with different time registrations. The transition from state 1 to state 3 means that the shortest phoneme that is modeled by the model in [Figure 5](#) is one that is two frames long, or 20 ms. Such a phoneme would occupy state 1 for one frame and state 3 for one frame only. One explanation for the need for three states, in general, is that state 1 corresponds roughly to the left part of the phoneme, state 2 to the middle part, and state 3 to the right part. (More states can be used, but then more data would be needed to estimate their parameters robustly.)

Usually, there is one HMM for each of the phonetic contexts of interest. Although the different contexts could have different structures, usually all such models have the same structure as the one shown in [Figure 5](#); what makes them different are the transition and output probabilities.

## A HISTORICAL OVERVIEW

HMM theory was developed in the late 1960s by Baum and colleagues (Baum and Eagon, 1967) at the Institute for Defense Analyses (IDA). Initial work using HMMs for speech recognition was performed in the 1970s at IDA, IBM (Jelinek et al., 1975), and Carnegie-Mellon University (Baker, 1975). In 1980 a number of researchers in speech recognition in the United States were invited to a workshop in which IDA researchers reviewed the properties of HMMs and their use for speech recognition. That workshop prompted a few organizations, such as AT&T and BBN, to start working with HMMs (Levinson et al., 1983; Schwartz et al., 1984). In 1984 a program in continuous speech recognition was initiated by the Advanced Research Projects Agency (ARPA), and soon thereafter HMMs were shown to be supe

rior to other approaches (Chow et al., 1986). Until then, only a handful of organizations worldwide had been working with HMMs. Because of the success of HMMs and because of the strong influence of the ARPA program, with its emphasis on periodic evaluations using common speech corpora, the use of HMMs for speech recognition started to spread worldwide. Today, their use has dominated other approaches to speech recognition in dozens of laboratories around the globe. In addition to the laboratories mentioned above, significant work is taking place at, for example, the Massachusetts Institute of Technology's Lincoln Laboratory, Dragon, SRI, and TI in the United States; CRIM and BNR in Canada; RSRE and Cambridge University in the United Kingdom; ATR, NTT, and NEC in Japan; LIMSI in France; Philips in Germany and Belgium; and CSELT in Italy, to name a few. Comprehensive treatments of HMMs and their utility in speech recognition can be found in Rabiner (1989), Lee (1989), Huang et al. (1990), Rabiner and Juang (1993), and the references therein. Research results in this area are usually reported in the following journals and conference proceedings: *IEEE Transactions on Speech and Audio Processing*; *IEEE Transactions on Signal Processing*; *Speech Communication Journal*; IEEE International Conference on Acoustics, Speech, and Signal Processing; Eurospeech; and the International Conference on Speech and Language Processing.

HMMs have proven to be a good model of speech variability in time and feature space. The automatic training of the models from speech data has accelerated the speed of research and improved recognition performance. Also, the probabilistic formulation of HMMs has provided a unified framework for scoring of hypotheses and for combining different knowledge sources. For example, the sequence of spoken words can also be modeled as the output of another statistical process (Bahl et al., 1983). In this way it becomes natural to combine the HMMs for speech with the statistical models for language.

## TRAINING AND RECOGNITION

Figure 6 shows a block diagram of a general system for training and recognition. Note that in both training and recognition the first step in the process is to perform feature extraction on the speech signal.

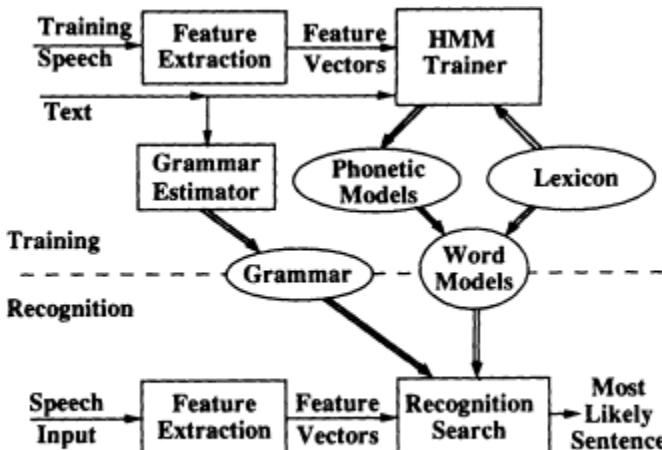


FIGURE 6 General system for training and recognition.

### Feature Extraction

In theory it should be possible to recognize speech directly from the signal. However, because of the large variability of the speech signal, it is a good idea to perform some form of feature extraction to reduce that variability. In particular, computing the envelope of the short-term spectrum reduces the variability significantly by smoothing the detailed spectrum, thus eliminating various source characteristics, such as whether the sound is voiced or fricated, and, if voiced, it eliminates the effect of the periodicity or pitch. The loss of source information does not appear to affect recognition performance much because it turns out that the spectral envelope is highly correlated with the source information.

One reason for computing the short-term spectrum is that the cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a nonlinear frequency scale (known as the Bark scale or the mel scale). This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. So, in the feature extraction, it is very common to perform a frequency warping of the frequency axis after the spectral computation.

Researchers have experimented with many different types of features for use with HMMs (Rabiner and Juang, 1993). Variations on the basic spectral computation, such as the inclusion of time and frequency masking, have been shown to provide some benefit in certain cases. The use of auditory models as the basis for feature extrac

tion has been useful in some systems (Cohen, 1989), especially in noisy environments (Hunt et al., 1991).

Perhaps the most popular features used for speech recognition with HMMs today are what are known as mel-frequency cepstral coefficients or MFCCs (Davis and Mermelstein, 1980). After the mel-scale warping of the spectrum, the logarithm of the spectrum is taken and an inverse Fourier transform results in the cepstrum. By retaining the first dozen or so coefficients of the cepstrum, one would be retaining the spectral envelope information that is desired. The resulting features are the MFCCs, which are treated as a single vector and are typically computed for every frame of 10 ms. These feature vectors form the input to the training and recognition systems.

## Training

Training is the process of estimating the speech model parameters from actual speech data. In preparation for training, what is needed is the text of the training speech and a lexicon of all the words in the training, along with their pronunciations, written down as phonetic spellings. Thus, a transcription of the training speech is made by listening to the speech and writing down the sequence of words. All the distinct words are then placed in a lexicon and someone has to provide a phonetic spelling of each word. In cases where a word has more than one pronunciation, as many phonetic spellings as there are pronunciations are included for each word. These phonetic spellings can be obtained from existing dictionaries or they can be written by anyone with minimal training in phonetics.

## Phonetic HMMs and Lexicon

Given the training speech, the text of the speech, and the lexicon of phonetic spellings of all the words, the parameters of all the phonetic HMMs (transition and output probabilities) are estimated automatically using an iterative procedure known as the Baum-Welch or forward-backward algorithm (Baum and Eagon, 1967). This algorithm estimates the parameters of the HMMs so as to maximize the likelihood (probability) that the training speech was indeed produced by these HMMs. The iterative procedure is guaranteed to converge to a local optimum. Typically, about five iterations through the data are needed to obtain a reasonably good estimate of the speech model. (See the paper by Jelinek in this volume for more details on the HMM training algorithm.)

It is important to emphasize the fact that HMM training does not

require that the data be labeled in detail in terms of the location of the different words and phonemes, that is, no time alignment between the speech and the text is needed. Given a reasonable initial estimate of the HMM parameters, the Baum-Welch training algorithm performs an implicit alignment of the input spectral sequence to the states of the HMM, which is then used to obtain an improved estimate. All that is required in addition to the training speech is the text transcription and the lexicon. This is one of the most important properties of the HMM approach to recognition. Training does require significant amounts of computing but does not require much in terms of human labor.

In preparation for recognition it is important that the lexicon contain words that would be expected to occur in future data, even if they did not occur in the training. Typically, closed-set word classes are filled out—for example, days of the week, months of the year, numbers.

After completing the lexicon, HMM word models are compiled from the set of phonetic models using the phonetic spellings in the lexicon. These word models are simply a concatenation of the appropriate phonetic HMM models. We then compile the grammar (which specifies sequences of words) and the lexicon (which specifies sequences of phonemes for each word) into a single probabilistic grammar for the sequences of phonemes. The result of the recognition is a particular sequence of words, corresponding to the recognized sequence of phonemes.

## Grammar

Another aspect of the training that is needed to aid in the recognition is to produce the grammar to be used in the recognition. Without a grammar, all words would be considered equally likely at each point in an utterance, which would make recognition difficult, especially with large vocabularies. We, as humans, make enormous use of our knowledge of the language to help us recognize what a person is saying. A grammar places constraints on the sequences of the words that are allowed, giving the recognition fewer choices at each point in the utterance and, therefore, improving recognition performance.

Most grammars used in speech recognition these days are statistical Markov grammars that give the probabilities of different sequences of words—so-called n-gram grammars. For example, bigram grammars give the probabilities of all pairs of words, while trigram grammars give the probabilities of all triplets of words in the lexicon. In practice, trigrams appear to be sufficient to embody much of the

natural constraints imposed on the sequences of words in a language. In an n-gram Markov grammar, the probability of a word is a function of the previous  $n - 1$  words. While this assumption may not be valid in general, it appears to be sufficient to result in good recognition accuracy. Furthermore, the assumption allows for efficient computation of the likelihood of a sequence of words.

A measure of how constrained a grammar is is given by its *perplexity* (Bahl et al., 1983). Perplexity is defined as 2 raised to the power of the Shannon entropy of the grammar. If all words are equally likely at each point in a sentence, the perplexity is equal to the vocabulary size. In practice, sequences of words have greatly differing probabilities, and the perplexity is often much less than the vocabulary size, especially for larger vocabularies. Because grammars are estimated from a set of training data, it is often more meaningful to measure the perplexity on an independent set of data, or what is known as test-set perplexity (Bahl et al., 1983). Test-set perplexity  $Q$  is obtained by computing

$$Q = P(w_1 w_2 \dots w_M)^{-1/M}$$

where  $w_1 w_2 \dots w_M$  is the sequence of words obtained by concatenating all the test sentences and  $P$  is the probability of that whole sequence. Because of the Markov property of n-gram grammars, the probability  $P$  can be computed as the product of consecutive conditional probabilities of n-grams.

## Recognition

As shown in [Figure 6](#), the recognition process starts with the feature extraction stage, which is identical to that performed in the training. Then, given the sequence of feature vectors, the word HMM models, and the grammar, the recognition is simply a large search among all possible word sequences for that word sequence with the highest probability to have generated the computed sequence of feature vectors. In theory the search is exponential with the number of words in the utterance. However, because of the Markovian property of conditional independence in the HMM, it is possible to reduce the search drastically by the use of dynamic programming (e.g., using the Viterbi algorithm). The Viterbi algorithm requires computation that is proportional to the number of states in the model and the length of the input sequence. Further approximate search algorithms have been developed that allow the search computation to be reduced further, without significant loss in performance. The most com

monly used technique is the beam search (Lowerre, 1976), which avoids the computation for states that have low probability.

## STATE OF THE ART

In this section we review the state of the art in continuous speech recognition. We present some of the major factors that led to the relatively large improvements in performance and give sample performance figures under different conditions. We then review several of the issues that affect performance, including the effects of training and grammar, speaker-dependent versus speaker-independent recognition, speaker adaptation, nonnative speakers, and the inclusion of new words in the vocabulary. Most of the results and examples below have been taken from the ARPA program, which has sponsored the collection and dissemination of large speech corpora for comparative evaluation, with specific examples taken from work most familiar to the authors.

### Improvements in Performance

The improvements in speech recognition performance have been so dramatic that in the ARPA program the word error rate has dropped by a factor of 5 in 5 years! This unprecedented advance in the state of the art is due to four factors: use of common speech corpora, improved acoustic modeling, improved language modeling, and a faster research experimentation cycle.

### Common Speech Corpora

The ARPA program must be given credit for starting and maintaining a sizable program in large-vocabulary, speaker-independent, continuous speech recognition. One of the cornerstones of the ARPA program has been the collection and use of common speech corpora for system development and testing. (The various speech corpora collected under this program are available from the Linguistic Data Consortium, with offices at the University of Pennsylvania.) The first large corpus was the Resource Management (RM) corpus (Price et al., 1988), which was a collection of read sentences from a 1000-word vocabulary in a naval resource management domain. Using this corpus as the basis for their work, the various sites in the program underwent a series of tests of competing recognition algorithms every 6 to 9 months. The various algorithms developed were shared with the other participants after every evaluation, and the successful

ones were quickly incorporated by the different sites. In addition to the algorithms developed by the sites in the program, other algorithms were also incorporated from around the globe, especially from Europe and Japan. This cycle of algorithm development, evaluation, and sharing of detailed technical information led to the incredible reduction in error rate noted above.

### Acoustic Modeling

A number of ideas in acoustic modeling have led to significant improvements in performance. Developing HMM phonetic models that depend on context, that is, on the left and right phonemes, have been shown to reduce the word error rate by about a factor of 2 over context-independent models (Chow et al., 1986). Of course, with context-dependent models, the number of models increases significantly. In theory, if there are 40 phonemes in the system, the number of possible triphone models is  $40^3 = 64,000$ . However, in practice, only a few thousand of these triphones might actually occur. So, only models of the triphones that occur in the training data are usually estimated. If particular triphones in the test do not occur in the training, the allophone models used may be the diphones or even the context-independent models. One of the properties of HMMs is that different models (e.g., context-independent, diphone, and triphone models) can be interpolated in such a way as to make the best possible use of the training data, thus increasing the robustness of the system.

Because most systems are implemented as word recognition systems (rather than phoneme recognition systems), it is usually not part of the basic recognition system to deal with cross-word contextual effects and, therefore, including those effects in the recognition can increase the computational burden substantially. The modeling of cross-word effects is most important for small words, especially function words (where many of the errors occur), and can reduce the overall word error rate by about 20 percent.

In addition to the use of feature vectors, such as MFCCs, it has been found that including what is known as delta features—the change in the feature vector over time—can reduce the error rate by a factor of about 2 (Furui, 1986). The delta features are treated like an additional feature vector whose probability distribution must also be estimated from training data. Even though the original feature vector contains all the information that can be used for the recognition, it appears that the HMM does not take full advantage of the time evolution of the feature vectors. Computing the delta parameters is a

way of extracting that time information and providing it to the HMM directly (Gupta et al., 1987).

Proper estimation of the HMM parameters—the transition and output probabilities—from training data is of crucial importance. Because only a small number of the possible feature vector values will occur in any training set, it is important to use probability estimation and smoothing techniques that not only will model the training data well but also will model other possible occurrences in future unseen data. A number of probability estimation and smoothing techniques have been developed that strike a good compromise between computation, robustness, and recognition accuracy and have resulted in error rate reductions of about 20 percent compared to the discrete HMMs presented in the section titled "Hidden Markov Models" (Bellegarda and Nahamoo, 1989; Gauvain and Lee, 1992; Huang et al., 1990; Schwartz et al., 1989).

## Language Modeling

As mentioned above, statistical n-gram grammars, especially word trigrams, have been very successful in modeling the likely word sequences in actual speech data. To obtain a good language model, it is important to use as large a text corpus as possible so that all the trigrams to be seen in any new test material are seen in the training with about the same probability. Note that only the text is needed for training the language model, not the actual speech. Typically, millions of words of text are used to develop good language models. A number of methods have been developed that provide a robust estimate of the trigram probabilities (Katz, 1987; Placeway et al., 1993).

For a large-vocabulary system, there is little doubt that the completeness, accuracy, and robustness of the language model can play a major role in the recognition performance of the system. Since one cannot always predict what new material is possible in a large-vocabulary domain, it will be important to develop language models that can change dynamically as the input data change (Della Pietra et al., 1992).

## Research Experimentation Cycle

We have emphasized above the recognition improvements that have been possible with innovations in algorithm development. However, those improvements would not have been possible without the proper computational tools that have allowed the researcher to shorten the research experimentation cycle. Faster search algorithms, as well

as faster workstations, have made it possible to run a large experiment in a short time, typically overnight, so that the researcher can make appropriate changes the next day and run another experiment. The combined increases in speed with better search and faster machines have been several orders of magnitude.

### Sample Performance Figures

[Figure 7](#) gives a representative sampling of state-of-the-art continuous speech recognition performance. The performance is shown in terms of the word error rate, which is defined as the sum of word substitutions, deletions, and insertions, as a percentage of the actual number of words in the test. All training and test speakers were native speakers of American English. The error rates are for speaker-independent recognition, that is, test speakers were different from the speakers used for training. All the results in [Figure 7](#) are for laboratory systems; they were obtained from the following references (Bates et al., 1993; Cardin et al., 1993; Haeb-Umbach et al., 1993; Huang et al., 1991; Pallett et al., 1993).

The results for four corpora are shown: the TI connected-digit corpus (Leonard, 1984), the ARPA Resource Management corpus, the ARPA Airline Travel Information Service (ATIS) corpus (MADCOW, 1992), and the ARPA Wall Street Journal (WSJ) corpus (Paul, 1992).

Corpus	Training Data		Vocabulary		Test Data		Word Error Rate
	Type	Amount	Size	Open/Closed	Type	Perplexity	
TI Digits	Read	4 hrs	10	Closed	Read	11	0.3%
ARPA Resource Management	Read	4 hrs	1000	Closed	Read	80	4%
ARPA Airline Travel	Spontaneous	13 hrs	1800	Open	Spontaneous	12	4%
ARPA Wall Street Journal Dictation	Read	12 hrs	5000	Closed	Read	45	5%
	Read	12 hrs	20,000	Open	Read	200	13%
	Read	12 hrs	20,000	Open	Spontaneous	255	26%

FIGURE 7 State of the art in speaker-independent, continuous speech recognition.

The first two corpora were collected in very quiet rooms at TI, while the latter two were collected in office environments at several different sites. The ATIS corpus was collected from subjects trying to access airline information by voice using natural English queries; it is the only corpus of the four presented here for which the training and test speech are spontaneous instead of being read sentences. The WSJ corpus consists largely of read sentences from the *Wall Street Journal*, with some spontaneous sentences used for testing. Shown in [Figure 7](#) are the vocabulary size for each corpus and whether the vocabulary is closed or open. The vocabulary is closed when all the words in the test are guaranteed to be in the system's lexicon, while in the open condition the test may contain words that are not in the system's lexicon and, therefore, will cause errors in the recognition. The perplexity is the test-set perplexity defined above. Strictly speaking, perplexity is not defined for the open vocabulary condition, so the value of the perplexity that is shown was obtained by making some simple assumptions about the probability of n-grams that contain the unknown words.

The results shown in [Figure 7](#) are average results over a number of test speakers. The error rates for individual speakers vary over a relatively wide range and may be several times lower or higher than the average values shown. Since much of the data were collected in relatively benign conditions, one would expect the performance to degrade in the presence of noise and channel distortion. It is clear from [Figure 7](#) that higher perplexity, open vocabulary, and spontaneous speech tend to increase the word error rate. We shall quantify some of these effects next and discuss some important issues that affect performance.

### Effects of Training and Grammar

It is well recognized that increasing the amount of training data generally decreases the word error rate. However, it is important that the increased training be representative of the types of data in the test. Otherwise, the increased training might not help.

With the RM corpus, it has been found that the error rate is inversely proportional to the square root of the amount of training data, so that quadrupling the training data results in cutting the word error rate by a factor of 2. This large reduction in error rate by increasing the training data may have been the result of an artifact of the RM corpus, namely, that the sentence patterns of the test data were the same as those in the training. In a realistic corpus, where the sentence patterns of the test can often be quite different from the

training, such improvements may not be as dramatic. For example, recent experiments with the WSJ corpus have failed to show significant reduction in error rate by doubling the amount of training. However, it is possible that increasing the complexity of the models as the training data are increased could result in larger reduction in the error rate. This is still very much a research issue.

Word error rates generally increase with an increase in grammar perplexity. A general rule of thumb is that the error rate increases as the square root of perplexity, with everything else being equal. This rule of thumb may not always be a good predictor of performance, but it is a reasonable approximation. Note that the size of the vocabulary as such is not the primary determiner of recognition performance but rather the freedom in which the words are put together, which is represented by the grammar. A less constrained grammar, such as in the WSJ corpus, results in higher error rates.

### **Speaker-Dependent vs. Speaker-Independent Recognition**

The terms speaker-dependent (SD) and speaker-independent (SI) recognition are often used to describe different modes of operation of a speech recognition system. SD recognition refers to the case when a single speaker is used to train the system and the same speaker is used to test the system. SI recognition refers to the case where the test speaker is not included in the training. HMM-based systems can operate in either SD or SI mode, depending on the training data used. In SD mode training speech is collected from a single speaker only, while in SI mode training speech is collected from a variety of speakers.

SD and SI modes of recognition can be compared in terms of the word error rates for a given amount of training. A general rule of thumb is that, if the total amount of training speech is fixed at some level, the SI word error rates are about four times the SD error rates. Another way of stating this rule of thumb is that, for SI recognition to have the same performance as SD recognition, requires about 15 times the amount of training data (Schwartz et al., 1993). These results were obtained when one hour of speech was used to compute the SD models. However, in the limit, as the amount of training speech for SD and SI models is made larger and larger, it is not clear that any amount of training data will allow SI performance to approach SD performance.

The idea behind SI recognition is that the training is done once, after which any speaker can use the system with good performance. SD recognition is seen as an inconvenience for potential users. How

ever, one must keep in mind that SI training must be performed for each new use of a system in a different domain. If the system is used in a domain in which it was not trained, the performance degrades. It has been a historical axiom that, for optimal SI recognition performance, it is best to collect training speech from as many speakers as possible in each domain. For example, instead of collecting 100 utterances from each of 100 speakers, it was believed that it is far superior to collect, say, 10 utterances from each of 1000 speakers. Recent experiments have shown that, for some applications, collecting speech from only a dozen speakers may be sufficient for good SI performance. In an experiment with the WSJ corpus, for a fixed amount of training data, it was shown that training with 12 speakers gave basically the same SI performance as training with 84 speakers (Schwartz et al., 1993). This is a welcome result; it makes it easier to develop SI models in a new domain since collecting data from fewer speakers is cheaper and more convenient.

The ultimate goal of speech recognition research is to have a system that is domain independent (DI), that is, a system that is trained once and for all so that it can be used in any new domain and for any speaker without retraining. Currently, the only method used for DI recognition is to train the system on a very large amount of data from different domains. However, preliminary tests have shown that DI recognition on a new domain not included in the training can increase the error rate by a factor of 1.5 to 2 over SI recognition when trained on the new domain, assuming that the grammar comes from the new domain (Hon, 1992). If a good grammar is not available from the new domain, performance can be several times worse.

### Adaptation

It is possible to improve the performance of an SI or DI system by incrementally adapting to the voice of a new speaker as the speaker uses the system. This would be especially needed for atypical speakers with high error rates who might otherwise find the system unusable. Such speakers would include speakers with unusual dialects and those for whom the SI models simply are not good models of their speech. However, incremental adaptation could require hours of usage and a lot of patience from the new user before the performance becomes adequate.

A good solution to the atypical speaker problem is to use a method known as rapid speaker adaptation. In this method only a small amount of speech (about two minutes) is collected from the new speaker before using the system. By having the same utterances collected

previously from one or more prototype speakers, methods have been developed for deriving a speech model for the new speaker through a simple transformation on the speech model of the prototype speakers (Furui, 1989; Kubala and Schwartz, 1990; Nakamura and Shikano, 1989). It is possible with these methods to achieve average SI performance for speakers who otherwise would have several times the error rate.

One salient example of atypical speakers are nonnative speakers, given that the SI system was trained on only native speakers. In a pilot experiment where four nonnative speakers were tested in the RM domain in SI mode, there was an eight-fold increase in the word error rate over that of native speakers! The four speakers were native speakers of Arabic, Hebrew, Chinese, and British English. By collecting two minutes of speech from each of these speakers and using rapid speaker adaptation, the average word error rate for the four speakers decreased five-fold.

### **Adding New Words**

Out-of-vocabulary words cause recognition errors and degrade performance. There have been very few attempts at automatically detecting the presence of new words, with limited success (Asadi et al., 1990). Most systems simply do not do anything special to deal with the presence of such words.

After the user realizes that some of the errors are being caused by new words and determines what these words are, it is possible to add them to the system's vocabulary. In word-based recognition, where whole words are modeled without having an intermediate phonetic stage, adding new words to the vocabulary requires specific training of the system on the new words (Bahl et al., 1988). However, in phonetically based recognition, such as the phonetic HMM approach presented in this paper, adding new words to the vocabulary can be accomplished by including their phonetic pronunciations in the system's lexicon. If the new word is not in the lexicon, a phonetic pronunciation can be derived from a combination of a transcription and an actual pronunciation of the word by the speaker (Bahl et al., 1990a). The HMMs for the new words are then automatically compiled from the preexisting phonetic models, as shown in [Figure 6](#). The new words must also be added to the grammar in an appropriate manner.

Experiments have shown that, without additional training for the new words, the SI error rate for the new words is about twice that with training that includes the new words. Therefore, user-specified vocabulary and grammar can be easily incorporated into a speech

recognition system at a modest increase in the error rate for the new words.

## REAL-TIME SPEECH RECOGNITION

Until recently, it was thought that to perform high-accuracy, real-time, continuous speech recognition for large vocabularies would require either special-purpose VLSI hardware or a multiprocessor. However, new developments in search algorithms have sped up the recognition computation at least two orders of magnitude, with little or no loss in recognition accuracy (Austin et al., 1991; Bahl et al., 1990b; Ney, 1992; Schwartz and Austin, 1991; Soong and Huang, 1991). In addition, computing advances have achieved two-orders-magnitude increase in workstation speeds in the past decade. These two advances have made software-based, real-time, continuous speech recognition a reality. The only requirement is that the workstation must have an A/D converter to digitize the speech. All the signal processing, feature extraction, and recognition search is then performed in software in real-time on a single-processor workstation.

For example, it is now possible to perform a 2000-word ATIS task in real-time on workstations such as the Silicon Graphics Indigo R3000 or the Sun SparcStation 2. Most recently, a 20,000-word WSJ continuous dictation task was demonstrated in real-time (Nguyen et al., 1993) on a Hewlett-Packard 735 workstation, which has about three times the power of an SGI R3000. Thus, the computation grows much slower than linear with the size of the vocabulary.

The real-time feats just described have been achieved at a relatively small cost in word accuracy. Typically, the word error rates are less than twice those of the best research systems.

The most advanced of these real-time demonstrations have not as yet made their way to the marketplace. However, it is possible today to purchase products that perform speaker-independent, continuous speech recognition for vocabularies of a few thousand words. Dictation of large vocabularies of about 30,000 words is available commercially, but the speaker must pause very briefly between words, and the system adapts to the voice of the user to improve performance. For more on some of the available products and their applications, the reader is referred to other papers in this volume.

## ALTERNATIVE MODELS

HMMs have proven to be very good for modeling variability in time and feature space and have resulted in tremendous advances in

continuous speech recognition. However, some of the assumptions made by HMMs are known not to be strictly true for speech—for example, the conditional independence assumptions in which the probability of being in a state is dependent only on the previous state and the output probability at a state is dependent only on that state and not on previous states or previous outputs. There have been attempts at ameliorating the effects of these assumptions by developing alternative speech models. Below, we describe briefly some of these attempts, including the use of segmental models and neural networks. In all these attempts, however, significant computational limitations have hindered the full exploitation of these methods and have resulted only in relatively small improvements in performance so far.

### Segmental Models

Phonetic segmental models form a model of a whole phonetic segment, rather than model the sequence of frames as with an HMM. Segmental models are not limited by the conditional assumption of HMMs because, in principle, they model dependencies among all the frames of a segment directly. Furthermore, segmental models can incorporate various segmental features in a straightforward manner, while it is awkward to include segmental features in an HMM. Segmental features include any measurements that are made on the whole segment or parts of a segment, such as the duration of a segment.

There have been few segmental models proposed, among them stochastic segment models and segmental neural networks (to be described in the next section). Stochastic segment models view the sequence of feature vectors in a phonetic segment as a single long feature vector (Ostendorf et al., 1992). The major task is then to estimate the joint probability density of the elements in the feature vector. However, because the number of frames in a segment is variable, it is important first to normalize the segment to a fixed number of frames. Using some form of interpolation, typically quasi-linear, a fixed number of frames are generated that together form the single feature vector whose probability distribution is to be estimated. Typically, the distribution is assumed to be multidimensional Gaussian and its parameters are estimated from training data. However, because of the large size of the feature vector and the always limited amount of training data available, it is not possible to obtain good estimates of all parameters of the probability distribution. Therefore, assumptions are made to reduce the number of parameters to be estimated.

Because segmental models model a segment directly, a segmentation of the speech into phonetic segments must be performed prior to

modeling and recognition. In theory one would try all possible segmentations, compute the likelihood of each segmentation, and choose the one that results in the largest likelihood given the input speech. However, to try all possible segmentations is computationally prohibitive on regular workstations. One solution has been to use an HMM-based system to generate likely candidates of segmentations, which are then rescored with the segment models.

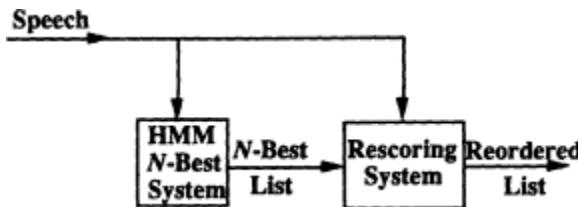


FIGURE 8 N-best paradigm for combining knowledge sources.

Figure 8 shows the basic idea of what is known as the *N-Best Paradigm* (Ostendorf et al., 1991; Schwartz and Chow, 1990). First, an HMM-based recognition system is used to generate not only the top-scoring hypothesis but also the top  $N$ -scoring hypotheses. Associated with each hypothesis is a sequence of words and phonemes and the corresponding segmentation. For each of these  $N$  different segmentations and labelings, a likelihood is computed from the probability models of each of the segments. The individual segmental scores are then combined to form a score for the whole hypothesis. The hypotheses are then reordered by their scores, and the top-scoring hypothesis is chosen. Typically, the segmental score is combined with the HMM score to improve performance.

Using the  $N$ -best paradigm with segmental models, with  $N = 100$ , has reduced word error rates by as much as 20 percent. The  $N$ -best paradigm has also been useful in reducing computation whenever one or more expensive knowledge sources needs to be combined, for example, cross-word models and  $n$ -gram probabilities for  $n > 2$ .  $N$ -best is a useful paradigm as long as the correct sentence has a high probability of being among the top  $N$  hypotheses. Thus far the paradigm has been shown to be useful for vocabularies up to 5,000 words, even for relatively long sentences.

### Neural Networks

Whatever the biological motivations for the development of "artificial neural networks" or neural nets (Lippmann, 1987), the utility of

neural nets is best served by understanding their mathematical properties (Makhoul, 1991). We view a neural net as a network of interconnected simple nonlinear computing units and the output of a neural net as just a complex nonlinear function of its inputs. Figure 9 shows a typical feedforward neural network, that is, it has no feedback elements. Although many different types of neural nets have been proposed, the type of network shown in Figure 9 is used by the vast majority of workers in this area. Shown in the figure is a three-layer network, with each layer consisting of inputs, a number of nodes, and interconnecting weights. (The term "hidden" has been used to describe layers that are not connected directly to the output.) All the nodes are usually identical in structure as shown in the figure. The inputs  $u$  to a node are multiplied by a set of weights  $v$  and summed to form a value  $z$ . A nonlinear function of  $z$ ,  $g(z)$ , is then computed. Figure 9 shows one of the most typical nonlinear functions used, that of a sigmoid. The output  $y$  of the network is then a nonlinear function of the input vector. In general, the network may have a vector of outputs as well.

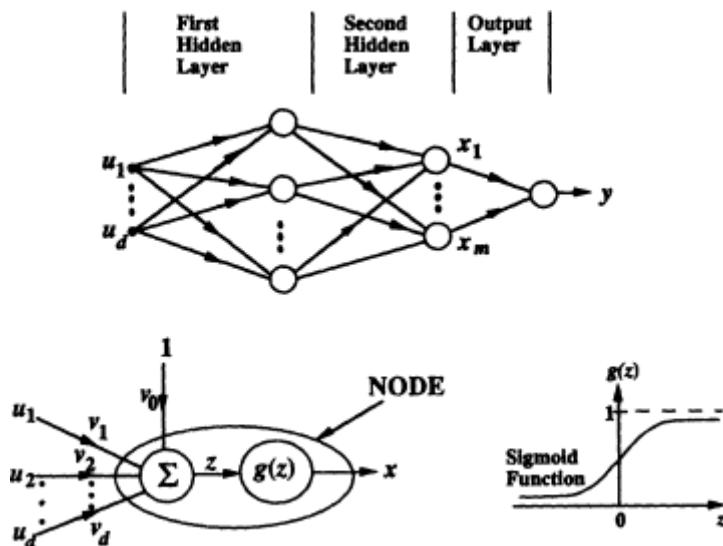


FIGURE 9 Feedforward neural network.

There are two important mathematical properties of neural nets that form the cornerstones upon which successful applications have been developed. The first is a function approximation property: it

has been shown that a two-layer neural net is capable of approximating any function arbitrarily closely in any finite portion of the input space (Cybenko, 1989). One could use more than two layers for the sake of parsimony, but this property says that two layers are sufficient (with a possibly large number of nodes in the hidden layer) to approximate any function of the input. For applications where some nonlinear function of the input is desired, the neural net can be trained to minimize, for example, the mean squared error between the actual output and the desired output. Iterative nonlinear optimization procedures, including gradient descent methods, can be used to estimate the parameters of the neural network (Rumelhart et al., 1986).

The second important property of neural nets relates to their use in classification applications: a neural net can be trained to give an estimate of the posterior probability of a class, given the input. One popular method for training the neural net in this case is to perform a least squares minimization where the desired output is set to 1 when the desired class is present at the input and the desired output is set to 0 otherwise. One can show that by performing this minimization the output will be a least squares estimate of the probability of the class given the input (White, 1989). If the classification problem deals with several classes, a network is constructed with as many outputs as there are classes, and, for a given input, the class corresponding to the highest output is chosen.

As mentioned above, estimating the parameters of a neural net requires a nonlinear optimization procedure, which is very computationally intensive, especially for large problems such as continuous speech recognition.

Neural nets have been utilized for large-vocabulary, continuous speech recognition in two ways:

- They have been used to model the output probability density for each state in an HMM (Renals et al., 1992).
- Segmental neural nets have been used to model phonetic segments directly by computing the posterior probability of the phoneme given the input (Austin et al., 1992).

In both methods the neural net system is combined with the HMM system to improve performance. In the case of segmental neural nets, the N-best paradigm is used to generate likely segmentations for the network to score. Using either method, reductions in word error rate by 10 to 20 percent have been reported. Other neural net methods have also been used in various continuous speech recognition experiments with similar results (Hild and Waibel, 1993; Nagai et al., 1993).

## CONCLUDING REMARKS

We are on the verge of an explosion in the integration of speech recognition in a large number of applications. The ability to perform software-based, real-time recognition on a workstation will no doubt change the way people think about speech recognition. Anyone with a workstation can now have this capability on their desk. In a few years, speech recognition will be ubiquitous and will enter many aspects of our lives. This paper reviewed the technologies that made these advances possible.

Despite all these advances, much remains to be done. Speech recognition performance for very large vocabularies and larger perplexities is not yet adequate for useful applications, even under benign acoustic conditions. Any degradation in the environment or changes between training and test conditions causes a degradation in performance. Therefore, work must continue to improve robustness to varying conditions: new speakers, new dialects, different channels (microphones, telephone), noisy environments, and new domains and vocabularies. What will be especially needed are improved mathematical models of speech and language and methods for fast adaptation to new conditions.

Many of these research areas will require more powerful computing resources—more workstation speed and more memory. We can now see beneficial utility for a two-orders-magnitude increase in speed and in memory. Fortunately, workstation speed and memory will continue to grow in the years to come. The resulting more powerful computing environment will facilitate the exploration of more ambitious modeling techniques and will, no doubt, result in additional significant advances in the state of the art.

## REFERENCES

- Asadi, A., R. Schwartz, and J. Makhoul, "Automatic Detection of New Words in a Large-Vocabulary Continuous Speech Recognition System," IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, pp. 125128, April 1990.
- Austin, S., R. Schwartz, and P. Placeway, "The Forward-Backward Search Algorithm," IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, pp. 697-700, 1991.
- Austin, S., G. Zavaliagkos, J. Makhoul, and R. Schwartz, "Speech Recognition Using Segmental Neural Nets," IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 1-625-628, March 1992.
- Bahl, L. R., F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. Pat. Anal. Mach. Intell., Vol. PAMI-5, No. 2, pp. 179-190, March 1983.

- Bahl, L., P. Brown, P. de Souza, R. Mercer, and M. Picheny, "Acoustic Markov Models used in the Tangora Speech Recognition System," IEEE International Conference on Acoustics, Speech, and Signal Processing , New York, pp. 497-500, April 1988.
- Bahl, L., S. Das, P. deSouza, M. Epstein, R. Mercer, B. Merialdo, D. Nahamoo, M. Picheny, and J. Powell, "Automatic Phonetic Baseform Determination," Proceedings of the DARPA Speech and Natural Language Workshop, Hidden Valley, Pa., Morgan Kaufmann Publishers, pp. 179-184, June 1990a.
- Bahl, L., P. de Souza, P. S. Gopalakrishnan, D. Kanevsky, and D. Nahamoo, "Constructing Groups of Acoustically Confusable Words," IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, pp. 85-88, April 1990b.
- Baker, J. K., "Stochastic Modeling for Automatic Speech Understanding," in Speech Recognition, R. Reddy, Ed., Academic Press, New York, pp. 521-542, 1975.
- Bates, M., R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwartz, and D. Stallard, "The BBN/HARC Spoken Language Understanding System," IEEE International Conference on Acoustics, Speech, and Signal Processing , Minneapolis, pp. 11-111-114, April 1993.
- Baum, L. E., and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model of Ecology," Am. Math. Soc. Bull., Vol. 73, pp. 360-362, 1967.
- Bellegarda, J. R., and D. H. Nahamoo, "Tied Mixture Continuous Parameter Models for Large-Vocabulary Isolated Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, pp. 13-16, May 1989.
- Cardin, R., Y. Normandin, and E. Millien, "Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, pp. 11-243246, April 1993.
- Chow, Y. L., R. M. Schwartz, S. Roucos, O. A. Kimball, P. J. Price, G. F. Kubala, M. O. Dunham, M. A. Krasner, and J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System," IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, pp. 1593-1596, April 1986.
- Cohen, J., "Application of an Auditory Model to Speech Recognition," J. Acoust. Soc. Am., Vol. 85, No. 6, pp. 2623-2629, June 1989.
- Cybenko, G. "Approximation by Superpositions of a Sigmoidal Function," Math. Contr. Signals Sys., pp. 303-314, Aug. 1989.
- Davis, S., and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-28, No. 4, pp. 357-366, August 1980.
- Della Pietra, S., V. Della Pietra, R. Mercer, and S. Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 1-633-636, March 1992.
- Forney, G. D., "The Viterbi Algorithm," Proc. IEEE, Vol. 61, pp. 268-278, 1973.
- Furui, S., "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics," IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, pp. 1991-1994, 1986.
- Furui, S., "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, paper S6.9, May 1989.

- Gauvain, J. L., and C-H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Comm.*, Vol. 11, Nos. 2-3, 1992.
- Gupta, V. N., M. Lennig, and P. Mermelstein, "Integration of Acoustic Information in a Large-Vocabulary Word Recognizer," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, pp. 697-700, April 1987.
- Haeb-Umbach, R., D. Geller, and H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, pp. 11-239-242, April 1993.
- Hild, H., and A. Waibel, "Multi-Speaker/Speaker-Independent Architectures for the Multi-State Time Delay Neural Network," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, pp. 11-255-258, April 1993.
- Hon, H. W., "Vocabulary-Independent Speech Recognition: The VOCIND System," Doctoral Thesis, Carnegie-Mellon University, Pittsburgh, Pa., 1992.
- Huang, X. D., Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, 1990.
- Huang, X. D., K. F. Lee, H. W. Hon, and M-Y. Hwang, "Improved Acoustic Modeling with the SPHINX Speech Recognition System," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, Vol. SI, pp. 345-347, May 1991.
- Hunt, M., S. Richardson, D. Bateman, and A. Piau, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 881-884, May 1991.
- Jelinek, F., L. R. Bahl, and R. L. Mercer, "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," *IEEE Trans. Info. Theory*, Vol. 21, No. 3, pp. 250-256, May 1975.
- Katz, S., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 35, No. 3, pp. 400-401, March 1987.
- Kubala, F., and R. Schwartz, "Improved Speaker Adaptation Using Multiple Reference Speakers," *International Conference on Speech and Language Processing*, Kobe, Japan, pp. 153-156, Nov. 1990.
- Lee, K.-F., *Automatic Speech Recognition: The Development of the Sphinx System*, Kluwer Academic Publishers, 1989.
- Leonard, R. G., "A Database for Speaker-Independent Digit Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Diego, paper 42.11, March 1984.
- Levinson, S. E., L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell Sys. Tech. J.*, Vol. 62, No. 4, pp. 1035-1073, April 1983.
- Lippmann, R. P., "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, pp. 4-22, April 1987.
- Lowerre, B. T., "The Harpy Speech Recognition System," Doctoral Thesis, Carnegie-Mellon University, Pittsburgh, Pa., 1976.
- MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, Harriman, N.Y., Morgan Kaufmann Publishers, pp. 7-14, Feb. 1992.
- Makhoul, J., "Pattern Recognition Properties of Neural Networks," *Neural Networks*

- for Signal Processing—Proceedings of the 1991 IEEE Workshop, IEEE Press, New York, pp. 173-187, 1991.
- Makhoul, J., S. Roucos, and H. Gish, "Vector Quantization in Speech Coding," Proc. IEEE, Vol. 73, No. 11, pp. 1551-1588, Nov. 1985.
- Nagai, A., K. Yamaguchi, S. Sagayama, and A. Kurematsu, "ATREUS: A Comparative Study of Continuous speech Recognition Systems at ATR," IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, pp. 11-139142, April 1993.
- Nakamura, S., and K. Shikano, "Speaker Adaptation Applied to HMM and Neural Networks," IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, paper S3.3, May 1989.
- Ney, H., "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 1-9-12, March 1992.
- Nguyen, L., R. Schwartz, F. Kubala, and P. Placeway, "Search Algorithms for Software-Only Real-Time Recognition with Very Large Vocabularies," Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufmann Publishers, Princeton, N.J., pp. 91-95, March 1993.
- Ostendorf, M., A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek, "Integration of Diverse Recognition Methodologies through Reevaluation of *N*-Best Sentence Hypotheses," Proceedings of the DARPA Speech and Natural Language Workshop, Monterey, Calif., Morgan Kaufmann Publishers, pp. 83-87, February 1991.
- Ostendorf, M., I. Bechwati, and O. Kimball, "Context Modeling with the Stochastic Segment Model," IEEE International Conference on Acoustics, Speech, and Signal Processing , San Francisco, pp. 1-389-392, March 1992.
- Pallett, D., J. Fiscus, W. Fisher, and J. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program," Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufmann Publishers, Princeton, N.J., pp. 7-18, March 1993.
- Paul, D., "The Design for the Wall Street Journal-based CSR Corpus," Proceedings of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, pp. 357-360, Feb. 1992.
- Placeway, P., R. Schwartz, P. Fung, and L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora," IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, pp. 11-33-36, April 1993.
- Price, P., W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, New York, pp. 651-654, April 1988.
- Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989.
- Rabiner, L. R., and B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.
- Renals, S., N. Morgan, M. Cohen, and H. Franco, "Connectionist Probability Estimation in the Decipher Speech Recognition System," IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 1-601-603, March 1992.
- Rumelhart, D., C. Hinton, and R. Williams, "Learning Representations by Error Propagation," in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, D. Rumelhart and J. McClelland (eds.), MIT Press, Cambridge, Mass., Vol. 1, pp. 318-362, 1986.

- Schwartz, R., and S. Austin, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, pp. 701-704, 1991.
- Schwartz, R., and Y. L. Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the  $N$  Most Likely Sentence Hypotheses," IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, pp. 81-84, April 1990.
- Schwartz, R. M., Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, pp. 35.6.1-35.6.4, March 1984.
- Schwartz, R., O. Kimball, F. Kubala, M. Feng, Y. Chow, C. Barry, and J. Makhoul, "Robust Smoothing Methods for Discrete Hidden Markov Models," IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, paper S10b.9, May 1989.
- Schwartz, R., A. Anastasakos, F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparative Experiments on Large-Vocabulary Speech Recognition," Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufmann Publishers, Princeton, N.J., pp. 75-80, March 1993.
- Soong, F., and E. Huang, "A Tree-Trellis Based Fast Search for Finding the  $N$  Best Sentence Hypotheses in Continuous Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, pp. 705-708, 1991.
- White, H., "Learning in Artificial Neural Networks: A Statistical Perspective," Neural Computation, pp. 425-464, 1989.

# Training and Search Methods for Speech Recognition

*Frederick Jelinek*

## SUMMARY

Speech recognition involves three processes: extraction of acoustic indices from the speech signal, estimation of the probability that the observed index string was caused by a hypothesized utterance segment, and determination of the recognized utterance via a search among hypothesized alternatives. This paper is not concerned with the first process.

Estimation of the probability of an index string involves a model of index production by any given utterance segment (e.g., a word). Hidden Markov models (HMMs) are used for this purpose (Makhoul and Schwartz, this volume). Their parameters are state transition probabilities and output probability distributions associated with the transitions. The Baum algorithm that obtains the values of these parameters from speech data via their successive reestimation will be described in this paper.

The recognizer wishes to find the most probable utterance that could have caused the observed acoustic index string. That probability is the product of two factors: the probability that the utterance will produce the string and the probability that the speaker will wish to produce the utterance (the language model probability).

Even if the vocabulary size is moderate, it is impossible to search for the utterance exhaustively. One practical algorithm is described (Viterbi) that, given the index string, has a high likelihood of finding the most probable utterance.

## INTRODUCTION

It was pointed out by Makhoul and Schwartz (this volume) that the problem of speech recognition can be formulated most effectively as follows:

Given observed acoustic data  $A$ , find the word sequence  $\hat{W}$  that was the most likely cause of  $A$ .

The corresponding mathematical formula is:

$$\hat{W} = \arg \max P(A \mid W)P(W) \quad (1)$$

$W$

$P(W)$  is the a priori probability that the user will wish to utter the word sequence  $W = w_1, w_2, \dots, w_n$  ( $w_i$  denotes the individual words belonging to some vocabulary  $V$ ).  $P(A \mid W)$  is the probability that if  $W$  is uttered, data  $A = a_1, a_2, \dots, a_k$  will be observed (Bahl et al., 1983).

In this simplified presentation the elements  $a_i$  are assumed to be symbols from some finite alphabet  $A$  of size  $|A|$ . Methods of transforming the air pressure process (speech) into the sequence  $A$  are of fundamental interest to speech recognition but not to this paper. From my point of view, the transformation is determined and we live with its consequences.

It has been pointed out elsewhere that the probabilities  $P(A \mid W)$  are computed on the basis of a hidden Markov model (HMM) of speech production that, in principle, operates as follows: to each word  $l$  of vocabulary  $V$ , there corresponds an HMM of speech production. A concrete example of its structure is given in [Figure 1](#). The model of speech production of a sequence of words  $W$  is a concatenation of models of individual words  $w_i$  making up the sequence  $W$  (see [Figure 2](#)).

We recall that the HMM of [Figure 1](#) starts its operation in the initial state  $S_I$  and ends it when the final state  $S_F$  is reached. A transi



FIGURE 1 Structure of a hidden Markov model for a word.

tion out of state  $s$  into a next state is either performed in a fixed unit of time along a solid arc and results in an output  $a$  from the alphabet  $A$ , or is performed instantaneously along an interrupted arc (a *null* transition) and results in no output.

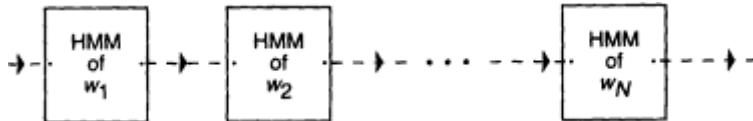


FIGURE 2 Schematic diagram of a hidden Markov model for a sequence of words.

It is assumed that the different HMMs corresponding to various words  $v$  have the same transition structure (e.g., that of Figure 1) and that they differ from each other in the number of states they have, in the values of the probabilities  $p(t)$  of the interstate transitions, and in the distributions  $q(a \mid t)$  of generating output  $a$  when the (nonnull) transition  $t$  is taken.

The first task, discussed in the second and third sections of this paper, is to show how the values of the parameters  $p(t)$  and  $q(a \mid t)$  may be estimated from speech data, once the structure of the models (including the number of states) is selected by the system designer.

Returning to formula (1), we are next interested in modeling  $P(W)$ . We will assume that

$$P(W) = \prod_{i=1}^n P[w_i \mid \phi(w_1, w_2, \dots, w_{i-1})]$$

where  $\phi(w_1, w_2, \dots, w_{i-1})$  denotes the equivalence class (for the purpose of predicting the next word  $w_i$ ) to which the history  $w_1, w_2, \dots, w_{i-1}$  belongs. Essentially without loss of generality we will consider only a finite alphabet  $f$  of equivalence classes, so that

$$\phi(w_1, w_2, \dots, w_{i-1}) = \phi_i \in \Phi$$

A popular classifier example is:

$$\phi(w_1, w_2, \dots, w_{i-1}) = w_{i-1}$$

(the *bigram* model) or

$$\phi(w_1, w_2, \dots, w_{i-1}) = w_{i-2}, w_{i-1}$$

(the *trigram* model). In this paper it is assumed that the equivalence

classifier  $\varphi$  was selected by the designer, who also estimated the *language model* probabilities  $P(v \mid \varphi)$  for all words  $v \in V$  and classes  $\varphi \in \Phi$  (Such estimation is usually carried out by processing a large amount of appropriately selected text.)

The second task of this article, addressed in the third and fourth sections, is to show one possible way to search for the recognizer sequence  $W$  that maximizes the product  $P(A \mid W) P(W)$  [see formula (1)].

This tutorial paper is not intended to be a survey of available training and search methods. So-called Viterbi training (Rabiner and Juang, 1993) is useful in special circumstances. As to search, various generalizations of the Stack algorithm (Jelinek, 1969) are very efficient and have, in addition, optimality properties that the beam search presented here does not possess. However, these methods (Bahl et al., 1993; Paul, 1992;) are quite difficult to explain and so are not presented here.

### ESTIMATION OF STATISTICAL PARAMETERS OF HMMS

We will arrive at the required algorithm (known in the literature variously as the Baum, or Baum-Welch, or Forward-Backward algorithm) by intuitive reasoning. Proofs of its convergence, optimality, or other characteristics can be found in many references (e.g., Baum, 1972) and will be omitted here.

It is best to gain the appropriate feeling for the situation by means of a simple example that involves the basic phenomenon we wish to treat. We will consider the HMM of [Figure 3](#) that produces binary sequences. There are three states and six transitions,  $t_3$  being a null transition. The transition probabilities satisfy constraints

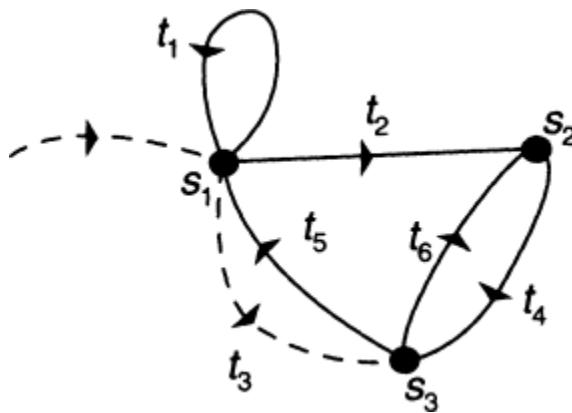


FIGURE 3 Sample three-state hidden Markov model.

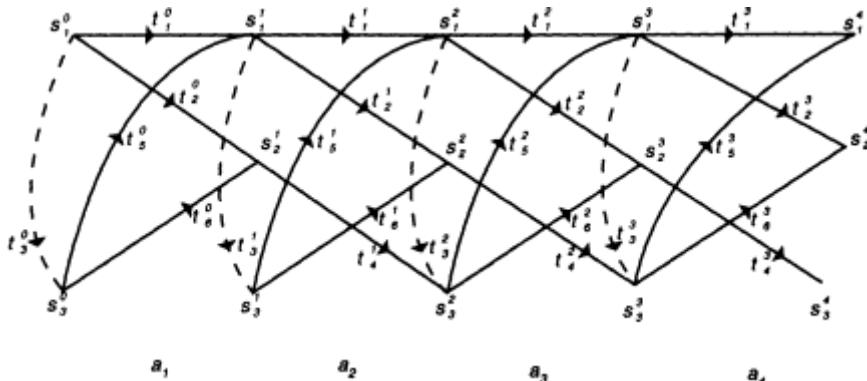


FIGURE 4 Trellis for the hidden Markov model of Figure 3.

$$p(t_1) + p(t_2) + p(t_3) = 1; p(t_4) = 1; p(t_5) + p(t_6) = 1 \quad (6)$$

We will wish to estimate the probabilities  $p(t_i)$  and  $q(a \mid t_i)$ , where  $a \in \{0,1\}$  and  $i \in \{1, \dots, 6\}$ . We will designate  $s_1$  as the starting state. There is no natural final state.

Suppose we knew that the HMM produced the (long) sequence  $a_1, a_2, \dots, a_k$ . How would we estimate the required parameters?

A good way to see the possible operation of the HMM is to develop it in time by means of a *lattice*, such as the one in Figure 4, corresponding to the production of  $a_1, a_2, a_3, a_4$  by the HMM of Figure 3. The lattice has *stages*, one for each time unit, each stage containing all states of the basic HMM. The states of successive stages are connected by the nonnull transitions of the HMM because they take a unit of time to complete. The states within a stage are connected by the null transitions because these are accomplished instantaneously. The starting, 0<sup>th</sup> stage contains only the starting state ( $s_1$ ) and those states that can be reached from it instantaneously ( $s_3$ , connected to  $s_1$  by a null transition). In Figure 4 superscripts have been added to the transitions and states to indicate the stage from which the transitions originate where the states are located.

We now see that the output sequence can be produced by any path leading from  $s_1$  in the 0<sup>th</sup> stage to any of the states in the final stage.

Suppose we knew the actual transition sequence  $T = t_1, t_2, \dots, t_n$  ( $n \geq k$ , because some transitions may be null transitions and  $k$  outputs were generated) that caused the observed output. Then the so-called maximum likelihood parameter estimate could be obtained by counting. Define the indicator function,

$$I(t^i, t) = \begin{cases} 1 & \text{if one of the } i^{\text{th}} \text{ stage transitions in } T \text{ is } t \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

and using it,

$$c(t) = \sum_{i=0}^{k-1} I(t^i, t) \quad (8)$$

for all transitions  $t$ , as well as

$$c(a, t') = \sum_{i=0}^{k-1} I(t^i, t') \delta(a_{i+1}, a) \quad (9)$$

for nonnull transitions  $t'$ . In Equation (9),  $\delta( , )$  denotes the Kronecker delta function.  $c(t)$  is then the number of times the process went through transition  $t$ , and  $c(a, t)$  is the number of times the process went through  $t$  and produced the output  $a$ .

The "natural" estimates of the probability parameters then would be

$$\hat{q}(a | t) = \frac{c(a, t)}{c(t)} \quad t \text{ nonnull} \quad (10)$$

$$\hat{p}(t_i) = \begin{cases} \frac{c(t_i)}{c(t_1) + c(t_2) + c(t_3)} & i = 1, 2, 3 \\ 1 & i = 4 \\ \frac{c(t_i)}{c(t_5) + c(t_6)} & i = 5, 6 \end{cases} \quad (11)$$

Of course, the idea of us knowing the transition sequence  $t_1, t_2, \dots, t_n$  is absurd. But what if we knew the probabilities  $P\{t^i = t\}$  that transition  $t$  was taken out of the  $i^{\text{th}}$  stage of the process? Then it would seem intuitively reasonable to define the "counts" by ( $t'$  is restricted to nonnull transitions)

$$c(t) = \sum_{i=0}^{k-1} P\{t^i = t\} \quad (8')$$

$$c(a, t') = \sum_{i=0}^{k-1} P\{t^i = t'\} \delta(a_{i+1}, a) \quad (9')$$

and use these in the estimation formulas (10) and (11).

Since the production of each output  $a_i$  corresponds to some nonnull

transition from stage  $i - 1$  to stage  $i$ , then  $\sum P\{t^i = t''\} = 1$ , where the sum is over all nonnull transitions  $t''$ . Thus, while in case (8), the counter of exactly one of the nonnull transitions is increased by 1 for each output, this same contribution is simply distributed by (8') among the various counters belonging to nonnull transitions.

The estimates arising from (8') and (9') will be convenient only if it proves to be practically sensible to compute the probabilities  $P\{t^i = t\}$ . To derive the appropriate method, we need a more precise notation:

---

$L(t)$	the initial state of transition $t$ [e.g., $L(t_0) = s_3$ ]
$R(t)$	the final state of transition $t$ [e.g., $R(t_6) = s_2$ ]
$P\{t^i = t\}$	the probability that $a_1, a_2, \dots, a_k$ was produced and the transition out of the $i^{\text{th}}$ stage was $t$
$P\{s^i = s\}$	the probability that $a_1, a_2, \dots, a_i$ was produced and the state reached at the $i^{\text{th}}$ stage was $s$
$P\{\text{rest }   s^i = s\}$	the probability that $a_{i+1}, a_{i+2}, \dots, a_k$ was produced when the state at the $i^{\text{th}}$ stage was $s$

---

Since the transition  $t$  can be taken only after the HMM reached the state  $L(t)$  and, since after it is taken, the rest of the action will start in state  $R(t)$ , we see immediately that

$$P\{t^i = t\} = \begin{cases} P\{s^i = L(t)\}p(t)q(a_{i+1} | t)P\{\text{rest } | s^{i+1} = R(t)\} & \text{if } t \text{ nonnull} \\ P\{s^i = L(t)\}p(t)P\{\text{rest } | s^i = R(t)\} & \text{if } t \text{ null} \end{cases} \quad (12)$$

It is also obvious that the following recursion holds:

$$\begin{aligned} P\{s^i = s\} &= \sum_{t \in N(s)} P\{s^{i-1} = L(t)\}p(t)q(a_i | t) \\ &\quad + \sum_{t \in N(s)} P\{s^i = L(t)\}p(t) \end{aligned} \quad (13)$$

Here  $N(s)$  is the set of all null transitions ending in  $s$ , and  $N(s)$  is the set of all nonnull transitions ending  $s$ . That is,

$$\begin{aligned} N(s) &= \{t: R(t) = s, t \text{ is null}\} \\ \overline{N}(s) &= \{t: R(t) = s, t \text{ is nonnull}\} \end{aligned} \quad (13')$$

Even though Equation (13) involves quantities  $P\{s^i = s\}$  on both sides of the equation, it can easily be evaluated whenever null transitions do not form a loop (as is the case in the example HMM), because in such a case the states can be appropriately ordered. We can also obtain a backward recursion:

$$\begin{aligned} P\{rest \mid s^i = s\} &= \sum_{t \in \overline{M}(s)} p(t) q(a_{i+1} \mid t) P\{rest \mid s^{i+1} = R(t)\} \\ &+ \sum_{t \in M(s)} p(t) P\{rest \mid s^i = R(t)\} \end{aligned} \quad (14)$$

where  $M(s)$  and  $\overline{M}(s)$  are the sets of null and nonnull transitions leaving  $s$ , respectively:

$$\begin{aligned} M(s) &= \{t: L(t) = s, t \text{ is null}\} \\ \overline{M}(s) &= \{t: L(t) = s, t \text{ is nonnull}\} \end{aligned} \quad (14')$$

Recursions (13) and (14) then lead directly to an algorithm computing the desired quantities  $P\{t^i = t\}$  via formula (12):

#### 1. The Forward Pass

Setting  $P\{s^0 = s_1\} = 1$  (in our HMM,  $s_1 = s_l$ ), use (13) to compute  $P\{s^i = s\}$  for all  $s$ , starting with  $i = 0$  and ending with  $i = k$ .

#### 2. The Backward Pass

Setting  $P\{rest \mid s^k = s\} = 1$  for all  $s$ , use (14) to compute  $P\{rest \mid s^i = s\}$ , starting with  $i = k - 1$  and ending with  $i = 0$ .

#### 3. Transition Probability Evaluation

Using formula (12), evaluate probabilities  $P\{t^i = t\}$  for all  $t$  and  $i = 0, 1, \dots, k - 1$ .

#### 4. Parameter Estimation

Use the counts (8') and (9') to get the parameter estimates

$$\hat{q}(a \mid t) = \frac{c(a, t)}{c(t)} \quad t \text{ nonnull} \quad (10)$$

$$\hat{p}(t) = \frac{c(t)}{\sum_{L(t')=L(t)} c(t')} \quad (11')$$

There is only one flaw, seemingly a fatal one, to our procedure: formulas (12), (13), and (14) use the values of the parameters  $p(t)$  and  $q(a \mid t)$  that we are trying to estimate! Fortunately, there is a good way out: we put the above algorithm into a loop, starting with a guess at  $p(t)$  and  $q(a \mid t)$ , obtaining a better estimate (this can be proved!) with (10) and (11'), plugging these back in for  $p(t)$  and  $q(a \mid t)$ , obtaining an even better estimate, etc.

## REMARKS ON THE ESTIMATION PROCEDURE

The heuristic derivation of the HMM parameter estimation algorithm of the previous section proceeded from the assumption that the observed data were actually produced by the HMM in question. Actually, the following maximum likelihood properties are valid regardless of how the data were produced:

Let  $P_\lambda(A)$  denote the probability that the HMM defined by parameters  $\lambda$  produced the observed output  $A$ . If  $\lambda'$  denotes the parameter values determined by (10) and (11') when parameters  $\lambda$  were used in (12), (13), and (14), then  $P_{\lambda'}(A) \geq P_\lambda(A)$ .

The previous section showed how to estimate HMM parameters from data. We did not, however, discuss the specific application to speech word models. We do this now.

First, it must be realized that the word models mentioned in the first section (e.g., see [Figure 1](#)) should be built from a relatively small number of building blocks that are used in many words of the vocabulary. Otherwise, training could only be accomplished using separate speech data for each and every word in the vocabulary. For instance, in the so-called fenonic case (Bahl et al., 1988), there is an alphabet of 200 elementary HMMs (see [Figure 5](#)), and a word model is specified by a string of symbols from that alphabet. The word HMM is then built out of a concatenation of the elementary HMMs corresponding to the elements of the defining string. The training data  $A = a_1, a_2, \dots, a_k$  is produced by the user, who is told to read an extended text (which, however, contains a small minority of words in the vocabulary). The corresponding HMM to be trained is a concatenation of models of words making up the text. As long as the text HMM contains a sufficient number of each of the elementary HMMs, the resulting estimation of the  $p(t)$  and  $q(a | t)$  parameters will be successful, and the HMMs



FIGURE 5 Elementary hidden Markov model for the fenonic case.

corresponding to all the words of the entire vocabulary can then be specified, since they are made up of the same elementary HMMs as those that were trained.

## FINDING THE MOST LIKELY PATH

Our second task is to find the most likely word string  $W$  given the observed data  $A$ . There are many methods to do this, but all are based on one of two fundamental methods: the Viterbi algorithm or the Stack algorithm ( $A^*$  search). In this paper the first method is used.

Consider the following basic problem:

Given a fully specified HMM, find the sequence of transitions that were the most likely "cause" of given observed data  $A$ .

To solve the problem, we again use the trellis representation of the HMM output process (see [Figure 4](#)). Let  $\pi$  be the desired most likely path through the trellis, that is, the one that is the most probable cause of  $A$ . Suppose  $\pi$  passes through some state  $s_j^i$ , and let  $\pi'$  be the initial segment of  $\pi$  ending in  $s_j^i$ , and  $\pi^*$  the final segment starting in  $s_j^i$ . Then  $\pi'$  is the most likely of all paths ending in  $s_j^i$ , because if  $\pi''$  were more likely, then  $\pi''\pi^*$  would be more likely than  $\pi = \pi'\pi^*$ , contradicting our assumption that  $\pi$  is the most likely path.

The direct consequence of this observation is that, if there are multiple paths leading into  $s_j^i$ , only the most probable of them may be the initial segment of the most likely total path. Hence, at each stage of the trellis, we need to maintain only the most probable path into each of the states of a stage; none of the remaining paths can be the initial segment of the most probable complete path.

Let  $P_m\{s^i = s\}$  be the probability of the most likely path ending in state  $s$  at stage  $i$ . We have the following recursion:

$$\begin{aligned} P_m\{s^i = s\} &= \max\left\{\max_{t \in N(s)} p(t)q(a_{i-1} | t)P_m\{s^{i-1} = L(t)\}, \right. \\ &\quad \left. \max_{t \in \bar{N}(s)} p(t)P_m\{s^i = L(t)\}\right\} \end{aligned} \quad (15)$$

where the transition sets  $N(s)$  and  $\bar{N}(s)$  were defined in (13'). The desired Viterbi algorithm (Viterbi, 1967) then is:

1. Setting  $P_m\{s^0 = s_1\} = 1$ , use (15) to compute  $P_m\{s^i = s\}$  for all  $s$ , starting with  $i = 0$  and ending with  $i = k$ . For each state  $s$  at stage  $i$  keep a pointer to the previous state along that transition in formula (15) that realized the maximum.

2. Find ( $k$  is the length of the output data)

$$\hat{s} = \arg \max_s P_m\{s^k = s\}$$

3. The trace-back from  $s$  along the pointers saved in 1 above defines the most likely path.

### DECODING: FINDING THE MOST LIKELY WORD SEQUENCE

We are now ready to describe the method of search for the most likely word sequence  $W$  to have caused the observed acoustic sequence  $A$ . This method, like all the others for large-vocabulary sizes  $N$ , will be approximate. However, it gives very good results from the practical point of view [i.e., it is very rare that the sequence  $IW$  defined by Equation (1) is the one actually spoken *and* will not be found by our search method].

First, consider the simplest kind of a language model, one in which all histories are equivalent [see Eq. (2)]. Then

$$P(W) = \prod_{i=1}^n P(w_i) \quad (16)$$

It is immediately obvious that in this case finding  $W$  amounts to searching for the most likely path through the graph of [Figure 6](#). This statement must be interpreted literally—we do not care what happens inside the HMM boxes.

The slightly more complicated bigram language model [see Eq. (4)] results in

$$P(w) = P(w_1) \prod_{i=2}^n P(w_i \mid w_{i-1}) \quad (17)$$

and  $W$  is defined by the most likely path through the graph of [Figure 7](#). Note that while [Figure 7](#) is somewhat more complicated than [Figure 6](#), the number of states is the same in both. It is proportional to the vocabulary size  $N$ . Thus, except in cases where estimates  $P(v_i \mid v_j)$  cannot be obtained, one would always search [Figure 7](#) rather than [Figure 6](#).

For a trigram language model [see Eq. (5)],

$$P(w) = P(w_1)P(w_2 \mid w_1) \prod_{i=3}^n P(w_i \mid w_{i-2}, w_{i-1}) \quad (18)$$

the graph is considerably more complex—the number of states is proportional to  $N^2$ . The situation for  $N=2$  is illustrated in [Figure 8](#).

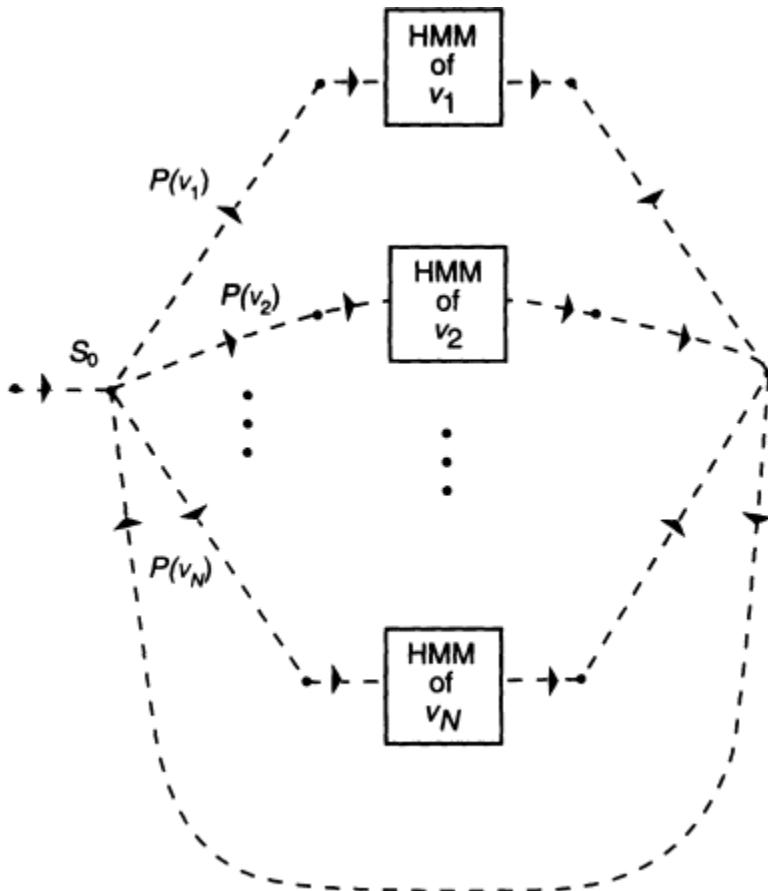


FIGURE 6 Schematic structure of the hidden Markov model for a unigram language model.

In general, similar graphs can be derived for any equivalence classifier  $\varphi$ , as long as the number of equivalence classes is finite.

How do we find the most likely paths through these graphs? There exist no practical algorithms finding the exact solution. However, if we replace the boxes in Figures 6 through 8 with the corresponding HMM models, all the figures simply represent (huge) HMMs in their own right. The most likely path through an HMM is found by the Viterbi algorithm described in the previous section!

The only problem is that for practical vocabularies (e.g.,  $N = 20,000$ ) even the HMM of Figure 7 has too many states. One solution is the so-called *beam search* (Lowerre, 1976). The idea is simple. Imagine the

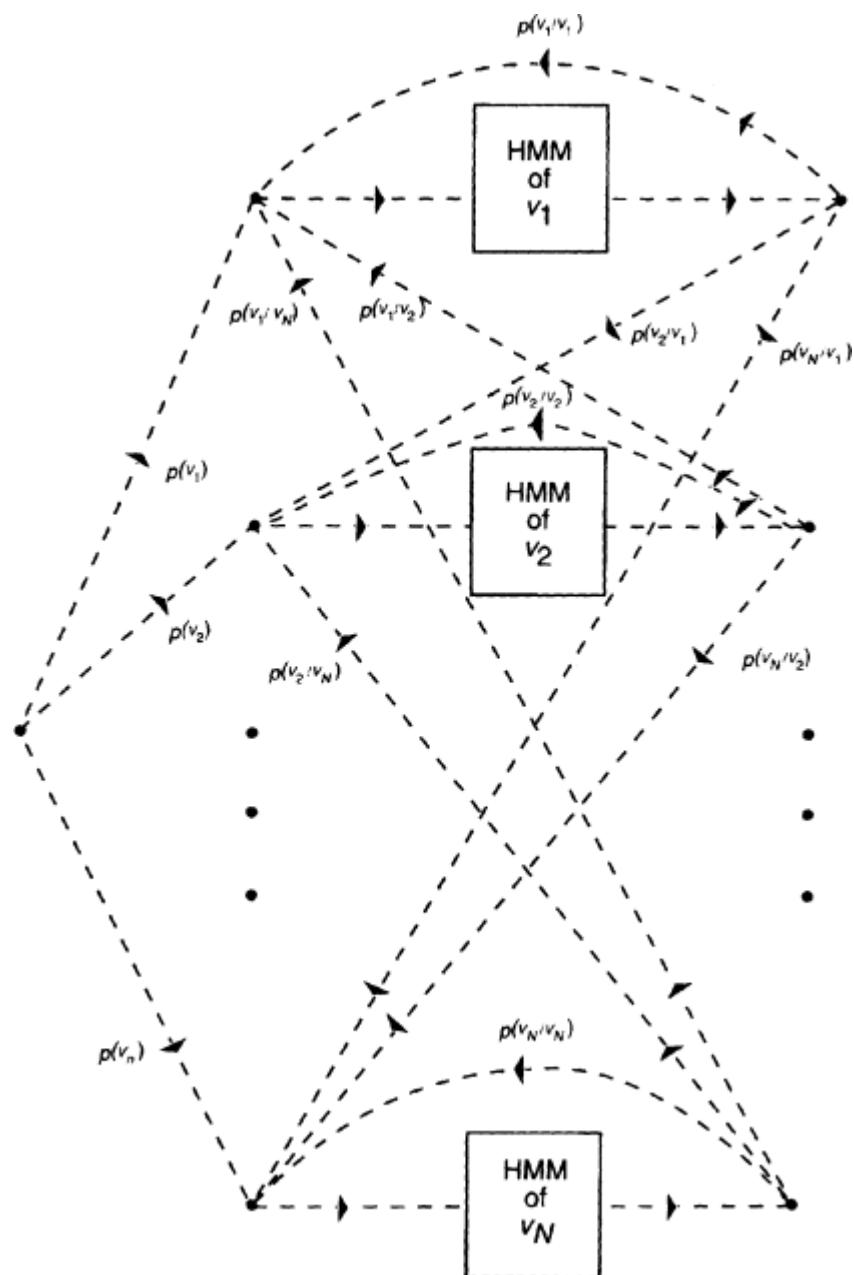


FIGURE 7 Schematic structure of the hidden Markov model for a bigram language model.

trellis (see the third section of this paper) corresponding to [Figure 7](#). Before carrying out the Viterbi *purge* (15) at stage  $i$ , we determine the maximal probability  $P_{i-1}^m$  of the states at stage  $i - 1$ .

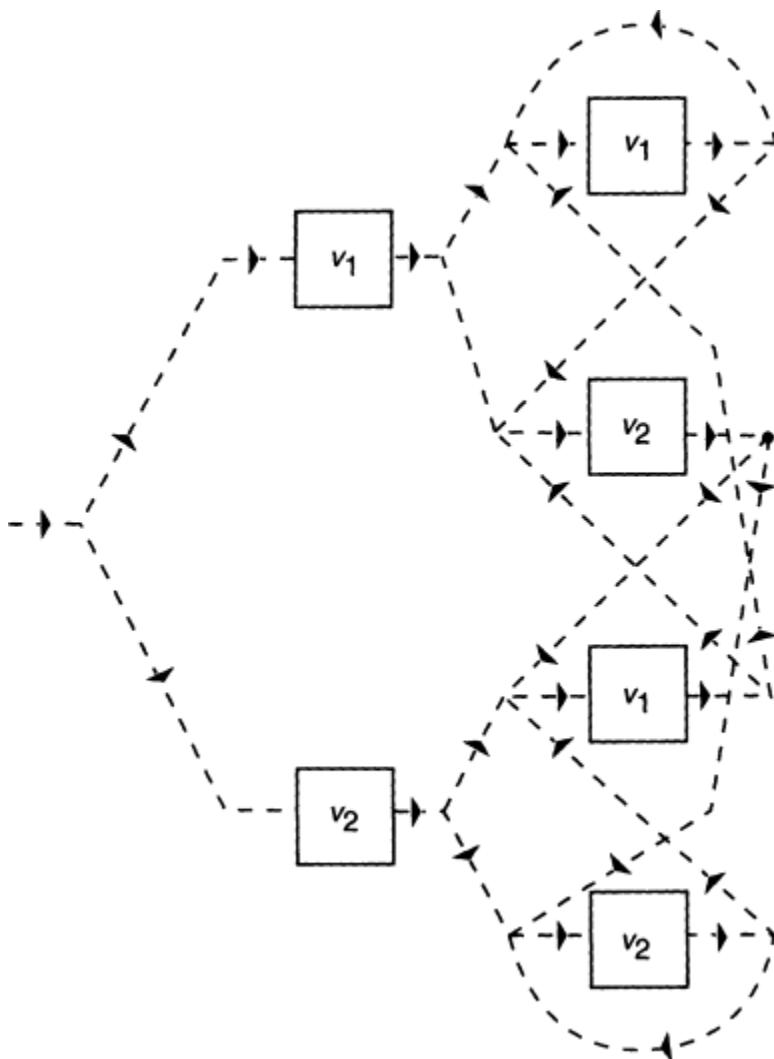


FIGURE 8 Schematic structure of the hidden Markov model for a trigram language model corresponding to a two-word vocabulary.

$$P_{i-1}^m = \max_s P_m\{s^{i-1} = s\} \quad (19)$$

We then establish a dynamic threshold

$$\tau_{i-1} = P_{i-1}^m / K \quad (20)$$

where  $K$  is a suitably chosen number. Finally, we eliminate from the trellis all states  $s$  on level  $i - 1$  such that

$$P_m(s^{i-1} = s) < \tau_{i-1} \quad (21)$$

This procedure is capable of drastically reducing the number of states entering the comparison implied by the *max* function in (15), without significantly affecting the values  $P_m\{s^i = s\}$  and makes the Viterbi algorithm a practical one, at least for the case of bigram language models.

Is there any expedient way to implement the search for  $W$  for a trigram language model? One ingenious method has recently been suggested by researchers at SRI (Murveit et al., 1993). I will describe the main idea.

The HMMs for each word in the vocabulary have a final state (see the right-most column of states in [Figure 7](#)). Final states of words therefore occur at various stages of the trellis. As the beam search proceeds, some of these states will be killed by the thresholding (21). Consider now those final states in the trellis that remain alive. The trace-back (see fourth section) from any of these final states, say, of word  $v$  at stage  $j$ , will lead to the initial state of the same word model, say at stage  $i$ . The pair  $(i, j)$  then identifies a time interval during which it is reasonable to admit the hypothesis that the word  $v$  was spoken. To each word  $v$  there will then correspond a set of time intervals:  $\{[i_1(v), j_1(v)], [i_2(v), j_2(v)], \dots [i_m(v), j_m(v)]\}$ . We can now hypothesize that word  $v'$  could conceivably follow word  $v$  if and only if there exist intervals  $[i_k(v), j_k(v)]$  and  $[i_t(v'), j_t(v')]$  such that  $i_k(v) < i_t(v')$ ,  $j_k(v) \geq i_t(v')$ , and  $j_k(v) < j_t(v')$ . We can, therefore, construct a directed graph whose intermediate states correspond to words  $v \in V$  that will have two properties: (1) Any path from the initial to the final state of the graph will pass through states corresponding to a word sequence  $w_1, w_2, \dots, w_n$ , such that  $w_i$  is permitted to follow  $w_{i-1}$  by the word interval sets. (2) Arcs leading to a state corresponding to a word  $v'$  emanate only from states corresponding to one and the same word  $v$ . ([Figure 8](#) has this property if its boxes are interpreted as states).

To the arcs of this graph we can then attach trigram probabilities  $P(w_i | w_{i-2}w_{i-1})$ , and we can expand its states by replacing them with the HMMs of the corresponding word. We can then construct a trellis for the resulting overall trigram HMM that will have only a small

fraction of states of the trellis constructed directly for the trigram language model (cf. [Figure 8](#)).

Consequently, to find  $W$  we conduct two beam searches. The first, on the bigram HMM, results in word presence time intervals. These give rise to a trigram HMM over which the second beam search is conducted for the final  $W$ .

## REFERENCES

- Bahl, L. R., F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, pp. 179-190, March 1983.
- Bahl, L., P. Brown, P. de Souza, R. Mercer, and M. Picheny, "Acoustic Markov Models used in the Tangora speech recognition system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York, April 1988.
- Bahl, L. R., P. S. Gopalakrishnan, and R. L. Mercer, "Search Issues in Large-Vocabulary Speech Recognition," *Proceedings of the IEEE Workshop on Automatic Speech Recognition*, Snowbird, Utah, 1993.
- Baum, L., "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process," *Inequalities*, vol. 3, pp. 1-8, 1972.
- Jelinek, F., "A fast sequential decoding algorithm using a stack," *IBM Journal of Research Development*, vol. 13, pp. 675-685, Nov. 1969.
- Lowerre, B. T., "The Harpy Speech Recognition System," Ph.D. Dissertation, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa., 1976.
- Murveit, H., J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Spoken Language Systems Technology Workshop*, Massachusetts Institute of Technology, Cambridge, Mass., January 1993.
- Paul, D. B., "An Essential A\* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model," *1992 International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, March 1992.
- Rabiner, L. R., and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993, pp. 378-384.
- Viterbi, A. J., "Error bounds for convolutional codes and an asymmetrically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. IT-13, pp. 260-267, 1967.

## **NATURAL LANGUAGE UNDERSTANDING TECHNOLOGY**



# The Roles of Language Processing in a Spoken Language Interface

*Lynette Hirschman*

## SUMMARY

This paper provides an overview of the colloquium's discussion session on natural language understanding, which followed presentations by Bates and Moore. The paper reviews the dual role of language processing in providing *understanding* of the spoken input and an additional source of *constraint* in the recognition process. To date, language processing has successfully provided understanding but has provided only limited (and computationally expensive) constraint. As a result, most current systems use a loosely coupled, unidirectional interface, such as N-best or a word network, with natural language constraints as a postprocess, to filter or resort the recognizer output. However, the level of discourse context provides significant constraint on what people can talk about and how things can be referred to; when the system becomes an active participant, it can influence this order. But sources of discourse constraint have not been extensively explored, in part because these effects can only be seen by studying systems in the context of their use in interactive problem solving. This paper argues that we need to study interactive systems to understand what kinds of applications are appropriate for the current state of technology and how the technology can move from the laboratory toward real applications.

## INTRODUCTION

This paper provides an overview of the natural language understanding session at the Colloquium on Human-Machine Communication by Voice held by the National Academy of Sciences (NAS). The aim of the paper is to review the role that language understanding plays in spoken language systems and to summarize the discussion that followed the two presentations by Bates and Moore. A number of questions were raised during the discussion, including whether a single system could provide both understanding and constraint, what the future role of discourse should be, how to evaluate performance on interactive systems, and whether we are moving in the right direction toward realizing the goal of interactive human-machine communication.<sup>1</sup>

### Background: The ARPA Spoken Language Program

Much of the research discussed at the natural language understanding session was done in connection with the Advanced Research Projects Agency's (ARPA) Spoken Language Systems program. This program, which started in 1989, brought together speech and language technologies to provide speech interfaces for *interactive problem solving*. The goal was to permit the user to speak to the system, which would respond appropriately, providing (intelligent) assistance. This kind of interaction requires the system to have both input and output capabilities, that is, for speech, both recognition and synthesis, and for language, both understanding and generation. In addition, the system must be able to understand user input in context and carry on a coherent conversation. We still know relatively little about this complex process of interaction, although we have made significant progress in one aspect, namely spoken language understanding.<sup>2</sup>

In the ARPA Spoken Language Systems program, multiple contractors are encouraged to develop independent approaches to the core problem of spoken language interaction. To focus the research,

---

<sup>1</sup> I am indebted to the many contributors during the colloquium's discussion who raised interesting questions or provided important material. For the sake of the flow of this paper, I have folded these questions or comments into appropriate sections, rather than summarizing the discussion separately.

<sup>2</sup> Spoken language *understanding* focuses on understanding user input, as opposed to *communicating* with the user, which is a bidirectional process that requires synthesis and generation technologies.

*common evaluation* is used to compare alternate technical approaches within a common task domain. To ensure that test results are comparable across sites, the sites choose a task domain, collect a common corpus of training material, and agree on a set of evaluation metrics. The systems are then evaluated periodically on a set of (previously unseen) test data, using the agreed upon evaluation metrics. The evaluation makes it possible not only to compare the effectiveness of various technical approaches but also to track overall progress in the field.

For the Spoken Language Systems program, the Air Travel Information System (ATIS) (Price, 1990) was chosen as the application domain for the common evaluation. This is a database interface application, where the data were drawn from a nine-city subset of the Official Airline Guide, containing airline, schedule, and ground transportation information.<sup>3</sup> To support this effort, sites cooperated to collect a training corpus of 14,000 spontaneous utterances (Hirschman et al., 1992), and, to date, there have been four formal evaluations in this domain (Hirschman et al., 1993; Pallett, 1990, 1991; Pallett et al., 1992, 1993). At the start of the Spoken Language Systems program in 1989, an accepted metric had evolved for speech recognition, namely word accuracy (Pallett, 1989); however, no comparable metric was available for measuring understanding. Over the past 4 years, the research community has developed an understanding metric for database interface tasks, using either speech or typed input (Bates et al., 1990; Hirschman et al., 1992). To date, there is still no agreed upon metric for the rich multidimensional space of interactive systems, which includes the system's ability to communicate effectively with the user, as well as an ability to understand what the user is trying to accomplish.

The remainder of this paper is divided into four sections: "The Dual Role of Language Processing" discusses the role of language processing in providing both understanding and constraint; "The Role of Discourse" outlines several sources of discourse and conversational constraints that are available at the inter-sentential level; "Evaluation" returns to the issue of evaluation and how it has affected research; and the final section, "Conclusions," summarizes these issues in terms of how they affect the development of deployable spoken language systems.

---

<sup>3</sup> There is now an enlarged 46-city version of the ATIS database; it will be the focus of the next round of evaluation.

## THE DUAL ROLE OF LANGUAGE PROCESSING

At the outset of the ARPA Spoken Language Systems program, two roles were envisioned for natural language processing:

- Natural language processing would interpret the strings of words produced by the speech recognition system, to provide *understanding*, not just transcription.
- Natural language would provide an additional knowledge source to be combined with information from the recognizer, to improve understanding and recognition by rejecting nonsense word strings and by preferring candidate word strings that "made sense."

### Approaches to Spoken Language Understanding

To achieve reasonable coverage, a spoken language system must understand what the user says, even in the face of hesitations, verbal repairs, metonymy, and novel vocabulary, as illustrated by Moore by in this volume.<sup>4</sup> And the system must do this despite the noise introduced by using speech (as opposed to text) as the input medium. This means that it is not possible to use strict rules of grammar to rule out nongrammatical utterances. Doing so results in a significant degradation in coverage.

To achieve reasonable coverage, the language-processing components have developed techniques based on partial analysis—the ability to find the meaning-bearing phrases in the input and construct a meaning representation out of them, without requiring a complete analysis of the entire string.<sup>5</sup> The current approaches to language understanding in the ARPA community can be divided into two general types:

- A *semantics-driven* approach identifies semantic constructs on the basis of words or word sequences. Syntax plays a secondary role—to identify modifying phrases or special constructions such as dates.

---

<sup>4</sup> During the discussion, R. Schwartz (Bolt, Beranek, and Newman) made an interesting observation. He reported that, given recognizer output from people reading the *Wall Street Journal*, human observers could reliably distinguish correctly transcribed sentences from incorrectly transcribed ones. Given recognizer output from the ATIS task, the observers could *not* reliably distinguish correctly transcribed output from incorrectly transcribed output, due to the irregularities that characterize spontaneous speech.

<sup>5</sup> Actually, partial analysis is equally critical for large-scale text-understanding applications, as documented in the *Proceedings of the Fourth Message Understanding Conference* (1992) and the *Proceedings of the Fifth Message Understanding Conference* (1993).

These systems differ in their approach to building the semantic constructs and include recursive transition networks to model the semantic phrases (Ward, 1991; Ward et al., 1992), template-matching (Jackson et al., 1991), hidden Markov models (HMMs) to segment the input into semantic "chunks" (Pieraccini et al., 1992), and neural networks or decision trees (Cardin et al., 1993).

- *A syntax-driven* approach first identifies syntactic constituents, and semantic processing provides an interpretation based on the syntactic relations. Recent syntax-driven systems are coupled with a backup mechanism to make use of partial information (Dowding et al., 1993; Linebarger et al., 1993; Seneff, 1992a; Stallard and Bobrow, 1993). In this approach the syntactic component first tries to obtain a full parse; failing that, partial syntactic analyses are integrated into a meaning representation of the entire string.

Both styles of system have shown increasing coverage and robustness for the understanding task. The next section discusses how successful these systems have been in providing constraint for understanding or recognition.

### Interfacing Speech and Language

The architecture of a spoken language system, in particular the interface between the recognition and language components, is closely related to the role of the language-processing component. For understanding, the minimal interface requirement is a word string passed from the recognizer to the language-understanding component for interpretation. However, if the language-understanding component is to provide an additional knowledge source to help in choosing the "right answer," it must have access to multiple hypotheses from the recognizer. The N-best interface (Chow and Schwartz, 1990; Schwartz et al., 1992)<sup>6</sup> has proved to be a convenient vehicle for such experimentation: it makes it easy to interface the recognition and understanding components, it requires no change to either component, and it permits off-line exploration of a large search space. Also, there has recently been renewed interest in word networks as a compact representation of a large set of recognition hypotheses (Baggia et al., 1991; Glass et al., 1993).

The language-processing component can provide help either to

---

<sup>6</sup> The N-best interface produces the top N recognizer hypotheses, in order of recognizer score.

recognition (choosing the right words) or to understanding (given multiple inputs from the recognizer). For recognition there are several ways the language-processing component can distinguish good word string candidates from less good ones. First, it is possible that a linguistically based model could provide lower perplexity than simple n-gram models. For example, the layered bigram approach (Seneff et al., 1992) combines a linguistically based grammar with sibling-sibling transition probabilities within the parse tree, to produce a grammar with lower perplexity than a conventional trigram model.<sup>7</sup>

Another approach is to use the language-understanding component for *filtering* or, more accurately, for preference based on parsability: the system prefers a hypothesis that gets a full parse to one that does not. At Carnegie-Mellon University (CMU), Ward and Young recently reported interesting results based on the tight coupling of a set of recursive transition networks (RTNs) into the recognizer, in conjunction with a bigram language model (Ward and Young, 1993); use of the RTN provided a 20 percent reduction in both word error and understanding error, compared to the recognizer using the word-class bigram, followed by the RTN for understanding.<sup>8</sup>

In experiments at the Massachusetts Institute of Technology (MIT) the TINA language-understanding system was used in a loosely coupled mode to filter N-best output. This produced a very small decrease in word error (0.2 percent, from 12.7 percent for  $N = 1$  to 12.5 percent for  $N = 5$ ) but a somewhat larger decrease in sentence recognition error (1.7 percent, from 48.9 percent for  $N = 1$ , to 47.2 percent for  $N = 5$ ).<sup>9</sup>

Alternatively, if the language-processing system can provide scores for alternate hypotheses, the hypotheses could be (re)ranked by a weighted combination of recognition and language-understanding score. Use of an LR parser produced over 10 percent reduction in error rate for both sentence error and word error when used in this way (Goddeau, 1992).<sup>10</sup> In summary, there have been some preliminary successes

---

<sup>7</sup> This model has not yet been coupled to the recognizer, to determine its effectiveness in the context of a complete spoken language system.

<sup>8</sup> These results were obtained using an older version of the recognizer and a somewhat higher-perplexity language model. Future experiments will determine whether this improvement carries over to the newer, higher-accuracy recognizer.

<sup>9</sup> These experiments were run on the February 1992 test set using TINA to prefer to the first parseable sentence; the best results were achieved for  $N$  in the range of 5 to 10; as  $N$  grew larger, both word and sentence error began to increase again.

<sup>10</sup> This system was not used for understanding in these runs but only for improving the recognition accuracy.

using linguistic processing as an additional knowledge source for recognition. However, more experiments need to be done to demonstrate that it is possible to use the same language-processing system in both recognition and language understanding to produce word error rates that are better than those of the best current systems.

TABLE 1 Language Understanding Scores for N = 1 vs. N = 10

Category	N=1	N=10
T	498 (72%)	528 (77%)
F	80 (12%)	81 (12%)
N.A.	109 (16%)	78 (11%)
Weighted error	39	35

It is clearer that language processing can be used to improve understanding scores, given alternatives from the recognizer. Early results at MIT showed a significant improvement using the language component as a filter and an additional but smaller improvement by reordering hypotheses using a weighted combination of recognizer and parse score (Goodine et al., 1991; Hirschman et al., 1991). An additional improvement in the number of sentences correctly understood was obtained by tightly coupling the language processing into the recognizer.<sup>11</sup> In recent MIT results using the N-best interface and the TINA language-understanding system (Seneff, 1992a) as a filter, TINA produced a significant improvement in understanding results. Table 1 shows the results for a test set of 687 evaluable utterances (the February 1992 test set): the error rate (1 - % Correct) dropped from 28 percent for N = 1 to 23 percent for N = 10. Similarly, the

<sup>11</sup> These results were obtained in VOYAGER domain, using a word-pair grammar and a language-understanding system trained with probabilities that required a full parse. It is not clear that the improvement would have been as dramatic with a lower-perplexity language model (e.g., bigram) or with a robust parsing system.

<sup>12</sup> The weighted error is calculated in terms of obtaining a correct answer from the database; an incorrect answer is penalized more heavily than no answer:  $\text{Weighted Error} = \#(\text{No Answer}) + 2 * \#(\text{Wrong Answer})$ .

<sup>13</sup> Overall, using the top N hypotheses affected 41 of the 687 sentences; the system answered an additional 31 queries (26 correctly, 5 incorrectly); 7 queries went from "incorrect" to "correct" and 3 from "correct" to "incorrect."

weighted error rate<sup>12</sup> dropped from 39 percent for  $N = 1$  to 35 percent for  $N = 10$ .<sup>13</sup>

At other sites, BBN also reported success in using the language-understanding system to filter the N-best list (Stallard and Bobrow, 1993).

In summary, it is clear that natural language understanding has contributed significantly to choosing the best hypothesis for understanding. There are some promising approaches to using linguistically based processing to help recognition also, but in this area it is hard to compete with simple statistical n-gram language models. These models have low perplexity, are easily trained, and are highly robust. For these reasons it still seems quite reasonable to keep the loosely coupled approach that uses different techniques and knowledge sources for recognition and understanding. The n-gram models are highly effective and computationally efficient for recognition, followed by some more elaborate language-understanding "filtering" to achieve improved hypothesis selection. I return to the issue of how language processing systems can provide constraint in the section "The Role of Discourse" in moving beyond the sentence to the discourse and conversational levels.

### Progress in Spoken Language Understanding

One of the benefits of common evaluation is that it is possible to track the progress of the field over time. Figure 1 shows the decline of error rates for spoken language understanding since the first benchmark in June 1990. The figure plots error metrics for the best-scoring system at each evaluation, as scored on context-independent utterance—utterances that can be interpreted without dialogue context.<sup>14</sup> The data are taken from the Defense Advanced Research Project Agency (DARPA) Benchmark Evaluation summaries (Pallett, 1990, 1991; Pallett et al., 1992, 1993). The figure shows four distinct error metrics:

1. sentence error, which is a recognition measure requiring that all the words in a sentence be correctly transcribed;
2. natural language understanding, which uses the transcribed input to compute the understanding error rate (100 - % Correct);
3. spoken language understanding, which uses speech as input and the same understanding error metric; and

---

<sup>14</sup> This set was chosen because it is the set for which the most data points exist. During the most recent evaluation, the best results for all evaluable queries did not differ significantly from the results for the context-independent queries alone.

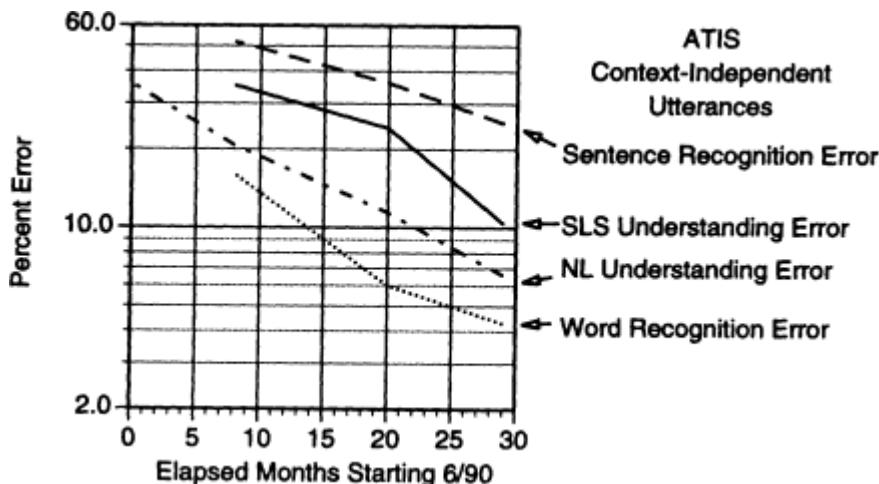


FIGURE 1 Error rate decrease over time for spoken language systems. (SLS, Spoken language system; NL, natural language.)

4. word error, which remains the basic speech recognition metric.

In Figure 1 we observe several interesting points. First, it is easier to understand sentences than to recognize them—that is, the sentence-understanding error is much lower than the sentence recognition error (10.4 percent compared to 25.2 percent). This is a significant difference and is due to the fact that many recognition errors do not significantly affect understanding for this task (e.g., confusing "a" and "the"). This means that the best current spoken language ATIS system understands almost 90 percent of its (evaluable) input.<sup>15</sup> Second, we can see how much of the error is due to understanding the perfectly transcribed input and how much is due to recognition errors. For the 1993 evaluation, the natural language error was at 6.6 percent, while the spoken language weighted error was at 10.4 percent. We can conclude that the natural language component was responsible for about 60 percent of the error and recognition for about 40 percent. This was borne out by a detailed error analysis furnished

<sup>15</sup> Only answers that have a database response can be evaluated using current evaluation technology. Therefore, the understanding figures all refer to evaluable utterances, which constitute approximately two-thirds of the corpus; the remainder of the utterances include system-initiated queries, polite forms that have no answer ("Thank you"), and other utterances that have no obvious database answer.

by Wayne Ward (CMU, personal communication, 1993) for these results. These figures lead to several conclusions:

- Both speech recognition and language understanding have made impressive progress since the first benchmarks were run in 1990; the understanding error rate has been reduced by a factor of 3 for both spoken language and natural language in less than 2 years. In speech recognition the word error rate has been reduced by a factor of 4 and the sentence recognition error by a factor of 2.
- Language-understanding technology has been able to accommodate to spoken language input by using robust processing techniques; these techniques have included heavy use of application-specific semantic information and relaxation of requirements to find a complete syntactic analysis of an entire utterance.
- To date, the loosely coupled N-best interface has proved adequate to provide multiple hypotheses to the language-understanding component; this allows statistical n-grams to provide constraint during recognition, followed by use of the language-understanding component to filter out N-best candidates. To make tighter coupling worthwhile, language-understanding systems will have to provide additional constraint without duplicating the "cheap" knowledge available from the statistical n-gram models and without losing robustness and coverage.

These figures paint a very optimistic picture of progress in spoken language understanding. However, we need to ask whether we have just gotten better at doing a very limited task (the ATIS task) or whether the field as a whole has made progress that will carry over in reasonable ways to new application domains. Because, to date, the research community has not developed any metrics in the area of portability, it is impossible to make any quantitative statements. However, there is evidence from the related field of natural language message understanding that systems are getting better, that they can be ported to new domains and even to new languages, and that the time to port is being reduced from person years to person months.<sup>16</sup> Because the natural language systems use similar techniques of partial or robust parsing and application-specific semantic rules, it seems reasonable to conclude that spoken language systems are not just getting better at doing the ATIS task but that these results will carry over to new

---

<sup>16</sup> This information comes from the recent *Proceedings of the Fifth Message Understanding Conference* (1993), in which a number of sites reported on systems that were built in one domain and then ported to both a new domain and to a new language, often by a single person working on it for a few months.

domains. However, there is a critical need to develop realistic measures of portability and to assess the degree to which current systems are now portable.

## THE ROLE OF DISCOURSE

The preceding section discussed the state of language understanding and the limited constraint it provided. The systems discussed above focused mainly on *within-sentence constraint*. It is clear that there is significant constraint *between sentences which results from constraints and coherence and conversational conventions*. This section outlines ways in which these higher levels provide constraint, based on recent work in the context of MIT's ATIS system (Seneff et al., 1991; Zue et al., 1992).

### Constraints on Reference

When the system is in an information-providing role, as in the ATIS application, it introduces new objects into the discourse in the course of answering user queries. Thus, if the user says, "Do you have any flights from Pittsburgh to Boston next Wednesday in the morning?", the system will respond with a set of flights, including airline code plus flight number, departure, and arrival time, etc., as shown in [Figure 2](#).

This display introduces new entities—"US674," "US732," and "US736"—into the conversation. Prior to this display, it would be relatively unlikely that a user would talk about one of these flights by flight number (unless he or she is a frequent flyer). However, after this display, it is quite likely that, if a flight number is mentioned, it will correspond to one of the displayed flights. The MIT system currently uses the list of previously displayed flight numbers to filter the N-best output. This is helpful since otherwise there is no way to choose among word string hypotheses that differ only in numbers (e.g., "U S six seventy" vs. "U S six seventy-four").

The same observation holds true for other displays—for example,

AIRL#	FROM	TO	LEAVE	ARRIVE	DURA	STOPS
US674	PIT	BOS	1200	1328	88	0
US732	PIT	BOS	710	839	89	0
US736	PIT	BOS	840	1006	86	0

FIGURE 2 System display introducing new referents.

fare restrictions: "What is AP slash fifty-seven?" is a common response to displaying the fare code restriction table. From these examples, it is clear that:

- certain classes of objects only become available for reference once the system has introduced them in the course of providing information to the user and
- certain kinds of abbreviations may elicit follow-on questions requesting clarification.

Both phenomena provide constraints on what the user is likely to say next. This kind of constraint is particularly useful since it allows the system to choose among syntactically equivalent strings (e.g., several numbers or several strings of letters), which is information not readily available from other knowledge sources.

### Constraints from Mixed Initiative

The previous discussion points to the system's role in contributing to the discourse. But the system can do far more than passively provide information to the user—it can take the lead in eliciting additional information to narrow the scope of the query or to converge more quickly on a solution. For example, to book a flight, the user must provide certain kinds of information: departure city, destination city, travel date, class of flight, etc. The system can take the initiative in requesting this information, asking for information on the date of travel, class of service, and so forth. The MIT system takes this kind of initiative in "booking mode" (Seneff et al., 1991). In a recent investigation I looked at the responses to five different system queries (request for place of departure, for destination, for one-way vs. roundtrip flight, for travel date, and for fare class). These responses to system-initiated queries occur at a rate of 6.4 percent in a training corpus of 4500 utterances. By knowing that a user is responding to a system query, we get both syntactic constraint and strong semantic constraint on the content of these responses: at the syntactic level, over 70 percent of the responses are response fragments, not full sentences. At the semantic level, the user provides the requested (and therefore predictable) semantic information in over 90 percent of the cases. This mode of interaction clearly imposes very strong constraints for recognition and understanding, but these constraints have not yet been incorporated into running systems in the ARPA community, in part because the evaluation metrics have discouraged use of interactive dialogue (see discussion in the section "[Evaluation](#)" below).

Several of the European SUNDIAL spoken language systems use system initiative and have achieved improved recognition accuracy by developing specialized language models for the set of system states corresponding to the different system queries. For example, in the SUNGerm train reservation system (Niedermaier, 1992), 14 such dialogue-dependent states were distinguished, which provided perplexity reduction ranging from 16 to 60 percent, depending on the state. This also resulted in improved recognition accuracy (ranging from 0 to 16 percent, again depending on the state). A similar approach was used in the French SUNDIAL system (Andry, 1992), which distinguished 16 dialogue states and used distinct word-pair grammars, resulting in significant improvement in word recognition and word accuracy.

### **Order in Problem Solving and Dialogue**

There is also constraint-derived implicitly from the task orientation of the application. People tend to perform tasks in a systematic way, which provides ordering to their exploration of the search space. For example, in looking at opening sentences for tasks in the ATIS domain, we find that 94 percent of the initial sentences contain information about departure city and destination city, as in "What is the cheapest one-way fare from Boston to Denver?" Information such as date and departure time is typically added before "fine tuning" the choice by looking at things like fare restrictions or ground transportation. It is clear that the structure of the problem and the way that information is stored in the database constrain the way in which the user will explore this space. There have been ongoing experiments at CMU (Young and Matessa, 1991) in this area, using hand-crafted knowledge bases to incorporate dialogue state information. There have also been some stochastic models of dialogue—for example, that of Nagata (1992), in which illocutionary-force trigrams (e.g., transitions from question to response) were computed from training data, resulting in significant reduction in "perplexity" for assignment of illocutionary force.

### **Discourse Constraints in a Spoken Language System**

The discourse-level sources of constraint described above appear promising. There is one potential drawback: current systems are not very portable to new domains. Presently, it requires a good system developer to create a new lexicon, a new set of semantic and syntactic rules, and a new interface to whatever application is chosen. If we

add to this another set of rules that must be handcrafted, namely discourse rules, portability is even more difficult. For this reason, approaches that can be automatically trained or ones that are truly domain and application independent should provide greater portability than approaches that require extensive knowledge engineering.

Even though these types of constraint look promising, it is always hard to predict which knowledge sources will complement existing knowledge sources and which will overlap, producing little improvement. Building spoken language systems is very much an iterative trial-and-error process, where different knowledge sources are exploited to see what effect they have on overall system performance. This leads into the next section, which raises the issue of evaluating system performance.

## EVALUATION

Evaluation has played a central role in the ARPA Spoken Language System program. The current evaluation method (Hirschman et al., 1992, 1993) provides an automated evaluation of the correctness of database query responses, presented as prerecorded (speech or transcribed) data in units of "scenarios."<sup>17</sup> The data are annotated for their correct reference answers, expressed as a set of minimal and maximal database tuples. The correct answer must include at least the information in the minimal answer and no more information than what is in the maximal answer. Annotation is done manually by a trained group of annotators. Once annotated, the data can be run repeatedly and answers can be scored automatically using the comparator program (Bates et al., 1990).

This methodology has evolved over four evaluations. In its current form, both context-independent utterances and context-dependent utterances are evaluated. The remaining utterances (about 25 to 35 percent of the data) are classified as unevaluatable because no well-defined database answer exists. For evaluation the data are presented one scenario at a time, with no side information about what utterances are context independent, context dependent, or unevaluatable. The availability of a significant corpus of transcribed and annotated training data (14,000 utterances of speech data, with 7500 utterances

---

<sup>17</sup> A scenario is the data from a single user solving a particular problem during one sitting.

annotated) has provided an infrastructure leading to very rapid progress in spoken language understanding—at least for this specific domain.

It is clear that this infrastructure has served the research community well. Figure 1 showed a dramatic and steady decrease in error rate in spoken language understanding over time. However, it is now time to look again to extending our suite of evaluation methods to focus research on new directions. The preceding section argued that natural language understanding could contribute more constraint if we go beyond individual sentences to look at discourse. Unfortunately, the present evaluation methods discourage such experimentation on several grounds. First, mixed initiative dialogue is explicitly disallowed in the evaluation because it may require an understanding of the system's side of the conversation. Because the current evaluation makes use of prerecorded data and assumes that the system will never intervene in a way to change the flow of the conversation, mixed initiative may disrupt the predictable flow of the conversation. If developers were allowed to experiment with alternative system response strategies, there would be no obvious way to use complete prerecorded sessions for evaluation. This is a serious problem—it is clearly desirable and useful to have a static set of data, with answers, so that experiments can be run repeatedly, either for optimization purposes or simply to experiment with different approaches. This kind of iterative training approach has proved highly successful in many areas. Nonetheless, we are at a crossroads with respect to system development. If we wish to exploit promising methods that use system-initiated queries and methods to model the flow of queries to solve a problem, we must be able to evaluate the system *with the human in the loop*, rather than relying solely on off-line recorded data.

Second, there is a potential mismatch between the notion of a canonically correct answer and a useful answer. There are useful answers that are not correct and, conversely, correct answers that are not useful. For example, suppose a user asks, "Show Delta flights from Boston to Dallas leaving after 6 p.m.;" furthermore, suppose the system does not understand p.m. and misinterprets the query to refer to flights leaving after 6 a.m. This happens to include those after 6 p.m., as is clear from the answer, shown in Figure 3.

The answer is not correct—it includes flights that depart before 6 p.m., but it also includes the explicitly requested information. If the user is able to interpret the display, there will be no need to reask the query.

On the other hand, it is quite possible to get information that is canonically correct but not useful. Suppose the user asks, "What are

User query:

Show Delta flights from Boston to Dallas leaving after six P M

System understands:

Show Delta flights from Boston to Dallas leaving after six the flights from San Francisco to Washington next Tuesday?" Furthermore, suppose the system misrecognizes "Tuesday" and hears "Thursday" instead. It then makes this explicit, using generation, so that the user knows what question the system is answering, and answers "Here are the San Francisco to Washington flights leaving Thursday" and shows a list of flights. If the set of Tuesday flights is identical to the set of Thursday flights, the answer is technically correct. However, the user *should* reject this answer and reask the question, since the system reported very explicitly that it was answering a different question. The user would have no reason to know that the Tuesday and Thursday flight schedules happened to be identical.

AIRL	NUMBER	FROM	TO	LEAVE	ARRIVE	STOPS
DELTA	1283	BOS	DFW	8:20 A.M.	11:05 A.M.	0
DELTA	169	BOS	DFW	11:20 A.M.	2:07 P.M.	0
DELTA	841	BOS	DFW	3:25 P.M.	6:10 P.M.	0
DELTA	487	BOS	DFW	6:45 P.M.	9:29 P.M.	0

FIGURE 3 Example of incorrectly understood but useful output.

The current evaluation methods have evolved steadily over the past 4 years, and we now need to move toward metrics that capture some notion of utility to the user. The initial evaluation began with only context-independent utterances and evolved into the evaluation of entire scenarios, including context-dependent utterances.<sup>18</sup> To push research toward making usable systems, we must take the next step and develop evaluation methods that encourage interactive dialogue and that reward systems for useful or helpful answers.

From mid-1992 to early 1993 there was an effort to investigate new evaluation methods that included interactivity and captured some notion of usability. During this period, four of the ARPA sites con

---

<sup>18</sup> Unevaluable utterances are still not counted in the understanding scores, but they are included as part of the input data; system responses for unevaluable queries are currently just ignored. For recognition rates, however, all sentences are counted as included in the evaluation.

ducted an experimental "end-to-end" evaluation (Hirschman et al., 1993) that introduced several new features:

- task completion metrics, where users were asked to solve travel planning tasks that had a well-defined answer set; this made it possible to judge the correctness of the user's solution and the time it took to obtain that solution;
- log-file evaluation by human judges, who reviewed the set of system-user interactions recorded on the session log file; the evaluators were asked to judge whether responses were correct or appropriate, thus making it possible to be more flexible in judging answer correctness and in evaluating interactive systems; and
- a user satisfaction questionnaire that asked the users to rate various aspects of the system.

This methodology still needs considerable work, particularly in factoring out variability due to different subjects using the different systems. However, this experiment was a first step toward a "whole-system" evaluation, away from evaluating only literal understanding.

## CONCLUSIONS

We can draw several conclusions from the preceding discussion about the role of language understanding in current spoken language systems:

- Current systems can correctly answer correctly almost 90 percent of the spoken input in a limited domain such as air travel planning. This indicates that natural language processing has become robust enough to provide useful levels of understanding for spoken language systems in restricted domains. Given the variability of the input, spontaneous speech effects, effects of unknown words, and the casual style of spontaneous speech, this is an impressive achievement. Successful understanding strategies include both semantic-based processing and syntactic-based processing, which relies on partial understanding as a backup when complete analysis fails.
- The N-best interface has proved to be an adequate interface between the speech and language components; this allows the language-understanding component to apply some constraint by filtering or reordering hypotheses. But as language systems become better at providing constraint, a tighter interface may prove worthwhile.
- The discourse and conversational levels of processing would appear to provide significant constraint that is not being fully ex

ploited by the ATIS systems. Some of the constraint derives from use of user-system interaction, however, and the current data collection and evaluation paradigms discourage system developers from exploring these issues in the context of the common ARPA evaluation.

- We need new evaluation methods to explore how to make systems more usable from the user's point of view. This kind of evaluation differs from those used to date in the ARPA community in that it would require evaluating system plus user as a unit. When evaluating interactive systems, there seems to be no obvious way of factoring the user out of the experiment. Such experiments require careful controls for subject variability, but without such experiments we may gain little insight into what techniques help users accomplish tasks.

It is clear that spoken language technology has made dramatic progress; to determine how close it is to being usable, we need to understand the many complex design trade-offs involved in matching the technology with potential applications. For example, response speed is an important dimension, but it can be traded off for response accuracy and/or vocabulary size. It is important to look at applications in the context of the application needs—some applications may require high-accuracy but only a small vocabulary, especially if the system can guide the user step by step through the interaction. Other applications may be more error tolerant but require greater flexibility to deal with unprompted user input.

The ATIS domain focused on collecting spontaneous input from the user. This may not be the best way to build a real air travel information system, as Crystal (DARPA) pointed out during the discussion at the NAS colloquium. However, it has been an effective way of gathering data about what people might do if they could talk to a (passive) system. To drive the research forward, we may want to move on to larger vocabulary domains, or more conceptually complex applications, or real applications, rather than mock-ups, such as ATIS. However, even as we extend our abilities to recognize and understand more complex input, we must not lose sight of the "other side," which is the machine's communication with the user.

For the machine to be a usable conversational partner, it must help to keep the conversation "synchronized" with the user. This may mean providing paraphrases to let the user know what question it is answering—or it may involve giving only short answers to avoid unnecessary verbosity. To date, there has been very little work on appropriate response generation, how to use these responses to help the user detect system errors, or how the system's response may aid the user in assimilating the information presented.

One of the major weaknesses of current systems, raised by Neuberg (Institute for Defense Analysis) during the NAS colloquium discussion, is that they do not know what they do not know—in terms of vocabulary and knowledge. For example, if you ask an ATIS system about whether you can pay with a credit card, you would like it to tell you that it does not know about credit cards, rather than just not understanding what was said or asking you to repeat your query; systems are just now beginning to incorporate such capabilities.<sup>19</sup> Keeping the user "in bounds" is an important aspect of overcoming fragility in the interface and may have a strong effect on user satisfaction in real systems.

Another major stumbling block to deployment of the technology is the high cost of porting to a new application. For language technology, this is still largely a manual procedure. Even for recognition that uses automated procedures for training, a significant amount of application-specific training data are required.<sup>20</sup> To support widespread use of spoken language interfaces, it is crucial to provide low-cost porting tools; otherwise, applications will be limited to those few that have such a high payoff that it is profitable to spend significant resources building the specific application interface.

In conclusion, as the technology begins to move from the laboratory to real applications, it is critical that we address the system in its "ecological context," that is, the context of its eventual use. This means looking not only at the recognition and understanding technologies but at the interface and interactive dimensions as well. This can best be accomplished by bringing together technology deployers and technology developers, so that developers can study the tradeoffs among such dimensions as speed, accuracy, interactivity, and error rate, given well-defined criteria for success provided by the technology deployers. This should lead to a better understanding of spoken language system technology and should provide a range of systems appropriate to the specific needs of particular applications.

## REFERENCES

- Andry, F., "Static and Dynamic Predictions: A Method to Improve Speech Understanding in Cooperative Dialogues," ICSLP-92 Proceedings, pp. 639-642, Banff, October 1992.

---

<sup>19</sup> At the Spoken Language Technology Applications Day (April 13, 1993), the CMU ATIS system demonstrated its ability to handle a variety of "out-of-domain" questions about baggage, frequent flyer information, and cities not in the database.

<sup>20</sup> For example, for ATIS, over 14,000 utterances have been collected and transcribed.

- Baggia, P., E. Fissore, E. Gerbino, E. Giachin, and C. Rullent, "Improving Speech Understanding Performance Through Feedback Verification," *Eurospeech-91*, pp. 211214, Genoa, September 1991.
- Bates, M., S. Boisen, and J. Makhoul, "Developing an Evaluation Methodology for Spoken Language Systems," *Proceedings of the Third DARPA Speech and Language Workshop*, R. Stern (ed.), Morgan Kaufmann, June 1990.
- Cardin, R., Y. Cheng, R. De Mori, D. Goupil, R. Kuhn, R. Lacouture, E. Millien, Y. Normandin, and C. Snow, "CRIM's Speech Understanding System for the ATIS Task," presented at the Spoken Language Technology Workshop, Cambridge, Mass., January 1993.
- Chow, Y.-L., and R. Schwartz, "The N-best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses," *ICASSP-90*, Toronto, Canada, pp. 697-700, 1990.
- Dowding, J., J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A Natural Language System for Spoken Language Understanding," *Proceedings of the Human Language Technology Workshop*, M. Bates (ed.), Princeton, N.J., March 1993.
- Glass, J., D. Goddeau, D. Goodine, L. Hetherington, L. Hirschman, M. Phillips, J. Polifroni, C. Pao, S. Seneff, and V. Zue, "The MIT ATIS System: January 1993 Progress Report," presented at the Spoken Language Technology Workshop, Cambridge, Mass., January 1993.
- Goddeau, D., "Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems," *ICSLP-92* pp. 321-324, October 1992.
- Goodine, D., S. Seneff, L. Hirschman, and V. Zue, "Full Integration of Speech and Language Understanding," *Eurospeech-91*, pp. 845-848, Genoa, Italy, September 1991.
- Hirschman, L., S. Seneff, D. Goodine, and M. Phillips, "Integrating Syntax and Semantics into Spoken Language Understanding," *Proceedings of the DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, pp. 366-371, Asilomar, February 1991.
- Hirschman, L., et al., "Multi-Site Data Collection for a Spoken Language Corpus," *ICSLP-92*, Banff, Canada, October 1992.
- Hirschman, L., M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann, "Multi-Site Data Collection and Evaluation in Spoken Language Understanding," *Proceedings of the Human Language Technology Workshop*, M. Bates (ed.), Princeton, N.J., March 1993.
- Jackson, E., D. Appelt, J. Bear, R. Moore, and A. Podlozny, "A Template Matcher for Robust NL Interpretation," *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.
- Linebarger, M., L. Norton, and D. Dahl, "A Portable Approach to Last Resort Parsing and Interpretation," *Proceedings of the ARPA Human Language Technology Workshop*, M. Bates (ed.), Princeton, March 1993.
- Nagata, M., "Using Pragmatics to Rule Out Recognition Errors in Cooperative Task-Oriented Dialogues," *ICSLP-92*, pp. 647-650, Banff, October 1992.
- Niedermair, G., "Linguistic Modeling in the Context of Oral Dialogue," *ICSLP-92*, pp. 635-638, Banff, 1992.
- Pallett, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *ICASSP-89*, pp. 536-539, IEEE, Glasgow, Scotland, 1989.
- Pallett, D., "DARPA ATIS Test Results June 1990," *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 114-121, R. Stern (ed.), Morgan Kaufmann, 1990.
- Pallett, D., "Session 2: DARPA Resource Management and ATIS Benchmark Test Poster"

- Session", Proceedings of the Fourth DARPA Speech and Natural Language Workshop Workshop, P. Price (ed.), Morgan Kaufmann, 1991.
- Pallett, D., et al. "February 1992 DARPA ATIS Benchmark Test Results Summary," Proceedings of the Fifth DARPA Speech and Natural Language Workshop, M. Marcus (ed.), Morgan Kaufmann, 1992.
- Pallett, D., J. Fiscus, W. Fisher, and J. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program," Proceedings of the Human Language Technology Workshop, M. Bates (ed.), Princeton, N.J., March 1993.
- Pieraccini, R., E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.-H. Lee, and J. Wilpon, "A Speech Understanding System Based on Statistical Representation of Semantics," ICASSP-92, IEEE, San Francisco, 1992.
- Price, P., "Evaluation of Spoken Language Systems: The ATIS Domain," Proceedings of the Third DARPA Speech and Language Workshop, R. Stern (ed.), Morgan Kaufmann, June 1990.
- Proceedings of the Fourth Message Understanding Conf., Morgan Kaufmann, McLean, June 1992.
- Proceedings of the Fifth Message Understanding Conf., Baltimore, August 1993.
- Schwartz, R., S. Austin, F. Kubala, J. Makhloul, L. Nguyen, P. Placeway, and G. Zavaliagkos, "New Uses for the N-Best Sentence Hypotheses Within the Byblos Speech Recognition System," ICASSP-92, Vol. I, pp. 5-8, San Francisco, March 1992.
- Seneff, S., "Robust Parsing for Spoken Language Systems," ICASSP-92, pp. 189-192, San Francisco, Calif., March 1992a.
- Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," Computational Linguistics Vol. 18, No. 1, pp. 61-86, March 1992b.
- Seneff, S., L. Hirschman, and V. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," Proceedings of the Third DARPA Speech and Natural Language Workshop, P. Price (ed.), pp. 354-359, Asilomar, February 1991.
- Seneff, S., H. Meng, and V. Zue, "Language Modeling for Recognition and Understanding Using Layered Bigrams," ICSLP-92, pp. 317-320, October 1992.
- Stallard, D., and R. Bobrow, "The Semantic Linker—A New Fragment Combining Method," Proceedings of the ARPA Human Language Technology Workshop, M. Bates (ed.), Princeton, N.J., March 1993.
- Ward, W., "Understanding Spontaneous Speech: The Phoenix System," ICASSP-91, pp. 365-367, May 1991.
- Ward, W., and S. Young, "Flexible Use of Semantic Constraints in Speech Recognition," ICASSP-93, Minneapolis, April 1993.
- Ward, W., S. Issar, X. Huang, H.-W. Hon, M.-Y. Hwang, S. Young, M. Matessa, F.-H Liu, and R. Stern, "Speech Understanding In Open Tasks," Proceedings of the Fifth DARPA Speech and Natural Language Workshop, M. Marcus (ed.), Morgan Kaufmann, 1992.
- Young, S., and M. Matessa, "Using Pragmatic and Semantic Knowledge to Correct Parsing of Spoken Language Utterances," Eurospeech-91, pp. 223-227, Genoa, September 1991.
- Zue, V., J. Glass, D. Goddeau, D. Goodine, L. Hirschman, M. Philips, J. Polifroni, and S. Seneff, "The MIT ATIS System: February 1992 Progress Report," Proceedings of the Fifth DARPA Speech and Natural Language Workshop, M. Marcus (ed.), February 1992.

# Models of Natural Language Understanding

*Madeleine Bates*

## SUMMARY

This paper surveys some of the fundamental problems in natural language (NL) understanding (syntax, semantics, pragmatics, and discourse) and the current approaches to solving them. Some recent developments in NL processing include increased emphasis on corpus-based rather than example- or intuition-based work, attempts to measure the coverage and effectiveness of NL systems, dealing with discourse and dialogue phenomena, and attempts to use both analytic and stochastic knowledge.

Critical areas for the future include grammars that are appropriate to processing large amounts of real language; automatic (or at least semiautomatic) methods for deriving models of syntax, semantics, and pragmatics; self-adapting systems; and integration with speech processing. Of particular importance are techniques that can be tuned to such requirements as full versus partial understanding and spoken language versus text. Portability (the ease with which one can configure an NL system for a particular application) is one of the largest barriers to application of this technology.

## INTRODUCTION

Natural language (NL) understanding by computer began in the 1950s as a discipline closely related to linguistics. It has evolved to

incorporate aspects of many other disciplines (such as artificial intelligence, computer science, and lexicography). Yet it continues to be the Holy Grail of those who try to make computers deal intelligently with one of the most complex characteristics of human beings: language.

Language is so fundamental to humans, and so ubiquitous, that fluent use of it is often considered almost synonymous with intelligence. Given that, it is not surprising that computers have difficulty with natural language. Nonetheless, many people seem to think it should be easy for computers to deal with human language, just because they themselves do so easily.

Research in both speech recognition (i.e., literal transcription of spoken words) and language processing (i.e., understanding the meaning of a sequence of words) has been going on for decades. But quite recently, speech recognition started to make the transition from laboratory to widespread successful use in a large number of different kinds of systems. What is responsible for this technology transition?

Two key features that have allowed the development of successful speech recognition systems are (1) a simple general description of the speech recognition problem (which results in a simple general way to measure the performance of recognizers) and (2) a simple general way to automatically train a recognizer on a new vocabulary or corpus. Together, these features helped to open the floodgates to the successful, widespread application of speech recognition technology. Many of the papers in this volume, particularly those by Makhoul and Schwartz, Jelinek, Levinson, Oberteuffer, Weinstein, and Wilpon attest to this fact.

But it is important to distinguish "language understanding" from "recognizing speech," so it is natural to ask, why the same path has not been followed in natural language understanding. In natural language processing (NLP), as we shall see, there is no easy way to define the problem being solved (which results in difficulty evaluating the performance of NL systems), and there is currently no general way for NL systems to automatically learn the information they need to deal effectively with new words, new meanings, new grammatical structures, and new domains.

Some aspects of language understanding seem tantalizingly similar to problems that have been solved (or at least attacked) in speech recognition, but other aspects seem to emphasize differences that may never allow the same solutions to be used for both problems. This paper briefly touches on some of the history of NLP, the types of NLP and their applications, current problem areas and suggested solutions, and areas for future work.

## A BRIEF HISTORY OF NLP

NLP has a long, diverse history. One way of looking at that history is as a sequence of application areas, each of which has been the primary focus of research efforts in the computational linguistics community, and each of which has produced different techniques for language understanding. A number of excellent references are available that survey the field in various ways (Allen, 1978; Gazdar and Mellish, 1989; Smith, 1991; Weischedel et al., 1990; Winograd, 1983).

In the 1950s, machine translation was the first area to receive considerable attention, only to be abandoned when it was discovered that, although it was easy to get computers to map one word string to another, the problem of translating between one natural language and another was much too complex to be expressible as such a mapping.

In the 1960s the focus turned to question answering. To "understand" and respond to typed questions, most NL systems used a strongly knowledge-based approach, attempting to encode knowledge for use by a system capable of producing an in-depth analysis of the input question. That analysis would then be used to retrieve the answer to the question from a database.

In the 1970s interest broadened from database interfaces to other kinds of application systems, but the focus was still on the kinds of natural language that would be produced by a person interacting with a computer system—typed queries or commands issued one at a time by the person, each of which needed to be understood completely in order to produce the correct response. That is, virtually every word in the input had some effect on the meaning that the system produced. This tended to result in systems that, for each sentence they were given, either succeeded perfectly or failed completely.

The 1980s saw the first commercialization of research that was done in the previous two decades: natural language database interfaces and grammar and style checkers. For the first time, widespread interest began to be paid to systems that dealt with written language in paragraphs or larger chunks, instead of typed interactions. There were even some attempts to generate natural language, not just understand it.

Another change during this decade was the beginning of a redefinition of the fundamental goal of NL systems, which had always been to process every word of the input as deeply as necessary to produce an understanding of the sentence as a whole. Researchers began to think that this goal was not just difficult to achieve but perhaps impossible, and perhaps even unnecessary! Instead, partial understanding (in which some words of the input were completely

ignored), which had been viewed as a failure and a problem that needed to be fixed, began to be seen as a meaningful and useful goal. It was discovered that systems which tried to extract at least some meaning from nearly every input could succeed better (at least for certain applications) than systems that tried (and often failed) to extract the complete meaning of every input. The old model of complete understanding or complete failure began to give way to the notion of partial correctness.

Today, in the 1990s, there is strong interest in a wide spectrum of tasks that require NL processing. Spoken language systems (SLSs) (which combine speech recognition with language understanding), language generation, message processing (the term used for systems that deal with bodies of written language in a noninteractive way, including document indexing and retrieval, text classification, and contents scanning, which is also called data extraction), and interactive NL interfaces are all important research areas. Even machine translation has come full circle and is now being reinvestigated using the results of more than 30 years of research, although this time around there is interest in doing speech translation (e.g., a translating telephone) as well as text. Reports of current research in all these areas can be found in the journal of the Association for Computational Linguistics and in the proceedings of various workshops that are included in the bibliography.

The emphasis has changed over the years not only in the type of applications that are of interest but also the "reality" of the language studied. Originally, toy problems with examples of language made up by researchers and linguists were all that a system could be expected to handle. But the development of large sharable corpora (often with additional attached information such as part-of-speech assignment, or structure, or question and answer) has revolutionized the study of language understanding. Now it is considered absolutely necessary for good research to examine "real" language, preferably a large linguistic corpus obtained using a real application in as natural a setting as possible. Many of these corpora and some systems to process them are available through the Linguistic Data Consortium at the University of Pennsylvania, the Consortium for Lexical Research at New Mexico State University, and the Treebank Project at the University of Pennsylvania.

## WHY IS NLP DIFFICULT?

One way to illustrate the problems of NL processing is to look at the difference between the fundamental goals of a speech recognition (SR) system and an NL system. As illustrated in [Figure 1](#), SR is well

defined in terms of input and output. The input is a speech signal, and the output is a word string (possibly a set or lattice of alternative word strings, possibly with scores or probabilities attached). Despite the fact that there are a few nontrivial problems in deciding what is a word, it is fairly easy for two or more speech researchers to come to agreement on what is considered a word (e.g., that BOOK and BOOKS are two different words and that AIR\_FARE is a collocation) and on metrics for evaluating the quality of SR systems.



FIGURE 1 Input/output for speech recognition is easy to define.

The word error rate, which incorporates insertions, deletions, and substitutions, has been the generally accepted metric for many years; it is widely accepted, easy to apply, and works so well that there is little reason for the speech research community to change it. Because the SR task is so well defined, it is fairly easy to tell whether an SR system is doing a good job or not, and it is very easy to tell, given two different SR systems with identical input, which performs better.

Computational linguists envy this straightforward problem definition and unambiguous criterion for success! It is quite a different matter in NL processing, which is extremely difficult precisely because the input/output characteristics of an NLP system are varied, hard to specify, difficult to get common agreement on, and resistant to the development of easily applied evaluation metrics.

The range of possible inputs to an NLP is quite broad. Language can be in a variety of forms, such as the (possibly imperfectly recognized) output of an SR system, paragraphs of text (possibly containing some material that is not natural language), commands that are typed directly to a computer system, etc.

The input language might be given to the system a sentence at a time or multiple sentences all at once. It might not be sentences at all in the sense of complete grammatical units but could be fragments of language or a mixture of sentences and fragments. The input might be grammatical, nearly grammatical, or highly ungrammatical. It might contain useful cues like capitalization and punctuation or (particularly if the input comes from a speech processor) all punctuation and even the sentence boundaries might be missing.

There is not even a good way to refer to the language that is input to an NL system. "Sentence" implies grammaticality, or at least unity of the words involved, and also implies a fairly small number of words. "Utterance" implies speech but does not imply any degree

of completeness or grammar. "Word string" might be better, but it seems to exclude speech input. I will use "input" as a general term that includes all the types of language input mentioned here.

The ultimate output from a system that incorporates an NLP might be an answer from a database, a command to change some data in a database, a spoken response, or some other action on the part of the system. But these are the output of the system as a whole, not the output of the NLP component of the system—a very important distinction.

This inability to specify a natural, well-defined output for an NLP system will cause problems in a variety of ways, as we shall see more of below.

Another example of why language processing is difficult is illustrated by a recent Calvin and Hobbes cartoon:

Calvin: I like to verb words.

Hobbes: What?

Calvin: I take nouns and adjectives and use them as verbs. Remember when "access" was a thing? Now it's something you do. It got verbed.

Calvin: Verbing weirds language.

Hobbes: Maybe we can eventually make language a complete impediment to understanding.

Understanding what Calvin meant by "Verbing weirds language" stretches the limits of human language performance. One of the reasons that NL is challenging to computational linguists is its variety. Not only are new words frequently introduced into any natural language, but old words are constantly reused with new meanings (not always accompanied by new morphology).

## WHAT IS IN AN NLP SYSTEM?

There are many good overviews of NL processing, including those by Allen (1987), Gazdar and Mellish (1989), Smith (1991), and Winograd (1983), and state-of-the-art research results, including those of Bates (1993) and Bates and Weischedel (1993). [Figure 2](#) shows a generic NLP system and its input-output variety. [Figure 3](#) shows a typical view of what might be inside the NLP box of [Figure 2](#). Each of the boxes in [Figure 3](#) represents one of the types of processing that make up an NL analysis.

Most NL systems have some kind of preprocessor that does morphological analysis, dictionary lookup, and lexical substitutions (to normalize abbreviations, for example), and part-of-speech assignment. The order in which these processes are performed, the techniques

used to perform them, and the format of the result are highly idiosyncratic.

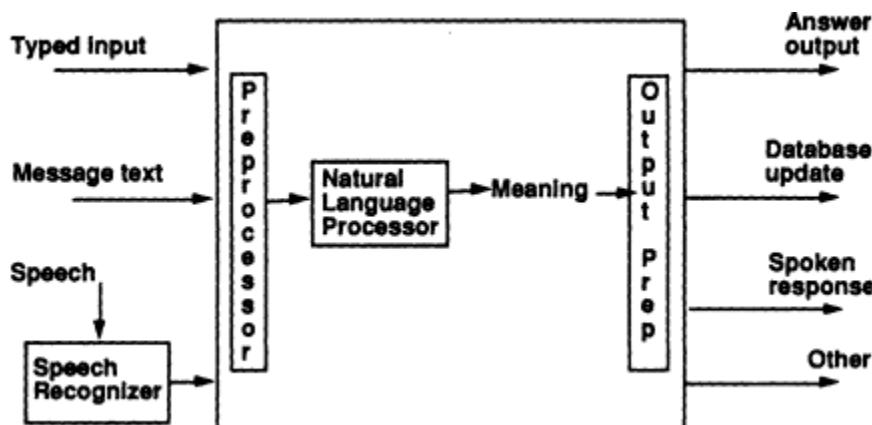


FIGURE 2 A generic NL system.

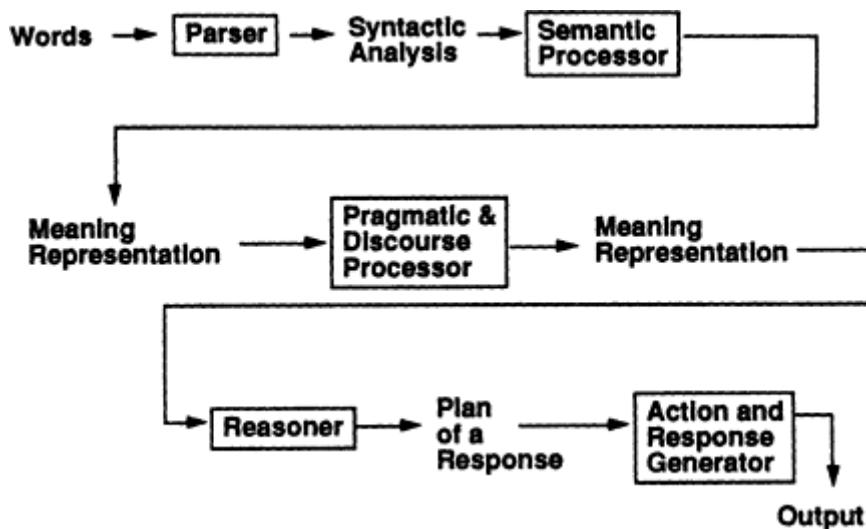


FIGURE 3 A pipeline view of the components of a generic NL system.

### Syntax

Syntactic processing is without doubt the most mature field of study in the NL field. Some think of syntax as just a way of checking

whether the input is well formed, but in fact syntactic analysis has at least two uses. One is to simplify the process of subsequent components as they try to extract meaning from the input. A second use of syntactic analysis is to help detect new or unusual meanings.

Without syntactic analysis it might be possible to use semantic probabilities to determine that a string containing "boy" and "dog" and "bit" means that a dog bit a boy, but syntax makes it easy to determine who bit whom in the input "boy bit dog." Calvin's observation that "Verbing weirds language" can be understood only by using morphological and syntactic cues, not by semantics alone.

Various syntactic formalisms have been developed and implemented in great detail, including mathematical analyses of their expressive power. Most are useful; all give incomplete accounts of the wide range of NL phenomena (as one linguist put it, "All grammars leak."). Augmented transition networks (a procedural language with most of the simplicity of context-free grammars but able to capture many context-sensitive aspects as well) were once quite popular, but in the mid 1980s a shift began toward declarative formalisms, such as the combination of context-free rules with unification.

## Semantics

Semantics is a more serious problem. The output of the semantic component is the "meaning" of the input. But how can this "meaning" be expressed? Not by anything as simple as a sequence of words. Many different "meaning representation languages" have been developed in an attempt to find a language that has the appropriate expressive power, but there is no uniform semantic representation language that can represent the meaning of every piece of NL.

Indeed, some would argue that there is no such thing as "the (i.e., unique) meaning" of a string of words, because the meaning can be greatly influenced by the context in which the words are used and by the purpose the words are intended to achieve.

Three general kinds of semantic representations are in wide use: propositional logic (most frame-based semantic representations are equivalent to this, since they do not allow quantification); First-Order Predicate Logic (FOPL, which does allow quantifiers); and various representations that can handle expressions not representable in FOPL (see, e.g., Montague, 1970).

First-order predicate logic is a good representation for some types of meaning, but it is unnecessarily complex for simple applications where quantifiers do not frequently occur, and it is not rich enough for others. And even FOPL requires prespecification of the atomic concepts from which the FOPL expressions are built. Even if it were

possible to express the meaning of any sentence in FOPL, the meaning would not be unambiguous.

Semantic processing methodologies are often chosen to match the characteristics of a particular application domain. For database access, meanings can generally be expressed in some form of predicate logic. For updating a database with information extracted from a body of text, it is crucial to be able to characterize in advance the kinds of information to be extracted; this allows the operation of the semantic component to be guided in part by the (very narrow) range of possible meanings.

Even when the problem is limited to a single application, it is very hard to get people to agree on the form of the output that a semantic component should produce.

### **Discourse and Pragmatics**

Modeling context, and using context appropriately, is one of the least well understood and most difficult aspects of NLP. Unlike context in speech, which is quite localized in time, NL context is all pervasive and extremely powerful; it can reach back (or forward) hundreds of words. It is the difficult task of the discourse and pragmatics component to determine the referents of pronouns and definite noun phrases and to try to understand elliptical sentence fragments, dropped articles, false starts, misspellings, and other forms of nonstandard language, as well as a host of other long-range language phenomena that have not even been adequately characterized much less conquered.

The pragmatic component must alter itself as a result of the meaning of previous inputs. Unlike speech, where the context that influences a particular bit of input is very close by, NL context can span multiple sentences, multiple paragraphs, or even multiple documents. Some NL expressions are forward referencing, so the relevant context is not always prior input.

But feedback across sentences is not limited to pragmatics alone. The reasoning component, or even the output generator might need to change the discourse state so that, for example, subsequent input will be understood in the context of the output that was returned to the user. This feedback is extremely important for NLP applications because real language rarely occurs in isolated sentences.

### **Reasoning, Response Planning, and Response Generation**

For straightforward inputs (e.g., a query or command like "List the flights from Boston to Topeka that leave on Saturday morning"),

it is usually possible to come up with a representation that captures the literal meaning well enough to answer the question or carry out the command, but sometimes additional reasoning is necessary.

For example, what should be done with the following input to a travel agent system: "I want to go from Pittsburgh to Boston"? Should that be interpreted as a request to list the flights from Pittsburgh to Boston, or should the effect merely be to change the discourse state so that subsequent queries will take into account the intended itinerary ("When is the first flight on Saturday morning?"), or should it cause the system to plan and produce a response to clarify the user's goals ("Do you want to see all the flights")?

Is the decision to treat the input as a command, or merely as information to the system, really part of the "NL-understanding" process, or part of the backend process that takes place after understanding, and is independent of it? The same questions can be asked about the response planner and response generator components in [Figure 2](#). Is it a part of the NL processing (a part that just is not used when the input comes from text instead of an interactive user) or is it part of the post-NLP system? Computational linguists do not agree on where to draw these boundaries or on how to represent the information that passes between them.

### Simplifying the Problem

Not every NLP system has or needs all of the components shown in [Figure 3](#). Some systems have a vestigial parser and attempt to extract meaning without using much if any syntactic information. Others combine the syntactic and semantic processing into one indistinguishable process. Some applications require little if any pragmatic or discourse processing.

A few systems attempt to leave out almost all of the components shown there and bravely try to go directly from words to a reasoner (perhaps an expert system) that attempts to produce a meaningful response without a detailed linguistic analysis at any level. This is no more extreme than eliminating the reasoner and response generator in applications where there are only a few allowable outputs.

Inside the NLP box, not every system has or needs all the pieces described above. In short, all NL systems work by simplifying some aspects of the problem they are trying to solve. The problem can be simplified on the input side (e.g., just typed questions to a database system) or on the output side (e.g., extract from multiple paragraphs of newspaper text just three pieces of information about company mergers; extract from a single spoken utterance one of six possible commands to an underlying display system). These problem simplifi

cations result in the simplification or elimination of one or more of the components shown above.

Progress in developing NLP systems will likely depend on training and evaluation (as has been the case with speech processing), but the multiplicity of components, each with its own input/output behavior that is not commonly agreed upon has made progress very difficult.

### Another View

Another way to look at the NLP problem, instead of boxes in sequential order, is as a series of independent processes, each of which uses particular kinds of knowledge bases and each of which contributes to an overall understanding of the input. This architecture is illustrated in [Figure 4](#).

In this view the lexical processor would use a dictionary to help it transform the input words into a structure with more meaning; the syntactic processor would use a grammar of the language; the semantic processor would use semantic interpretation rules and a domain model of concepts and relationships that defines the domain the system can understand; and discourse and pragmatics might use a task model that specifies the user's goals and the portions of those goals that have been achieved by previous inputs.

All of these knowledge sources are available to a process called the "understanding search," rather like the "recognition search" of speech recognition. It produces one or more outputs, such as an ordered list of possible meanings, perhaps with probabilities attached.

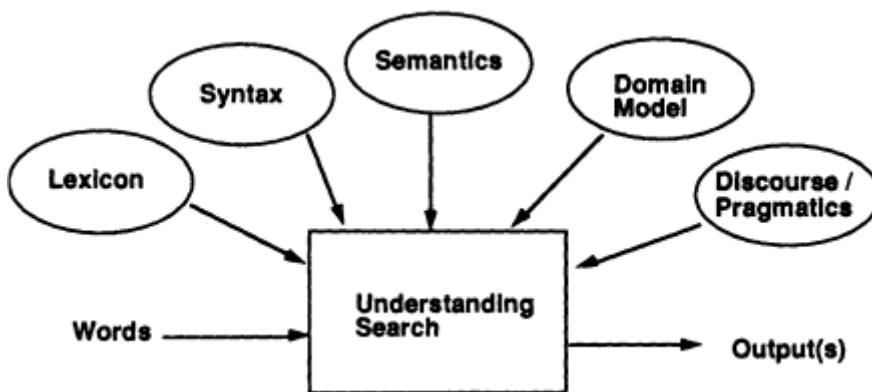


FIGURE 4 A cooperating process view of a generic NLP system.

See Marcus and Moore in this volume for more detailed descriptions of related work.

One advantage of this view of the problem is that it permits a new and very important component to be added: a learning algorithm that populates the knowledge sources by an automatic (or semiautomatic) process and an appropriately annotated corpus, as shown in Figure 5. The use of a single common understanding search process provides the framework for using all of the knowledge sources in ways that are similar enough for the results to be combined; in the old pipelined architecture (Figure 2), it would be much harder to have a uniform way of expressing the results of each component and thus much harder to develop a learning component for each of the knowledge sources.

The cooperating process view of language understanding holds the promise that the knowledge sources needed for NLP in a new domain can be created automatically. If this is true, it should become much easier to make an NLP system for a new domain by starting from a suitably annotated corpus and populating the knowledge sources so that an understanding search process could take place.

NLP systems based on this model are currently being developed, but it is too soon to assess their overall success. There are a number

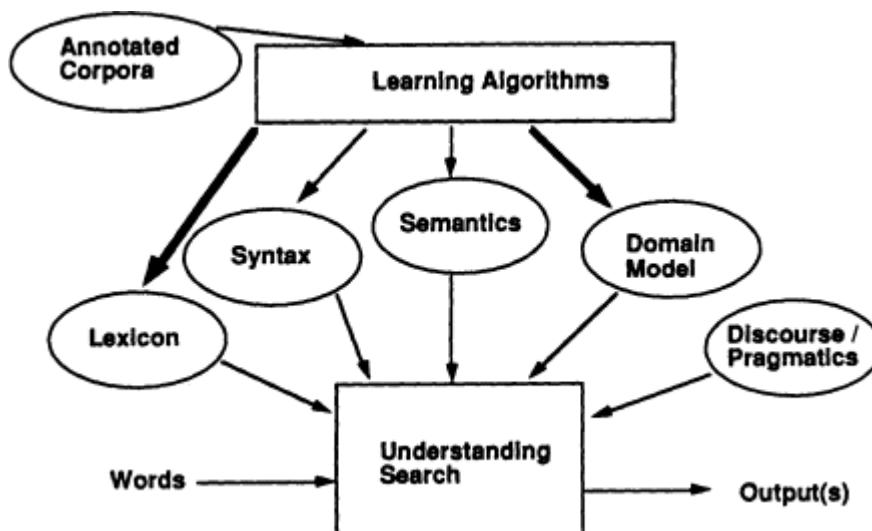


FIGURE 5 Learning from an annotated corpus.

of methods for learning the parameters necessary to predict the part of speech of an unknown word in text with only a 2 to 3 percent error rate. No systems yet create syntactic rules completely automatically, but some do use syntactic probabilities that have been trained on corpora. Semantics and domain model extraction are harder yet, but enough effort is being expended in this direction that it seems justified to expect considerable progress in these areas in the near future.

It is also possible to add probabilities to syntactic and semantic rules so that the most probable parse or the most likely interpretation is produced first. The basic operation of one system (Bobrow et al., 1992) is to determine the most probable set of concepts (word senses) and the semantic links between those concepts, given a set of local linguistic structures (called grammatical relations) and the a priori likelihood of semantic links between concepts. This domain-independent statistical search strategy works as effectively on fragments as on grammatical sentences, producing a meaning expression output for almost any input. Learning algorithms will follow. Currently, no attempt is being made to use these statistical techniques for discourse or pragmatic processing.

One factor that limits progress in using this model is that the cost of producing an appropriately annotated corpus for NL processing is significantly higher than the cost of producing an annotated corpus for speech processing. In the case of speech, annotation consists basically of transcription, which can be performed by almost anyone who knows the language being spoken. But for NL, annotation is mostly done by experts. Some lexical and syntactic annotations can be done by carefully trained people (who are not experts).

The semantic annotation that was performed for the Air Travel Information System (ATIS) data in the Advanced Research Project Agency's (ARPA) Spoken Language Systems program (Hirschman et al., 1993) is an example of a corpus that was expensive to collect and annotate, and it is not even useful for some types of research (e.g., pragmatics). Efforts are currently under way in the computational linguistics community to develop some kinds of annotation that are both useful and inexpensive.

## HOW CAN NL SYSTEMS BE APPLIED AND EVALUATED?

In speech there are a few factors, such as vocabulary size, perplexity, and training data, that roughly determine the performance of an SR system for a particular domain. In NLP there are a large number of such factors, not all easily quantifiable: vocabulary size, number of concepts in the domain, number of relations between those

concepts, amount of training data, type of annotations available, task complexity, amount of ambiguity, amount of ungrammaticality, errors in input, etc.

We need a metric similar to perplexity but one that takes into account the number of possible concepts and relations and the amount of overlap (potential ambiguity) among them. Ideally, such a metric would take into account domain and task complexity as well as syntactic and semantic complexity.

Even without a way to measure the difficulty of the input, there are a variety of ways to evaluate the performance of NL systems. Evaluation is a necessary part of any system, whether developed for research or for application. In the past few years, several methodologies have emerged for evaluating particular kinds of NL systems—spoken language systems (SLS) and message-processing systems foremost among them. The notion of domain-independent understanding and evaluation, while being actively explored by the research community, is only one of several approaches to evaluation. The methodologies that have actually been used thus far vary from one type of application (message processing, question answering, translation, information retrieval, etc.) to another.

The methodologies allow comparative evaluation of different NL systems, as well as tracking of the progress of a single NL system as it evolves over time. For example, the methodology for the SLS program can be used for both spoken language systems (with speech input) or just NL systems (by omitting the SR component and giving the NL system a word string as input). The SLS methodology works by comparing the predetermined "right answer" (the canonical answer) to answers that are produced when different SLS systems are given identical inputs and are required to use identical databases (Pallett et al., 1993). The "understanding error rate" is the percentage of utterances that are either answered incorrectly or not answered at all.

One of the current ways to evaluate NL systems (a way that has been very successful in several recent ARPA-sponsored evaluations of both text and spoken language) is to look at the output produced by a system with an NL component and try to determine whether the output is correct or incorrect. (A better metric might be appropriate or inappropriate; that is a topic of considerable discussion.) Either way, the complexity and variety of the types of output that can be produced by a system make evaluation based on output extremely difficult.

These evaluations are labor intensive and somewhat difficult to specify and carry out, but they are very important to the community.

of people doing research in this area. Current NL systems (components of SLS systems, operating in the ATIS domain, with vocabularies of around 2000 to 3000 words) achieve an understanding error rate of about 6 percent, which appears to be quite close to the threshold of real utility for applications. Detailed descriptions of the methodology, as well as the underlying databases and annotated corpora for the ATIS domain (as well as many other NL corpora), are available from the Linguistic Data Consortium at the University of Pennsylvania.

## CONCLUSIONS

What is the current state of the art in NL processing?

In question answering domains such as database interfaces, the understanding error rate is about 5 to 10 percent. The amount of effort needed to bring an NL system to this level of performance is still more substantial than we would like. In fact, it is currently the major bottleneck to the availability of NLP applications.

Portability can be defined as the ability to make an NL system (for a particular type of application, such as a database) usable in a new domain (with a new vocabulary, a new set of semantic concepts and relations). A system could be considered portable if it were possible to achieve moderate performance (perhaps 15 to 20 percent error) in a new domain using some automatic methods and a few person-weeks of human effort. The goal of portability is good performance with moderate effort.

The portability problem will probably be cracked by work that is being done in several areas simultaneously. Automatic learning (training) based on annotated corpora holds substantial promise. In addition, NL systems need to be adaptable to their users (i.e., a user should be able to tell a system when it understood something incorrectly and what the right interpretation is).

Other challenges in addition to portability include scaling up from demonstration to real applications; increasing robustness (how systems deal with unexpected novel input); feedback (what kind of help the system can give a user when an interpretation goes wrong); and determining what is an acceptable level of performance for a new NLP system.

As has been the case in speech processing, the biggest payoff comes when machines can perform at or near the level of human performance. NL systems still have a long, long way to go, but the goal (for limited domains) will soon be within our grasp. The result will be a paradigm shift that will enable designers and developers of

many types of systems (but particularly interactive systems) to incorporate NLP into their systems. Users will begin to expect their systems to understand spoken or typed commands and queries, to be able to classify bodies of text, and to extract various kinds of information from bodies of text.

## REFERENCES

- Allen, J., Natural Language Understanding, The Benjamin/Cummings Publishing Co., Menlo Park, Calif., 1987.
- Bates, M. (ed.), Proceedings of the ARPA Workshop on Human Language Technology, Princeton, N.J., March 1993, Morgan Kaufmann Publishing Co., 1993.
- Bates, M., and R. M. Weischedel (eds.), Challenges in Natural Language Processing, Cambridge University Press, Cambridge, 1993.
- Bobrow, R., R. Ingria, and D. Stallard, "Syntactic/Semantic Coupling in the BBN DELPHI System," DARPA Speech and Natural Language Workshop, Harriman, N.Y., Morgan Kaufmann Publishers, 1992.
- Gazdar, G., and C. Mellish, Natural Language Processing in LISP, Addison-Wesley, Reading, Mass., 1989.
- Hirschman, L., et al., "Multisite Data Collection and Evaluation in Spoken Language Understanding," in Bates (ed.), Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufmann, 1993.
- Montague, R., Pragmatics and Intensional Logic, *Synthese* vol. 22, pp. 68-94, 1970.
- Pallett, D., et al., "Benchmark Tests for the DARPA Spoken Language Program," in Bates (ed.), Proceedings of the ARPA Workshop on Human Language Technology, 1993.
- Smith, G. W., Computers and Human Language, Oxford University Press, Oxford, U.K., 1991.
- Weischedel, R. M., J. Carbonell, B. Grosz, M. Marcus, R. Perrault, and R. Wilensky, "Natural Language Processing," Annual Review of Computer Science, Vol. 4, 1990.
- Winograd, T., Language as a Cognitive Process, Addison-Wesley, Reading, Mass., 1983.

## BIBLIOGRAPHY

- Grosz, B., D. E. Appelt, P. Martin, and F. Pereira, TEAM: An Experiment in the Design of Transportable Natural Language Interfaces. *Artificial Intelligence*, 1985.
- Lehnert, W. G., and B. Sundheim, "A Performance Evaluation of Text Analysis Technologies," *AI Magazine*, Fall, pp. 81-94, 1991.
- Neal, J., and S. Walter, S. (eds.), Natural Language Processing Systems Evaluation Workshop, Rome Laboratory, 1991.
- Sundheim, B. (ed.), Proceedings of the Third Message Understanding Conference, 1991.
- Sundheim, B. (ed.), Proceedings of the Fourth Message Understanding Conference, 1992.
- Sundheim, B. (ed.), Proceedings of the Fifth Message Understanding Conference, 1993.

# Integration of Speech with Natural Language Understanding\*

*Robert C. Moore*

## SUMMARY

The integration of speech recognition with natural language understanding raises issues of how to adapt natural language processing to the characteristics of spoken language; how to cope with errorful recognition output, including the use of natural language information to reduce recognition errors; and how to use information from the speech signal, beyond just the sequence of words, as an aid to understanding. This paper reviews current research addressing these questions in the Spoken Language Program sponsored by the Advanced Research Projects Agency (ARPA). I begin by reviewing some of the ways that spontaneous spoken language differs from standard written language and discuss methods of coping with the difficulties of spontaneous speech. I then look at how systems cope with errors in speech recognition and at attempts to use natural language information to reduce recognition errors. Finally, I discuss how prosodic information in the speech signal might be used to improve understanding.

---

\* Preparation of this paper was supported by the Advanced Research Projects Agency under Contract No. N00014-93-C-0142 with the Office of Naval Research. The views and conclusions contained here are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency of the U.S. Government.

## INTRODUCTION

The goal of integrating speech recognition with natural language understanding is to produce spoken-language-understanding systems—that is, systems that take spoken language as their input and respond in an appropriate way depending on the meaning of the input. Since speech recognition (Makhoul and Schwartz, in this volume) aims to transform speech into text, and natural-language-understanding systems (Bates, in this volume) aim to understand text, it might seem that spoken-language-understanding systems could be created by the simple serial connection of a speech recognizer and a natural-language-understanding system. This naive approach is less than ideal for a number of reasons, the most important being the following:

- Spontaneous spoken language differs in a number of ways from standard written language, so that even if a speech recognizer were able to deliver a perfect transcription to a natural-language-understanding system, performance would still suffer if the natural language system were not adapted to the characteristics of spoken language.
- Current speech recognition systems are far from perfect transcribers of spoken language, which raises questions about how to make natural-language-understanding systems robust to recognition errors and whether higher overall performance can be achieved by a tighter integration of speech recognition and natural language understanding.
- Spoken language contains information that is not necessarily represented in written language, such as the distinctions between words that are pronounced differently but spelled the same, or syntactic and semantic information that is encoded prosodically in speech. In principle it should be possible to extract this information to solve certain understanding problems more easily using spoken input than using a simple textual transcription of that input.

This paper looks at how these issues are being addressed in current research in the ARPA Spoken Language Program.

## COPING WITH SPONTANEOUS SPOKEN LANGUAGE

### Language Phenomena in Spontaneous Speech

The participants in the ARPA Spoken Language Program have adopted the interpretation of requests for air travel information as a common task to measure progress in research on spoken language

understanding. In support of this effort, over 15,000 utterances have been collected from subjects by using either a simulated or actual spoken language Air Travel Information System (ATIS). Interpreting these utterances requires dealing with a number of phenomena that one would not encounter often in dealing with linguists' examples or even real written texts.

Among the most common types of nonstandard utterances in the data are sentence fragments, sequences of fragments, or fragments combined with complete sentences:

*six thirty a m from atlanta to san francisco what type of aircraft  
on the delta flight number ninety eight what type of aircraft  
i would like information on ground transportation city of boston  
between airport and downtown*

A particular subclass of these utterances might be dubbed "afterthoughts." These consist of an otherwise well-formed sentence followed by a fragment that further restricts the initial request:

*i'd like a return flight from denver to atlanta evening flights  
i need the cost of a ticket going from denver to baltimore a first  
class ticket on united airlines  
what kind of airplane goes from philadelphia to san francisco monday  
stopping in dallas in the afternoon first class flight*

Another important group of nonstandard utterances can be classified as verbal repairs or self-corrections, in which the speaker intends that one or more words be replaced by subsequently uttered words. In the following examples, groups of words that are apparently intended for deletion are enclosed in brackets:

*i'd like [to] a flight from washington [to] that stops in denver and  
goes on to san francisco  
[do any of these flights] [do] are there any flights that arrive after five p m  
can you give me information on all the flights [from san francisco  
no] from pittsburgh to san francisco on monday*

Some utterances involve use of metonymy, in which a word or phrase literally denoting one type of entity is used to refer to a related type of entity:

*i need flight information between atlanta and boston*

*what is the flight number that leaves at twelve twenty p m  
what delta leaves boston for atlanta*

In the first two utterances, properties of flights are attributed to flight information and flight numbers; in the third, the name *delta* is used to refer to flights on Delta Airlines.

Some utterances that perhaps could be viewed as cases of metonymy might better be interpreted simply as slips of the tongue:

*does delta aircraft fly d c tens*

In this utterance *aircraft* has simply been substituted for *airlines*, perhaps because of the phonetic similarity between the two words and semantic priming from the information being requested.

Finally, some utterances are simply ungrammatical:

*what kinds of ground transportation is available in dallas fort worth*

*okay what type of aircraft is used on a flight between san francisco to atlanta*

*what types of aircraft can i get a first class ticket from Philadelphia to dallas*

*from those show me that serve lunch*

The first example in this list is a case of lack of number agreement between subject and verb; the subject is plural, the verb singular. The second example seems to confuse two different ways of expressing the same constraint; *between san francisco and atlanta* and *from san francisco to atlanta* have been combined to produce *between san francisco to atlanta*. The final pair of examples both seem to involve deletion of function words needed to make the utterances grammatical:

*what types of aircraft can i get a first class ticket from philadelphia to dallas (on)*

*from those show me (the ones) that serve lunch*

Of course, there are also utterances that combine several of these phenomena, for example:

*[flight number] [okay flight] [let us] you have a flight number*

*going [to] from san francisco to atlanta around eight a m*

This utterance appears to begin with three repairs; it finally gets going with a somewhat odd way of asking for a flight number, *you have a flight number...*, it involves the metonymy of talking about a flight

number going somewhere rather than a flight and includes yet another repair, replacing *to* with *from*.

### Strategies for Handling Spontaneous Speech Phenomena

Spoken-language-understanding systems use various strategies to deal with the nonstandard language found in spontaneous speech. Many of these phenomena, although regarded as nonstandard, are just as regular in their patterns of use as standard language, so they can be incorporated into the linguistic rules of a natural language system. For example, one may add grammar rules to allow a sequence of syntactically unconnected noun phrase modifiers to be a complete utterance if all the modifiers can be interpreted as predication of the same class of objects, so that an utterance like *from boston to dallas on tuesday* can be interpreted as a request for flights or fares. Nonstandard but regular uses of particular lexical items can be accommodated simply by extending the lexicon. For example, common cases of metonymy can be handled this way; *flight information* and *flight number* can be lexically coded to allow the same modifiers as *flight*.

Extending the linguistic rules of a system to include nonstandard but regular patterns still leaves disfluencies and truly novel uses of language unaccounted for. To deal with these, virtually all systems developed for the ARPA ATIS task incorporate some sort of language-understanding method that does not depend on deriving a complete analysis of the input that accounts for every word. Such methods are usually described as "robust," although they are actually robust only along certain dimensions and not others. A common strategy is to have predefined patterns (case frames or templates) for the most common sorts of queries in the ATIS task and to scan the input string for words or phrases that match the elements of the pattern, allowing other words in the utterance to be skipped. The Carnegie-Mellon University (CMU) Phoenix system (Ward et al., 1992) and SRI International's Template Matcher (Jackson et al., 1991) rely exclusively on this approach. In the SRI system, for example, a key word such as *flight*, *fly*, *go*, or *travel* is used to trigger the template for a request for flight information and phrases matching patterns such as *on 8date>*, *from 8city>*, and *to 8city>* are used to fill in constraints on the flights. This allows the system to ignore disfluencies or novel language if they do not occur in parts of the utterances that are crucial for recognizing the type of request or important constraints on the request. For instance, this approach can easily process the example given above of a sentence with multiple problematic features,

*[flight number] [okay flight] [let us] you have a flight number  
going [to] from san francisco to atlanta around eight a m*

because it is fundamentally a very simple type of request, and none of the disfluencies affect the phrases that express the constraints on the answer.

This template-based approach works well for the vast majority of utterances that actually occur in the ATIS data. In principle, however, the approach would have difficulties with utterances that express more complex relationships among entities or that involve long-distance dependencies, such as in

*what cities does united fly to from san francisco*

Here the separation of *what cities* and *to* would make this utterance difficult to interpret by template-based techniques, unless a very specific pattern were included to link these together. But in fact this is a made-up example, and things like this are extremely rare, if they occur at all, in the ATIS task. Nevertheless, such possibilities have led several groups to design systems that first try to carry out a complete linguistic analysis of the input, falling back on robust processing techniques only if the complete analysis fails. The Delphi system of BBN Systems and Technologies (Stallard and Bobrow, 1992), the TINA system of the Massachusetts Institute of Technology (MIT) (Seneff, 1992), and SRI International's Gemini+TM system [a combination of the Template Matcher with SRI's Gemini system (Dowding et al., 1993)] all work this way. In the case of SRI's systems, the combination of detailed linguistic analysis and robust processing seems to perform better than robust processing alone, with the combined Gemini+TM system having about four points better performance than the Template Matcher system alone for both speech and text input in the November 1992 ATIS evaluation, according to the weighted understanding error metric (Pallet et al., 1993).<sup>1</sup> It should be noted, however, that the best-performing system in the November 1992 ATIS evaluation, the CMU Phoenix system, uses only robust interpretation methods with no attempt to account for every word of an utterance.

The robust processing strategies discussed above are fairly general and are not specifically targeted at any particular form of disfluency. There has been recent work, however, aimed specifically at the detec

---

<sup>1</sup> This measure is equal to the percentage of utterances correctly answered minus the percentage incorrectly answered, with utterances not answered omitted, so as to punish a wrong answer more severely than not answering at all.

tion and correction of verbal repairs. Utterances containing repairs constitute about 6 percent of the ATIS corpus, although repair rates as high as 34 percent have been reported for certain types of human-human dialogue (Levett, 1983). A module to detect and correct repairs has been incorporated into SRI's Gemini system (Bear et al., 1992; Dowding et al., 1993) that is keyed to particular word patterns that often signal a repair. For example, in the utterance

*can you give me information on all the flights [from san francisco  
no] from pittsburgh to san francisco on monday*

the section *from san francisco no from pittsburgh* matches a pattern of a cue word, *no*, followed by a word (*from*) that is a repetition of an earlier word. Often, as in this case, this pattern indicates that text from the first occurrence of the repeated word through the cue word should be deleted. This kind of pattern matching alone generates many false positives, so in the Gemini system a repair edit based on pattern matching is accepted only if it converts an utterance that cannot be fully parsed and interpreted into one that can. Applying this method to a training corpus of 5873 transcribed utterances, Gemini correctly identified 89 of the 178 utterances that contained repairs consisting of more than just a word fragment. Of these, 81 were repaired correctly and 8 incorrectly. An additional 15 utterances were misidentified as containing repairs. Similar results were obtained in a fair test on transcriptions of the November 1992 ATIS test set. Gemini identified 11 of the 26 repair utterances out of 756 interpretable utterances, of which 8 were repaired correctly and 3 incorrectly; 3 other utterances were misidentified as containing repairs.

It should be noted that the general robust processing methods discussed above are insensitive to repairs if they occur in sections of the utterance that are not critical to filling slots in the pattern being matched. In addition, CMU's Phoenix system incorporates one other simple method for handling repairs. If the pattern matcher finds more than one filler for the same slot, it uses the last one, on the assumption that the fillers found earlier have been replaced by repairs. This method seems to work on many of the repairs actually found in the ATIS corpus. It is easy to make up cases where this would fail, but the more complex method used by Gemini would work:

*show me flights to boston no from boston*

However, it is not clear that such utterances occur with any frequency in the ATIS task.

## ROBUSTNESS TO RECOGNITION ERRORS

Since even the best speech recognition systems make at least some errors in a substantial proportion of utterances, coping with speech recognition errors is one of the major challenges for correctly interpreting spoken language. This problem can be approached in a number of ways. If there are particular types of errors that the recognizer is especially prone to make, the natural-language-understanding system can be modified to accommodate them. For instance, *four* can be allowed in place of *for*, or the deletion of short, frequently reduced words, such as *to*, can be permitted.

For the most part, however, current systems rely on their general robust understanding capability to cope with noncritical recognition errors. Although it has not been carefully measured, anecdotally it appears that a high proportion of recognition errors made by the better-performing recognizers for the ATIS task occur in portions of the utterance that are noncritical for robust interpretation methods. It may be conjectured that this is due to the fact that most of the critical key words and phrases are very common in the training data for the task and are therefore well modeled both acoustically and in the statistical language models used by the recognizers.

The degree of robustness of current ATIS systems to speech recognition errors can be seen by examining Table 1. This table compares three different error rates in the November 1992 evaluation (Pallet et al., 1993) of the ATIS systems developed by the principal ARPA contractors working on the ATIS task: BBN Systems and Technologies (BBN), Carnegie-Mellon University (CMU), the Massachusetts Institute of Technology (MIT), and SRI International (SRI). The error rates compared are (1) the percentage of queries not correctly answered when the natural language component of the system was presented with verified transcriptions of the test utterances, (2) the percentage of queries not correctly answered when the combined speech

TABLE 1 Comparison of Understanding Error with Recognition Error in November 1992 ATIS Evaluation

System	Understanding Error w/Text, %	Understanding Error w/Speech, %	Recognition Error, %
BBN	15.0	19.0	25.2
CMU	6.5	11.2	28.9
MIT	10.9	19.2	37.8
SRI	15.2	21.6	33.8

recognition and natural language understanding system was presented with the digitized speech signal for the same utterances, and (3) the percentage of queries for which the speech recognition component of the system made at least one word recognition error in transcribing the utterance. All of these error rates are for the subset of utterances in the test set that were deemed to constitute answerable queries.

The striking thing about these results is that for all of these systems the increase in understanding error going from text input to speech input is surprisingly small in comparison with the rate of utterance recognition error. For instance, for the CMU system the rate of understanding error increased by only 4.7 percent of all utterances when a verified transcription of the test utterances was replaced by speech recognizer output, even though 28.9 percent of the recognizer outputs contained at least one word recognition error. Moreover, the rate of speech-understanding errors was much lower than the rate of speech recognition errors for all systems, even though all systems had many language-understanding errors even when provided with verified transcriptions. This shows a remarkable degree of robustness to recognition errors using methods that were primarily designed to cope with difficult, but accurately transcribed, spoken language.

## NATURAL LANGUAGE CONSTRAINTS IN RECOGNITION

### Models for Integration

Despite the surprising degree of robustness of current ATIS systems in coping with speech recognition errors, Table 1 also reveals that rates for understanding errors are still substantially higher with speech input than with text input, ranging from 1.26 to 1.76 times higher, depending on the system. One possible way to try to close this gap is to use information from the natural language processor as an additional source of constraint for the speech recognizer. Until recently, most attempts to do this have followed what might be called "the standard model":

Pick as the preferred hypothesis the string with the highest recognition score that can be completely parsed and interpreted by the natural language processor.

This model was embodied in several systems developed under the original ARPA Speech Understanding Program of the 1970s (Klatt, 1977) and also in some of the initial research in the current ARPA Spoken Language Program (Boisen et al., 1989; Chow and Roukos,

1989; Moore et al., 1989; Murveit and Moore, 1990). However, the model depends on having a natural language grammar that accurately models the speech being recognized. For the kind of messy, ill-formed spoken language presented here in the section "Language Phenomena in Spontaneous Speech," this presents a serious problem. It is highly unlikely that any conventional grammar could be devised that would cover literally everything a person might actually utter.

The dilemma can be seen in terms of the kind of natural language system, discussed in the section "Strategies for Handling Spontaneous Speech Phenomena," that first attempts a complete linguistic analysis of the input and falls back on robust processing methods if that fails. If the grammar used in attempting the complete linguistic analysis is incorporated into the speech recognizer according to the standard model, the recognizer will be overconstrained and the robust processor will never be invoked because only recognition hypotheses that can be completely analyzed linguistically will ever be selected. This means that, for cases in which the robust processor should have been used, the correct recognition hypothesis will not even be considered. On the other hand, if the robust language processor were incorporated into the speech recognizer according to the standard model, it would provide very little information since it is designed to try to make sense out of almost any word string.

A number of modifications of the standard model have been proposed to deal with this problem. One method is to use a highly constraining grammar according to the standard model but to limit the number of recognition hypotheses the grammar is allowed to consider. The idea is that, if none of the top  $N$  hypotheses produced by the recognizer can successfully be analyzed by the grammar, it is likely that the correct recognition hypothesis is not allowed by the grammar, and a second attempt should be made to interpret the recognizer output using robust processing. The parameter  $N$  can be empirically estimated to give optimal results for a particular grammar on a particular task. Another possibility is to allow parsing a hypothesis as a sequence of grammatical fragments with a scoring metric that rewards hypotheses that can be analyzed using fewer fragments.

An additional problem with the standard model is that it does not take into account relative likelihoods of different hypotheses. All utterances that are allowed by the grammar are treated as equally probable. This differs from the N-gram statistical language models commonly used in speech recognition that estimate the probability of a given word at a particular point in the utterance based on the one or two immediately preceding words. Baker (1979) developed an automatic training method, the inside-outside algorithm, that allows

such techniques to be extended to probabilistic context-free grammars. Since most grammars used in natural-language-processing systems are based to some extent on context-free grammars, Baker's or related methods may turn out to be useful for developing probabilistic natural language grammars for use in speech recognition. This approach appears to leave at least two important problems to be addressed, however.

First, while the grammars used in natural-language-processing systems are usually based on the context-free grammars, they also usually have augmentations that go beyond simple context-free grammars. Recently, grammars based on the unification of grammatical categories incorporating features-value structures (Shieber, 1986) have been widely used. If the value spaces are finite for all the features used in a particular grammar, the grammar is formally equivalent to a context-free grammar, but for any realistic unification grammar for natural language the corresponding context-free grammar would be so enormous (due to all the possible combinations of feature values that would have to be considered) that it is extremely doubtful that enough training data could either be obtained or processed to provide reliable models via the inside-outside algorithm. This suggests that, at best, a carefully selected context-free approximation to the full grammar would have to be constructed.

A second problem derives from the observation (Church et al., 1989) that particular lexical associations, such as subject-verb, verb-object, or head-modifier, appear to be a much more powerful source of constraint in natural language than the more abstract syntactic patterns typically represented in natural language grammars. Thus, in predicting the likelihood of a combination such as *departure time*, one cannot expect to have much success by estimating the probability of such noun-noun combinations, independently of what the nouns are, and combining that with context-independent estimates of an arbitrary noun being *departure* or *time*. Yet in the model of probabilistic context-free grammar to which the inside-outside algorithm applies, this is precisely what will happen, unless the grammar is carefully designed to do otherwise. If probabilistic models of natural language are not constructed in such a way that lexical association probabilities are captured, those models will likely be of little benefit in improving recognition accuracy.

### Architectures for Integration

Whatever model is used for integration of natural language constraints into speech recognition, a potentially serious search problem

must be addressed. The speech recognizer can no longer simply find the best acoustic hypothesis; it must keep track of a set of acoustic hypotheses for the natural language processor to consider. The natural language processor similarly has to consider multiple recognition hypotheses, rather than a single determinate input string. Over the past 5 years, three principal integration architectures for coping with this search problem have been explored within the ARPA Spoken Language Program: word lattice parsing, dynamic grammar networks, and N-best filtering or rescoring.

## **Word Lattice Parsing**

Word lattice parsing was explored by BBN (Boisen et al., 1989; Chow and Roukos, 1989) in the early stages of the current ARPA effort. In this approach the recognizer produces a set of word hypotheses, with an acoustic score for each potential pair of start and end points for each possible word. A natural language parser is then used to find the grammatical utterance spanning the input signal that has the highest acoustic score. Word lattice parsing incorporates natural language constraints in recognition according to the standard model, but it results in extremely long processing times, at least in recent implementations. The problem is that the parser must deal with a word lattice containing thousands of word hypotheses rather than a string of just a few words. More particularly, the parser must deal with a large degree of word boundary uncertainty. Normally, a word lattice of adequate size for accurate recognition will contain dozens of instances of the same word with slightly different start and end points. A word lattice parser must treat these, at least to some extent, as distinct hypotheses. One approach to this problem (Chow and Roukos, 1989) is to associate with each word or phrase a set of triples of start points, end points, and scores. Each possible parsing step is then performed only once, but a dynamic programming procedure must also be performed to compute the best score for the resulting phrase for each possible combination of start and end points for the phrase.

## **Dynamic Grammar Networks**

Dynamic grammar networks (Moore et al., 1989; Murveit and Moore, 1990) were developed to address the computational burden in word lattice parsing posed by the need for the parser to deal with acoustic scores and multiple possible word start and end points. In this approach a natural language parser is used to incrementally generate

the grammar-state-transition table used in the standard hidden Markov model (HMM) speech recognition architecture. In an HMM speech recognizer, a finite-state grammar is used to predict what words can start in a particular recognition state and to specify what recognition state the system should go into when a particular word is recognized in a given predecessor state. Dynamic programming is used to efficiently assign a score to each recognition state the system may be in at a particular point in the signal.

In HMM systems the finite-state grammar is represented as a set of state-word-state transitions. Any type of linguistic constraints can, in fact, be represented as such a set, but for a nontrivial natural language grammar the set will be infinite. The dynamic-grammar-network approach computes the state-transition table needed by the HMM system incrementally, generating just the portion necessary to guide the pruned search carried out by the recognizer for a particular utterance. When a word is successfully recognized beginning in a given grammar state, the recognizer sends the word and the state it started in to the natural language parser, which returns the successor state. To the parser, such a state encodes a parser configuration. When the parser receives a state-word pair from the recognizer, it looks up the configuration corresponding to the state, advances that configuration by the word, creates a name for the new configuration, and passes back that name to the recognizer as the name of the successor state. If it is impossible, according to the grammar, for the word to occur in the initial parser configuration, the parser sends back an error message to the recognizer, and the corresponding recognition hypothesis is pruned out. Word boundary uncertainty in the recognizer means that the same word starting in the same state can end at many different points in the signal, but the recognizer has to communicate with the parser only once for each state-word pair. Because of this, the parser does not have to consider either acoustic scores or particular start and end points for possible words, those factors being confined to the recognizer.

The dynamic-grammar-net approach succeeds in removing consideration of acoustic scores and word start and end points from the parser, but it too has limitations. The state space tends to branch from left to right, so if a group of utterance hypotheses differ in their initial words but have a later substring in common, that substring will be analyzed multiple times independently, since it arises in different parsing configurations. Also, in the form presented here, the dynamic-grammar-net approach is tied very much to the standard model of speech and natural language integration. It is not immediately clear how to generalize it so as not to overconstrain the recognizer in cases where the utterance falls outside the grammar.

## N-best Filtering or Rescoring

N-best filtering and rescoring were originally proposed by BBN (Chow and Schwartz, 1989; Schwartz and Austin, 1990). This is a very simple integration architecture in which the recognizer enumerates the N-best full recognition hypotheses, which the natural language processor then selects from. The standard model of speech and natural language integration can be implemented by N-best filtering, in which the recognizer simply produces an ordered list of hypotheses, and the natural language processor chooses the first one on the list that can be completely parsed and interpreted. More sophisticated models can be implemented by N-best rescoring, in which the recognition score for each of the N-best recognition hypotheses is combined with a score from the natural language processor, and the hypothesis with the best overall score is selected.

The advantage of the N-best approach is its simplicity. The disadvantage is that it seems impractical for large values of  $N$ . The computational cost of the best-known method for exact enumeration of the N-best recognition hypotheses (Chow and Schwartz, 1989) increases linearly with  $N$ ; but an approximate method exists (Schwartz and Austin, 1990) that increases the computational cost of recognition only by a small constant factor independent of  $N$ . There is no reported method, however, for carrying out the required natural language processing in time less than linear in the number of hypotheses. In practice, there seem to be no experiments that have reported using values of  $N$  greater than 100, and the only near-real-time demonstrations of systems based on the approach have limited  $N$  to 5. To put this in perspective, in information theoretic terms, an N-best system that selects a single hypotheses from a set of 64 hypotheses would be providing at most 6 bits of information per utterance. On the other hand, it has not proved difficult to develop purely statistical language models for particular tasks that provide 6 bits or more of information per word.<sup>2</sup>

However, if the basic recognizer is good enough that there is a very high probability of the correct hypothesis being in that set of 64, those 6 bits per utterance may be enough to make a practical difference.

---

<sup>2</sup> For a 1000-word recognition task, a perplexity-15 language model reduces the effective vocabulary size by a factor of about 67, which is about a 6-bit per word reduction in entropy.

### Integration Results

In 1992, BBN (Kubala et al., 1992) and MIT (Zue et al., 1992) reported results on the integration of natural language constraints into speech recognition for the ATIS task. BBN used an N-best integration scheme in which strict grammatical parsing was first attempted on the top five recognition hypotheses, choosing the hypothesis with the best recognition score that could be fully interpreted. If none of those hypotheses was fully interpretable, the process was repeated using a robust processing strategy. On a development test set, BBN found that this reduced the weighted understanding error from 64.6 percent, when only the single best recognizer output was considered, to 56.6 percent—a 12.4 percent reduction in the error rate. Taking the top 20 hypotheses instead of the top five also improved performance compared with taking only the single top hypothesis, but the improvement was less than when consideration was limited to the top five.

The MIT experiment was more complex, because it used natural language constraints in two different ways. First, a probabilistic LR parser was integrated into the recognition search (using an architecture that somewhat resembled dynamic grammar nets) to incorporate a language probability into the overall recognition score. The LR parser was modified to allow parsing an utterance as a sequence of fragments if no complete parse allowed by the grammar could be found. Then the top 40 recognition hypotheses were filtered using the complete natural language system. This reduced the word recognition error from 20.6 to 18.8 percent (an 8.7 percent reduction) and the utterance recognition error from 67.7 to 58.1 percent (a 13.4 percent reduction) on a development test set, compared with the best version of their recognizer incorporating no natural language constraints.

### SPEECH CONSTRAINTS IN NATURAL LANGUAGE UNDERSTANDING

While most efforts toward integration of speech and natural language processing have focused on the use of natural language constraints to improve recognition, there is much information in spoken language beyond the simple word sequences produced by current recognizers that would be useful for interpreting utterances if it could be made available. Prosodic information in the speech signal can have important effects on utterance meaning. For example, since in the ATIS task the letters B, H, and BH are all distinct fare codes, the two utterances

*What do the fare codes BH and K mean?*

*What do the fare codes B, H, and K mean?*

would differ only prosodically but have very different meanings.

In a preliminary study of the use of prosodic information in natural language processing, Bear and Price (1990) reported on an experiment in which the incorporation of prosodic information into parsing reduced syntactic ambiguity by 23 percent on a set of prosodically annotated sentences. Another area where information from the speech signal would undoubtedly be useful in natural language processing is in the detection of verbal repairs. While no experiment has yet been reported that actually used speech information to help correct repairs, Bear et al., (1992) and Nakatani and Hirschberg (1993) have identified a number of acoustic cues that may be useful in locating repairs.

## CONCLUSIONS

Work on the ATIS task within the ARPA Spoken Language Program is ongoing, but a number of conclusions can be drawn on the basis of what has been accomplished so far. Perhaps the most significant conclusion is that natural-language-understanding systems for the ATIS task have proved surprisingly robust to recognition errors. It might have been thought a priori that spoken language utterance understanding would be significantly worse than utterance recognition, since recognition errors would be compounded by understanding errors that would occur even when the recognition was perfect. The result has been quite different, however, with the robustness of the natural language systems to recognition errors more than offsetting language-understanding errors.

Nevertheless, understanding error remains 20 to 70 percent higher with speech input than with text input. Attempts to integrate natural language constraints into recognition have produced only modest results so far, improving performance by only about 9 to 13 percent, depending on how performance is measured.

Is it possible to do better? So far, relatively few ideas for the incorporation of natural language constraints into recognition of spontaneous speech have been tested, and there may be many ways in which the approaches might be improved. For example, published reports of experiments conducted so far do not make it clear whether strong semantic constraints were used or how well important word associations were modeled. Compared with where the field stood

only 3 or 4 years ago, however, great progress has certainly been made, and there seems no reason to believe it will not continue.

## REFERENCES

- Baker, J. (1979), "Trainable Grammars for Speech Recognition," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, J. J. Wolf and D. H. Klatt, eds., Massachusetts Institute of Technology, Cambridge, Mass.
- Bear, J., and P. Price (1990), "Prosody, Syntax and Parsing," in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pa., pp. 17-22.
- Bear, J., J. Dowding, and E. Shriberg (1992), "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialogue," in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Del., pp. 56-63.
- Boisen, S., et al. (1989), "The BBN Spoken Language System," in *Proceedings of the Speech and Natural Language Workshop*, February 1989, Philadelphia, Pa., pp. 106-111, Morgan Kaufman Publishers, San Mateo, Calif..
- Chow, Y. L., and S. Roukos (1989), "Speech Understanding Using a Unification Grammar," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, pp. 727-730.
- Chow, Y. L., and R. Schwartz (1989), "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses," in *Proceedings of the Speech and Natural Language Workshop*, Cape Cod, Mass., pp. 199-202, Morgan Kaufman Publishers, San Mateo, Calif.
- Church, K., et al. (1989), "Parsing, Word Associations and Typical Predicate-Argument Relations," in *Proceedings of the Speech and Natural Language Workshop*, Cape Cod, Mass., pp. 75-81, Morgan Kaufman Publishers, San Mateo, Calif.
- Dowding, J., et al. (1993), "Gemini: A Natural Language System for Spoken-Language Understanding," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 54-61.
- Jackson, E., et al. (1991), "A Template Matcher for Robust NL Interpretation," in *Proceedings of the Speech and Natural Language Workshop*, Pacific Grove, Calif., pp. 190-194, Morgan Kaufman Publishers, San Mateo, Calif.
- Klatt, D. H. (1977), "Review of the ARPA Speech Understanding Project," *Journal of the Acoustical Society of America*, Vol. 62, No. 6, pp. 1345-1366.
- Kubala, F., et al. (1992), "BBN Byblos and HARC February 1992 ATIS Benchmark Results," in *Proceedings of the Speech and Natural Language Workshop*, Harriman, N.Y., pp. 72-77, Morgan Kaufman Publishers, San Mateo, Calif.
- Levitt, W. (1983), "Monitoring and Self-Repair in Speech," *Cognition*, Vol. 14, pp. 41104.
- Moore, R., F. Pereira, and H. Murveit (1989), "Integrating Speech and Natural-Language Processing," in *Proceedings of the Speech and Natural Language Workshop*, Philadelphia, Pa., pp. 243-247, Morgan Kaufman Publishers, San Mateo, Calif.
- Murveit, H., and R. Moore (1990), "Integrating Natural Language Constraints into HMM-Based Speech Recognition," in *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, N.M., pp. 573576.

- Nakatani, C., and J. Hirschberg (1993), "A Speech-First Model for Repair Detection and Correction," Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp. 46-53.
- Pallet, D. S., et al. (1993), "Benchmark Tests for the DARPA Spoken Language Program," Proceedings of the ARPA Workshop on Human Language Technology, Plainsboro, N.J., pp. 7-18, Morgan Kaufman Publishers, San Francisco, Calif.
- Schwartz, R., and S. Austin (1990), "Efficient, High-Performance Algorithms for N-Best Search," Proceedings of the Speech and Natural Language Workshop, Hidden Valley, Pa., pp. 6-11, Morgan Kaufman Publishers, San Mateo, Calif.
- Seneff, S. (1992), "A Relaxation Method for Understanding Spontaneous Speech Utterances," Proceedings of the Speech and Natural Language Workshop, Harriman, N.Y., pp. 299-304, Morgan Kaufman Publishers, San Mateo, Calif.
- Shieber, S. M., (1986), An Introduction to Unification-Based Approaches to Grammar, Center for the Study of Language and Information, Stanford University, Stanford, Calif.
- Stallard, D., and R. Bobrow (1992), "Fragment Processing in the DELPHI System," Proceedings of the Speech and Natural Language Workshop, Harriman, N.Y., pp. 305-310, Morgan Kaufman Publishers, San Mateo, Calif.
- Ward, W., et al. (1992), "Speech Recognition in Open Tasks," Proceedings of the Speech and Natural Language Workshop, Harriman, N.Y., pp. 78-83, Morgan Kaufman Publishers, San Mateo, Calif.
- Zue, V., et al. (1992), "The MIT ATIS System: February 1992 Progress Report." Proceedings of the Speech and Natural Language Workshop, Harriman, N.Y., pp. 8488, Morgan Kaufman Publishers, San Mateo, Calif.



## **APPLICATIONS OF VOICE- PROCESSING TECHNOLOGY I**



# A Perspective on Early Commercial Applications of Voice-Processing Technology for Telecommunications and Aids for the Handicapped

*Chris Seelbach*

## SUMMARY

The Colloquium on Human-Machine Communication by Voice highlighted the global technical community's focus on the problems and promise of voice-processing technology, particularly, speech recognition and speech synthesis. Clearly, there are many areas in both the research and development of these technologies that can be advanced significantly. However, it is also true that there are many applications of these technologies that are capable of commercialization now. Early successful commercialization of new technology is vital to ensure continuing interest in its development. This paper addresses efforts to commercialize speech technologies in two markets: telecommunications and aids for the handicapped.

## INTRODUCTION

Two voice-processing technologies, speech recognition and speech synthesis, have reached the point that they are ready for commercial application. Speech recognition applications using small vocabularies deliver significant cost reduction for the service providers and also expand markets to rotary telephone users. Speech synthesis provides cost reduction and expanded services for its users, despite "nonhuman" sound. For handicapped markets these technologies provide

users with increased mobility and control of devices, such as computers and telephones not otherwise available to them.

The industry must be able to work in this environment where significant technological advances are possible yet growing numbers of commercial applications deliver real benefits. To date, too many potential users of these technologies have been frightened away by examples of the improvements necessary to make the technology "really work." In this environment the challenge for system integrators is to learn how to apply developing technology in ways that deliver results, rather than disappointment, because the match between the job to be done and the application of the available technology was inappropriate.

In addition, the technical community must continue its efforts to utilize the successes and failures in a way that leads to successful identification of the "right" applications and commercialization of the "right" technology. Early successful commercialization of the right level of technology in the right applications will benefit users, service providers, and the research community.

In today's world of cost cutting, reengineering, and impatience with long lead time projects, investors are looking for a near-term payback for their interest. The research community does itself a disservice by not understanding this and providing for near-term commercial demonstrations. Examples are plentiful. It took 20 years for speech recognition to be deployed widely in a telecommunications application. Speech synthesis is only now being deployed on a large scale for reverse directory applications.

The initial large-scale, telephone-based commercial speech recognition application was a simple "yes" or "no" recognition. Because it was so "simple" the research community was not interested—researchers wanted to solve the really big problems. In addition, the application required working with users, systems integrators, and human factors professionals. And live, messy, real-world trials were essential to explore uncharted areas. These were not the normal directions for the research community, causing researchers to revert to ivory tower "real research."

But the payout for this simple application was understood to be large from the earliest days. The payout was recognized, but the direction was different and the technical research was not as challenging as others. As a result, the speech recognition community lost out on an opportunity to demonstrate the viability of this technology for at least 5 to 10 years.

This is not an isolated case, but these times call for recognizing

the need to get new technology into the marketplace early in order to:

- demonstrate its viability, even in simple uses;
- build credibility for more research;
- use systems integrators, human factors professionals, and others to broaden the research base; and
- make end users part of the process and use all features of the application to make the technology work better.

### **CURRENT COMMERCIAL APPLICATIONS: TELEPHONE BASED**

Successful application experience on a large scale is occurring in applications of speech recognition that deliver large cost reductions. Automation of operator services is the largest ongoing commercial application, at first using "yes" and "no" to save hundreds of millions of dollars a year for telephone companies, initially in the United States and Canada. The vocabularies have been expanded to include selection of paying choice (e.g., collect, bill-to-third-party) as well as help commands such as "operator."

Early deployment is increasing attention on user interface issues and spurring advances to meet additional early user challenges. For example, the initial applications of the early 1990s are being expanded to handle larger vocabularies, "out-of-vocabulary words," and the ability to speak over prompts (called "barge in"). In fact, deployment of "simple" technology is aiding the research community's efforts by providing large-scale use of the technology, which highlights areas for priority research that might not have been as high a priority before.

In addition, deployment of "simple" technology gets systems integrators involved with the technology earlier. In this case knowledge of the applications is vital to successful technology commercialization. Use of unique aspects of the applications vocabulary, the work process, and other aspects of the application can enhance the success of the technology. None of this would happen without the early involvement of application-knowledgeable systems integrators.

Speech recognition and synthesis technologies are affected more than other recent new technologies by specific applications factors and user interface issues. Successful commercialization of these technologies will not happen unless systems integrators and human factors professionals are involved at early stages. The technical research community is recognizing this, although later than it should have.

Other applications providing call routing, directory assistance, and speaker identification are being deployed by telephone companies worldwide. Initial deployment is in Canada and the United States, led by Northern Telecom and AT&T.

In addition, the use of speech recognition for access to information services is growing with both telephone and independent information service providers. The lack of Touch-Tone dialing is a big incentive to deploying speech recognition in this applications area. In the United States 30 percent of phones are rotary, while in Europe it varies from 25 percent in Scandinavia to 80 percent in Germany. A few early applications in Japan, the United States, and Europe have been deployed for 5 to 10 years, and service bureaus report that on services where speech recognition is advertised in the United States, 30 percent of the callers use it. Reports on the use of speech recognition in Japan and Europe exceed even these results.

Speech synthesis is less broadly deployed, primarily because of dissatisfaction with the "nonhuman" sound that is produced. Applications for internal company use such as dispatch are spreading, but where interaction with customers or the public is involved most organizations have been reluctant to use it. However, some believe that having a "nonhuman" sound is preferable in situations where people become confused with what they are being asked to do on telephone systems because they do not know they are speaking to a computer and thus must be more precise.

Applications for reverse directory assistance are on the horizon as a potential large-scale commercialization effort by a number of telephone companies. This will help the entire industry as experience is gained on what is acceptable and what is not in a well-chosen application done with care.

### CURRENT COMMERCIAL APPLICATIONS: AIDS TO THE HANDICAPPED

This market area uses signal-processing technologies to enhance hearing-aid performance. Hearing loss affects more people than any other disability—over 3 million people in the United States. In addition, voice-processing technologies are used to provide speech output for the blind and control by voice for devices, computers, and telephones for blind and physically handicapped people.

For disabled people, even limited speech recognition increases their control over such things as beds and wheelchairs and allows some to use computers and telephones. Speech synthesis also pro

vides much benefit to blind people in hearing the output of computers and other devices.

While this market is much more forgiving about imperfect technology because of the benefits offered, other attributes of the market have limited technology deployment. In applications of voice-processing technology for the disabled, many users have special, specific needs. These needs often require customized systems that are expensive to develop and do not lead to large enough markets for "generic" products to encourage widespread use. Thus, the costs to deliver benefits are often very high.

In addition, the incorporation of voice-processing technologies in large-scale applications to date has been expensive relative to the underlying cost of the system or device. So hospital beds and wheelchairs with speech control are still small specialized markets. However, this market shares with the telephone market the need to involve human factors professionals and systems integrators early in the commercialization process. With a broader market, lower costs, and more adaptable systems, the use of voice-processing technology will grow.

## CONCLUSION

While it is recognized that many improvements in voice-processing technologies are possible, the commercialization of current technologies is under way. Greater involvement of human factors professionals and systems integrators is enhancing the possibility of commercial success. The global research community needs to continue its impressive efforts at expanding the capability of the technologies while encouraging and learning from the commercialization efforts.

# Applications of Voice-Processing Technology in Telecommunications

*Jay G. Wilpon*

## SUMMARY

As the telecommunications industry evolves over the next decade to provide the products and services that people will desire, several key technologies will become commonplace. Two of these, automatic speech recognition and text-to-speech synthesis, will provide users with more freedom on when, where, and how they access information. While these technologies are currently in their infancy, their capabilities are rapidly increasing and their deployment in today's telephone network is expanding. The economic impact of just one application, the automation of operator services, is well over \$100 million per year. Yet there still are many technical challenges that must be resolved before these technologies can be deployed ubiquitously in products and services throughout the worldwide telephone network. These challenges include:

- *High level of accuracy*—The technology must be perceived by the user as highly accurate, robust, and reliable.
- *Easy to use*—Speech is only one of several possible input/output modalities for conveying information between a human and a machine, much like a computer terminal or Touch-Tone® pad on a telephone. It is *not* the final product. Therefore, speech technologies must be *hidden* from the user. That is, the burden of using the technology must be on the technology itself.

- *Quick prototyping and development of new products and services*—The technology must support the creation of new products and services based on speech in an efficient and timely fashion.

In this paper I present a vision of the voice-processing industry with a focus on the areas with the broadest base of user penetration: speech recognition, text-to-speech synthesis, natural language processing, and speaker recognition technologies. The current and future applications of these technologies in the telecommunications industry will be examined in terms of their strengths, limitations, and the degree to which user needs have been or have yet to be met. Although noteworthy gains have been made in areas with potentially small user bases and in the more mature speech-coding technologies, these subjects are outside the scope of this paper.

## INTRODUCTION

As the telecommunications industry evolves over the next decade to provide the products and services that people will desire, several key technologies will become commonplace. Two of these, automatic speech recognition (ASR) and text-to-speech synthesis (TTS), will provide users with more freedom regarding when, where, and how they can access information. Although these technologies are currently in their infancy, their capabilities are increasing rapidly and their use in today's telephone network is expanding.

Beginning with advances in speech coding, which now allows for high-speed transmission of audio signals, speech-processing technologies and telecommunications are the perfect marriage of a technology and an industry. Currently, the voice-processing market is projected to be over \$1.5 billion by 1994 and is growing at about 30 percent a per year (Meisel; Oberteuffer; The Yankee Group, 1991). [Figure 1](#) shows a plot of the projected growth of voice-processing equipment sales from 1989 to 1993. The two driving forces behind this growth are (1) the increased demand for interactive voice services such as voice response and voice messaging and (2) the rapid improvement in speech recognition and synthesis technologies.

[Figures 2](#) and [3](#) shows the current (as of December 1991) distribution of market share for the voice-messaging and voice response markets, respectively. These figures show Octel being the market leader for voice-messaging systems and AT&T the market leader in voice response systems. The data also indicate that there is no one dominant system provider in either product family. These data obviously represent a maturing industry and a mature technology (speech cod

ing). The contributions for speech recognition and text-to-speech synthesis technologies are minimal at present but are growing rapidly.

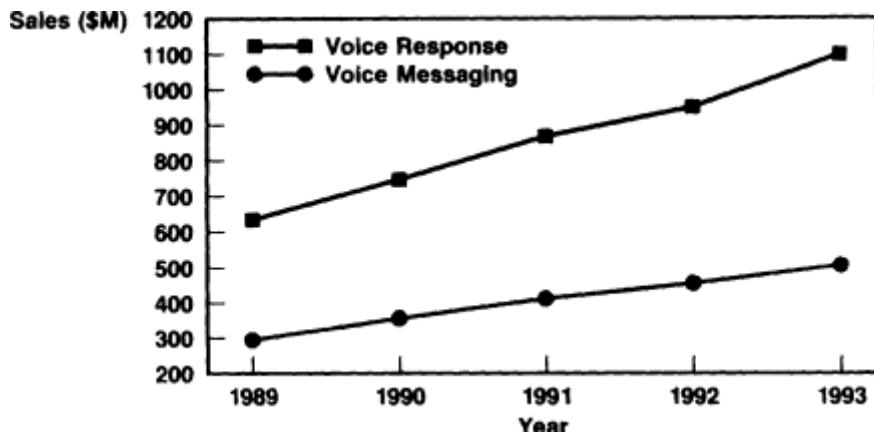


FIGURE 1 Plot of the growth in sales of voice processing equipment from 1989 to 1993 (from The Yankee Group, 1991).

Current applications using speech recognition and text-to-speech synthesis technologies center around two areas: those that provide cost reduction [e.g., AT&T and Bell Northern Research's (BNR) automation of some operator functions and NYNEX and BNR's attempt to automate portions of directory assistance] and those that provide for new revenue opportunities [e.g., AT&T's Who's Calling service,

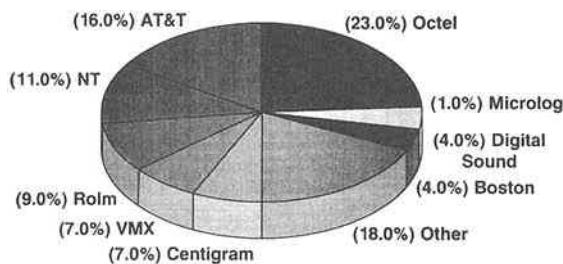


FIGURE 2 Distribution of the market share of the voice-messaging market in 1991 (from The Yankee Group, 1991).

NYNEX's directory assistance call completion service, BNR's stock quotation service, and Nippon Telegraph & Telephone's (NTT) banking by phone service].

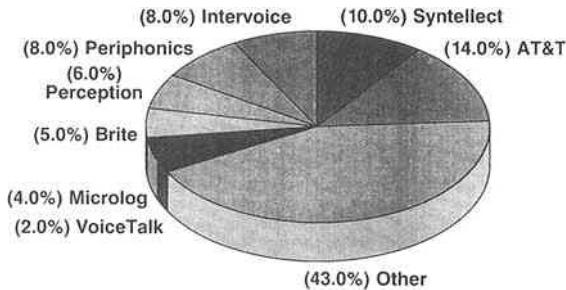


FIGURE 3 Distribution of the market share of the voice response market in 1991 (from The Yankee Group, 1991).

Yet in the face of this potentially large market, a quarter century ago the influential John Pierce wrote an article questioning the prospects of one technology, speech recognition, and criticizing the "mad inventors and unreliable engineers" working in the field. In his article entitled "Whither speech recognition," Pierce argued that speech recognition was futile because the task of speech understanding is too difficult for any machine (Pierce, 1969). Such a speech-understanding system would require tremendous advances in linguistics, natural language, and knowledge of everyday human experiences. In this prediction he was completely correct: there is still no speech recognizer that can transcribe natural speech as well as a trained stenographer, because no machine has the required knowledge and experience of human language. Furthermore, this ultimate goal is still not within sight in 1994. Pierce went on to describe the motivation for speech recognition research: "The attraction [of speech recognition] is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon." His influential article was successful in curtailing, but not stopping, speech recognition research.

What Pierce's article failed to foretell was that even limited success in speech recognition—simple, small-vocabulary speech recognizers—would have interesting and important applications, especially within

the telecommunications industry. In 1980 George Doddington, in another "Whither speech recognition?" article, pointed this out (Doddington, 1980). He emphasized that it was unnecessary to build the ultimate speech-understanding system with full human capabilities to get simple information over the telephone or to give commands to personal computers. In the decade since Doddington's article, tens of thousands of these "limited" speech recognition systems have been put into use, and we now see the beginnings of a telecommunications-based speech recognition industry (Bossemeyer and Schwab, 1991; Franco, 1993; Jacobs et al., 1992; Lennig, 1992; Oberteuffer; Yashchin et al., 1992). The economic impact of just one application, the automation of operator services, is well over \$100 million a year. However, there are still many technical challenges that must be surmounted before universal use of ASR and TTS technologies can be achieved in the telephone network. These challenges include:

- *High level of accuracy.* The technology must be perceived by the user as highly accurate, robust, and reliable. Voice-processing systems must operate under various conditions—from quiet living rooms to noisy airport terminals—while maintaining high performance levels for all conditions. Over time, as a worldwide wireless telecommunications network becomes a reality, speech technology *must* grow to provide the desired interface between the network and the communications services that users will demand.
- *Easy to use.* Speech is only one of several possible input/output modalities for conveying information between a human and a machine, much like a computer terminal or Touch-Tone pad on a telephone. It is *not* the final product. Therefore, speech technologies must be *hidden* from the user. That is, the burden of using the technology must be on the technology itself. For example, TTS systems must be natural and pleasant sounding, and ASR systems must be able to recognize predefined vocabulary words even when nonvocabulary words are also uttered.
- *Quick prototyping and development of new products and services.* The technology must support the creation of new product and service ideas based on speech in an efficient and timely fashion. Users should be not required to wait weeks or months for new products or services.

In this paper, I present a vision of the voice-processing industry, with a focus on the areas with the broadest base of user penetration: speech recognition, text-to-speech synthesis, natural language processing, and speaker recognition technologies. Current and future applications of these technologies in the telecommunications indus

try will be examined in terms of their strengths, limitations, and the degree to which user needs have been or have yet to be met. Basic research is discussed elsewhere in this volume. In this paper, I discuss only the basic technologies as they relate to telecommunications-based applications needs. Although noteworthy gains have been made in areas with potentially small user bases and in the more mature speech-coding technologies, these subjects are outside the scope of this paper.

## THE VISION

At AT&T we have developed a vision for voice processing in the telecommunications industry that will carry us into the next century:

*To have natural, continuous, two-way communication between humans and machines in any language, so that people have easier access to one another, to information, and to services—anytime and anywhere.*

This is a very ambitious vision and one that will take decades to achieve. *Natural, continuous, two-way communication*—speech recognition technology can currently support only small vocabularies spoken in a rather stylized fashion (Bossemeyer and Schwab, 1991; Jacobs et al., 1992; Lennig, 1990; Wilpon et al., 1990), and while a text-to-speech system can produce intelligible speech from practically any text it is presented with, it is anything but natural sounding. *Two-way communication* implies being able to translate speech from one language to another so that people can communicate across language barriers—a tall order for current state-of-the-art techniques (Hutchins and Somers, 1992; Morimoto et al., 1990; Rabiner and Juang, 1993; Roe et al., 1992; Waibel et al., 1991). *So that people have easier access to one another, to information, and to services* implies that we must be able to extract from a speech signal relevant information that can provide a computer with the data it needs to obtain, infer, create, or compute the information desired by the user. We are just beginning to understand how to incorporate natural language processing into the speech recognition world so that the meaning of a user's speech can be extracted. This research is in its infancy and may require more than a decade of work before viable solutions can be found, developed, and deployed (Hirschman et al., 1992; Marcus, 1992; Proceedings of the DARPA Speech and Natural Language Workshop, 1993). We are far from having such technology ready for deployment within the telecommunications industry. *Anytime and anywhere*—this, too, is a tall order. Technology must be robust enough to work equally well from the quietest ones (e.g., an office) to the noisiest ones (e.g., an airport).

or moving car). Users cannot be bothered with having to *think* about whether the technology will work. It either does and will become ubiquitous in society or it does not and will be relegated to niche applications.

Visions like this are what drives the speech community. Someday it will become a reality. It is important to understand that speech technology is *not* the final product. It is only another modality of input and output (much like keyboards and Touch-Tone pads), which will provide humans with an easier, friendlier interface to the services desired. While we wait for our vision to be realized, there are many so-called "low-hanging-fruit" telecommunications-based applications that current speech technologies can support that do not need the full capabilities just described. Many of these are discussed in the sections that follow.

## THE ART OF SPEECH RECOGNITION AND SYNTHESIS

Current speech recognition and text-to-speech synthesis practices encompass engineering art as well as scientific knowledge. Fundamental knowledge of speech and basic principles of pattern matching have been essential to the success of speech recognition over the past 25 years. Knowledge of basic linguistics and signal-processing techniques has done the same for synthesis. That said, the *art* of successful engineering is critically important for applications using these technologies. There is an important element of craftsmanship in designing a successful speech recognition or text-to-speech-based application. Knowledge of the task also helps ASR- and TTS-based applications be tuned to the user's requirements. Often, this engineering art is developed through trial and error. It should be emphasized that improving the engineering art is a proper and necessary topic for applied research.

The art of speech recognition and synthesis technology has improved significantly in the past few years, further opening up the range of possible applications (Roe and Wilpon, 1993). For speech recognition some of the advances are:

- *Wordspotting.* We are a long way from understanding fluently spoken spontaneous speech. However, some very simple elements of language understanding have been successfully developed and deployed. The ability to spot key sounds in a phrase is the first step toward understanding the essence of a sentence even if some words are not or cannot be recognized. For example, in the sentence *I'd like to make a collect call please*, the only word that must be recognized in

an operator services environment is the key word *collect*. Given that hundreds of millions of potential users will be able to pick up their telephones and use a speech recognizer to perform some task, to assume that the users will strictly adhere to the protocol of speaking only words that the recognizer understands is grossly naive. Thus, wordspotting, or the ability to recognize key words from sentences that contain both key words and nonkey words, is essential for any telecommunications-based application (Rohlicek et al., 1989; Rose and Hofstetter, 1992; Sukkar and Wilpon, 1993; Wilpon et al., 1990).

- "*Barge in.*" When talking with a person, it is desirable to be able to interrupt the conversation. In most current telephone-based voice response systems, it is possible to interrupt a prompt using Touch-Tones. This capability has been extended to allow users the option to speak during a prompt and have the system recognize them. "Barge in" provides a necessary, easy-to-use capability for customers and, as with wordspotting, is an essential technology for successful mass deployment of ASR into the telephone network (AT&T Conversant Systems, 1991).
- *Rejection.* An ability that we take for granted in conversation is the ability to detect when we do not understand what someone is saying. Unfortunately, this is a very difficult task for current speech recognition systems. While it is possible to determine when there are two (or more) possible words or sentences, it has been very difficult for systems to determine when people are saying something on a completely different subject. Given the diversity of customers in the telephone network that would use speech recognition capabilities, accurately rejecting irrelevant input is mandatory. Further research effort is needed in detecting this type of "none of the above" response (Rohlicek et al., 1989; Rose and Hofstetter, 1992; Sukkar and Wilpon, 1993; Wilpon et al., 1990).
- *Subword units.* It is now possible to build a speaker-independent dictionary of models comprised of constituent phonetic (or phoneme-like) statistical models. Initially, this work focused on supporting robust speech recognition for small, easily distinguishable vocabularies. More recently the effort has focused on supporting larger-vocabulary applications (Lennig, 1992). These subword pieces are then concatenated to build representative models for arbitrary words or phrases. Therefore, the effort and expense of gathering speech from many speakers for each new vocabulary word are eliminated, making the development and deployment of new and improved applications simple, quick, and efficient. Much of this work has relied on the knowledge gained from work in TTS. For example, the rules for describing new words in terms of subword units can be derived from

- the rules used by TTS systems to allow for proper pronunciation of words or phrases (Lennig et al., 1992; Rabiner and Juang, 1993).
- *Adaptation.* People can adapt quickly to dialects and accents in speech. It is rather naive to think that we can develop a set of models for a speech recognition system that can recognize all variations in speaking and channel conditions. Machines now have the beginnings of the capability to respond more accurately as they learn an individual voice, dialect, accent, or channel environment (Lee et al., 1991; Rosenberg and Soong, 1987; Schwartz et al., 1987).
  - *Noise immunity and channel equalization.* Better speech enhancement algorithms and channel modeling have made speech recognizers more accurate in noisy or changing environments, such as airports or automobiles (Acero, 1990; Hermansky et al., 1991; Hirsch et al., 1991; Murveit et al., 1992).

For text-to-speech synthesis, some advances in the engineering art include:

- *Proper name pronunciation.* In general, proper names do not follow the same prescribed rules for pronunciation as do other words. However, one of the major applications for TTS technology is to say people's names (e.g., directory assistance applications). Most current TTS systems have implemented techniques to determine the etymology of a name first and then pronounce the name given a set of rules based specifically on its origin. Therefore, *Weissman* would be pronounced with a long *i* (as is common in Germany) as opposed to a long *e* as would be common in English (e.g., as in *receive*) (Church, 1986).
- *Address, date, and number processing.* Addresses, dates, and numbers have different meanings and pronunciations depending on how they are used in an address or sentence. For example, does the abbreviation *St.* stand for *Street* or *Saint*? Is *Dr.* for *Drive* or *Doctor*? And what happens if no punctuation is provided with the text, in which case *oh* could mean *Ohio*. In the past decade, much work has gone into making TTS systems much more robust to these types of requirements. For specific applications, most current systems have no problems with this type of input. There is ongoing research in text analysis to improve the performance of TTS in the most general cases (Sproat et al., 1992).
- *Prosody.* While a natural-sounding voice is an obvious goal of TTS research, current technology still produces "machine"-sounding voices. However, in the past few years the incorporation of better prosodic modeling, such as pitch, duration, and rhythm, has greatly

increased the melodic flow or intonation of the TTS voice (Hirschberg, 1990; van Santen, in press).

The design of an easy-to-use dialogue with a computer system is a significant challenge. We know from experience that it is possible to design good human interfaces for computer dialogue systems. Unfortunately, it has also been verified that it is possible to design systems that aggravate people. At this time there are some general guidelines for good human interface designs, but there is no "cookbook" recipe that guarantees a pleasant and easy-to-use system (Kamm, in this volume). Thus, the art of speech recognition and TTS technologies need to be advanced while waiting for major research breakthroughs to occur.

### **APPLICATIONS OF SPEECH RECOGNITION AND SYNTHESIS**

It is important to bear in mind that the speech technologies described above, notwithstanding advances in reliability, remain error-prone. For this reason the first successful products and services will be those that have the following characteristics:

- *Simplicity.* Successful services will be natural to use. For instance, they may use speech recognition to provide menu capabilities using only small vocabularies (less than 10 words), rather than large vocabularies (more than 1000 words).
- *Evolutionary growth.* The first applications will be extensions of existing systems—for example, speech recognition as a Touch-Tone replacement for voice response systems or TTS for reading out information stored in a machine, such as for remote electronic mail access.
- *Tolerance of errors.* Given that any speech recognizer and synthesizer will make occasional errors, inconvenience to the user should be minimized. This means that careful design of human factors will be essential in providing suitable systems.

That said, there are some general questions that must be asked when considering an application using current speech technologies. The answers will help determine whether it is advisable or possible to design a quality application using speech technology. Some of these questions include:

- Are the potential users friendly and motivated? If so, they might accept a higher error rate in order to carry out the function they desire.

- What environment will the recognizer be expected to work in (e.g., a noisy airport or quiet home)?
- How robust is the algorithm performance in the face of adverse conditions?
- Has the speech technology been prototyped or is it still in the laboratory?
- Can the technology be "broken" by malicious users or hackers?
- How well thought out is the user interface to the technology?
- What accuracy will the user of this service expect?
- What is the maximum tolerable error rate?
- Are the ASR and TTS algorithms accurate enough to meet user expectations?
- Is natural-sounding speech required for the application?
- Does the benefit of using speech technology in this application outweigh its cost compared to alternative technologies?

## SPEECH TECHNOLOGY TELECOMMUNICATIONS MARKET

How do the speech technologies described above expand to telecommunications-based products and services? [Figure 4](#) graphically shows the main application areas for speech recognition, speaker recognition, natural language processing, and text-to-speech synthesis currently considered industry-wide. The figure shows that most of the broad application areas center around speech recognition, such as for menu-based transactions or for information access. The fuzzy lines indicate where overlapping technologies are needed. Applications in this area include the whole field of language translation and identification, where the interaction between natural language processing and speech recognition is essential. In the following sections, I will discuss many of the applications currently being deployed, trialed, or planned that fall into these different technology areas. [Table 1](#) gives a summary of all the applications discussed below.

### Cost Reduction vs. New Revenue Opportunities

There are two classes of applications that are beginning to appear. The first, *cost reduction applications*, are those for which a person is currently trying to accomplish a task by talking with a human attendant. In such applications the accuracy and efficiency of the computer system that replaces the attendant are of paramount concern. This is because the benefits of ASR technology generally reside with the corporation that is reducing its costs and not necessarily

with the end users. Hence, users may not be sympathetic to technology failures. Examples of such applications include (1) automation of operator services, currently being deployed by many telephone companies, including AT&T, Northern Telecom, Ameritech, and Bell Atlantic; (2) automation of directory assistance, currently being trialed by NYNEX and Northern Telecom; and (3) control of network fraud currently being developed by Texas Instruments (TI), Sprint, and AT&T.

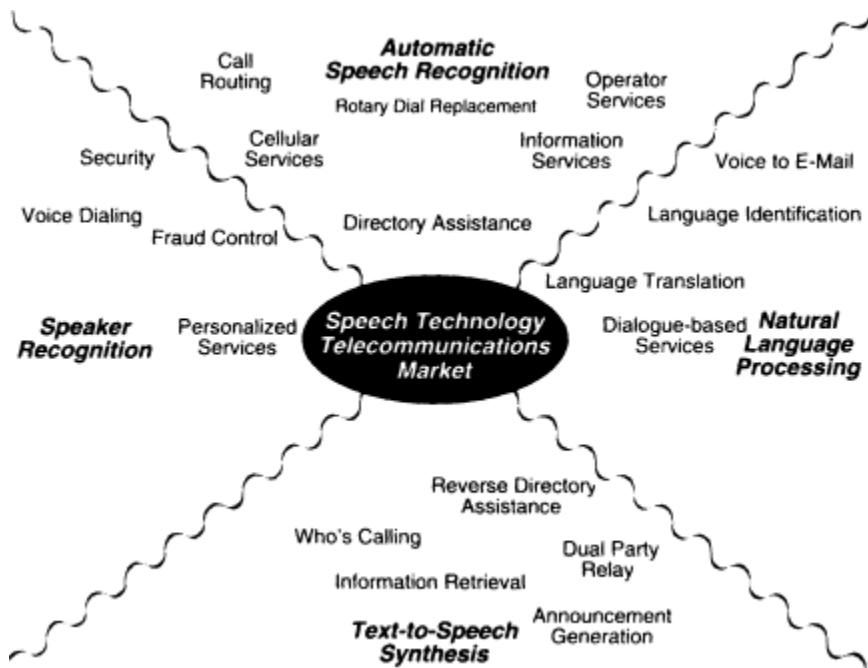


FIGURE 4 Plot showing the major application groups for speech-processing technologies in the telecommunications market.

The second class of applications are services that generate new revenues. For these applications the benefit of speech recognition technology generally resides with the end user. Hence, users may be more tolerant of the technology's limitations. This results in a win-win situation. The users are happy, and the service providers are happy. Examples include (1) automation of banking services using ASR offered since 1981 by NTT in Japan; (2) Touch-Tone and rotary phone replacement with ASR introduced by AT&T in 1992; (3) reverse directory assistance, in which a user enters a telephone number to retrieve a name and address provided by Ameritech and Bellcore.

TABLE 1 Network-Based, Voice-Processing-Based Services

Name	Company	Date	Technology	Task
ANSER	NTT	1981	Small-vocabulary, isolated-word ASR	Banking services
Automated Alternative Billing Services	BNR/Ameritech	1989	Small-vocabulary, isolated-word ASR	Automation of operator services
Intelligent Network	AT&T	1991	Small-vocabulary, wordspotting, barge-in, Spanish	Rotary telephone replacement in Spain
Voice Recognition Call Processing (VRCP)	AT&T	1991	Small-vocabulary, wordspotting, barge-in	Automation of operator services
Telephone Relay Services (TRS)	AT&T	1992	TTS	Enhancement of services to the hearing impaired
Directory Assistance Call Completion (DACC)	NYNEX, AT&T	1992-1993	Small-vocabulary ASR	Automation of directory services
Flex-Word	BNR	1993	Large-vocabulary isolated-word ASR	Stock quotations and automation of directory assistance
Reverse Directory Assistance (RDA)	Bellcore/Ameritech	1993	TTS	Name and address retrieval
Voice Dialing	NYNEX	1993	Small-vocabulary, speaker-dependent, isolated-word, ASR	Automatic name dialing
Voice Prompter	AT&T	1993	Small-vocabulary, wordspotting, barge-in	Rotary telephone replacement
Voice Interactive Phone (VIP)	AT&T	1992	Small-vocabulary, wordspotting, barge-in	Automated access to enhancement telephone features

beginning in 1992; and (4) information access services, such as a stock quotation service currently being trialed by BNR.

In general, the most desirable applications are those that are not gimmicks but provide real usefulness to customers. Since these technologies are in their infancy and have obvious limitations, careful planning and deployment of services must be achieved if mass deployment of the technologies is to occur.

### Automation of Operator Services

In 1985 AT&T began investigating the possibility of using limited-vocabulary, speaker-independent, speech recognition capabilities to automate a portion of calls currently handled by operators. The introduction of such a service would reduce operator workload while greatly increasing the overall efficiency of operator-handled calls. The exact task studied was automation of billing functions: *collect*, *calling card*, *person-to-person*, and *bill-to-third-number*. Customers would be asked to identify verbally the type of call they wished to make without speaking directly to a human operator. Could a simple five-word vocabulary (the function names and the word *operator* for human assistance) be designed, built, and deployed with such a degree of accuracy that customers would be willing to use the technology? Early trials in 1986 and 1987 seemed to indicate that the technology was indeed providing such performance levels (Wilpon et al., 1988).

In 1989 BNR began deploying AABS (Automated Alternate Billing Services) (Lennig, 1990) through local telephone companies in the United States, with Ameritech being the first. [Figure 5](#) shows a diagram of the voice service node platform that BNR used for AABS. For this service, ASR and Touch-Tone technologies are used to automate only the back end of collect and bill-to-third-number calls. That is, after the customer places a call, a speech recognition device is used to recognize the called party's response to the question: *You have a collect call. Please say yes to accept the charges or no to refuse the charges.* Using the two-word recognition system (with several yes/no synonyms), the system successfully automated about 95 percent of the calls that were candidates for automation by speech recognition (Bossemeyer and Schwab, 1991).

After extensive field trials in Dallas, Seattle, and Jacksonville during 1991 and 1992, AT&T announced that it would begin deploying VRCP (Voice Recognition Call Processing). This service would automate the front end as well as the back end of collect, calling card, person-to-person and bill-to-third-number calls. A typical call flow through VRCP is:

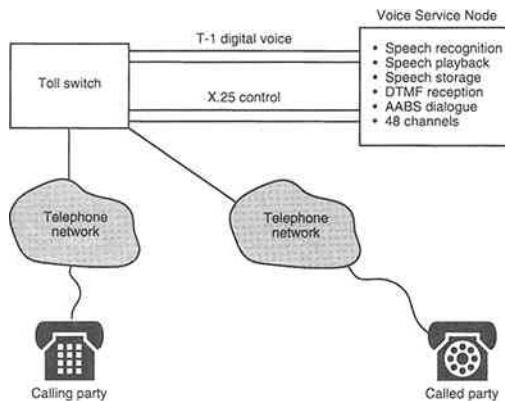


FIGURE 5 Diagram of the voice service node platform used by BNR to automate portions of operator-assisted calls.

System	Caller	Called Party
"Please say collect, calling, card . . ."	"I'd like to make a collect call please."	—
"Thank you for placing a collect call."	—	—
"At the tone please say your name < tone >."	John Smith	—
"Thank you, please wait."	—	—
"You have a collect call from < John Smith >. Please say yes if you accept the charges, no if you refuse the charges, or operator if you need assistance, now."	—	"Yes, I will."
"Thank you, for using AT&T."	—	—

These trials were considered successful not just from a technological point of view but also because customers were willing to use the

service (Franco, 1993). By the end of 1993, it is estimated that over 1 billion telephone calls each year will be automated by the VRCP service.

What differentiates the earlier BNR system from the AT&T system is the speech recognition technology. Analysis of the 1985 AT&T trials indicated that about 20 percent of user utterances contained not only the required command word but also extraneous sounds that ranged from background noise to groups of nonvocabulary words (e.g., "I want to make a *collect* call please"). These extraneous sounds violated a basic assumption for many speech recognition systems of that time: that the speech to be recognized consist solely of words from a predefined vocabulary. With these constraints, the burden of speaking correctly fell on users. In 1990 AT&T developed its wordspotting technology, which began to shift the burden from the users to the speech recognition algorithms themselves. This technology is capable of recognizing key words from a vocabulary list spoken in an unconstrained fashion (Wilpon et al., 1990). Results from recent field trials showed that about 95 percent of the calls that were candidates for automation with speech recognition were successfully automated when wordspotting was used to accommodate all types of user responses.

We expect that the capability to spot key words in speech will be a prerequisite for most telephone network applications. Also, the ability to recognize speech spoken over voice prompts, called *barge in*, is essential for mass deployment of ASR technology in the network (AT&T Conversant Systems, 1991). Both of these techniques have been deployed in VRCP.

Recently, Northern Telecom announced (Lennig, 1992) a trial for the automation of a second operator function, directory assistance. This service would rely on technology that the company calls *Flexible Vocabulary Recognition*. By entering the pronunciation of words in a more basic phonetic-like form, pattern-matching methods can be used to find sequences of subword units that match sequences in the pronunciation "dictionary." Thus, vocabularies of hundreds or thousands of words can, in principle, be recognized without having to record each word. This is especially convenient for vocabularies for which new words need to be added when the service is already in use, for instance, names in a telephone directory.

The directory assistance service would allow telephone customers to obtain telephone numbers via speech recognition, and only in difficult cases would a human operator be necessary. As currently envisioned, the customer would place a call to directory assistance, and hear a digitized voice asking whether the caller preferred French or English (the initial service is planned for Canada). After the caller

says "English" or "Français," subsequent speech prompts would be given in that language. Next the caller would be asked to say the name of the city. The customer's response, one of hundreds of city names in Canada, would be recognized using speaker-independent word recognition based on subword units. The caller would then be transferred to an operator, who would have the information spoken thus far displayed on a screen. Subsequent stages of the call—for instance, recognition of the names of major businesses—would be automated in later phases of the trial deployment.

### Voice Access to Information over the Telephone Network

It has been over a decade since the first widespread use of automatic speech recognition in the telephone network was deployed. In 1981 NTT combined speech recognition and synthesis technologies in a telephone information system called *Anser*—Automatic Answer Network System for Electrical Requests (Nakatsu, 1990). This system provides telephone-based information services for the Japanese banking industry. Anser is deployed in more than 70 cities across Japan serving over 600 banks. Currently, over 360 million calls a year are automatically processed through Anser, bringing in about \$30 million in revenue to NTT annually.

Using a 16-word lexicon consisting of the 10 Japanese digits and six control words, a speaker-independent, isolated-word, speech recognition system allows customers to make inquiries and obtain information through a well-structured dialogue with a computer over a standard telephone.

A typical transaction is of form:

Customer	System
(Calls Center)	"Hello, this is the NTT bank telephone service center. What is your service number?"
"one, one"	"You are asking for your account balance. What is your branch number?"
"one, two, ..."	"What is your account number?"
"three, four, ..."	"What is your secret number?"
"five, six, ..."	"Your current balance is 153,000 yen. If you would like to have your balance repeated, please say 'Once More'. If not, say 'OK'."
"OK"	"Thank you very much"

[Figure 6](#) shows a block diagram of the Anser system. Anser has

embedded a facsimile response unit and a personal computer control unit, so customers have alternate choices to interface with the system. Currently, about 25 percent of customers choose to use the ASR capabilities, with a reported recognition accuracy of 96.5 percent (Furui, 1992).

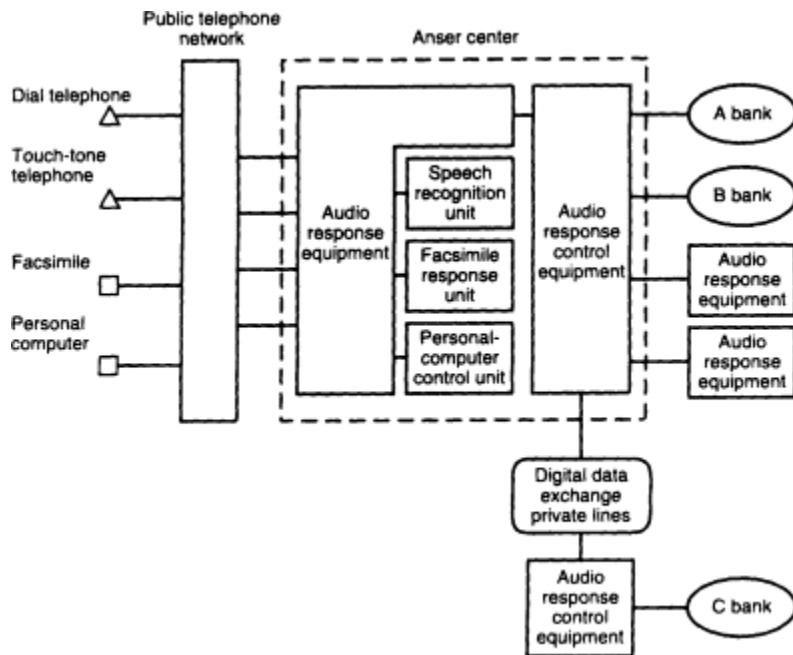


FIGURE 6 Block diagram of NTT's Anser system.

Anser provides a win-win scenario for both the service provider and the customer. From a customer's standpoint, the cost of obtaining information about bank accounts is low (about the cost of a local telephone call). Also, because most banks are on the Anser network, there is a uniformity across the banking industry. Therefore, customers can access any bank computer using the same procedures anytime of the day. For the banking industry, Anser allows the banks to provide a much-needed service to its customers without having to hire large numbers of people or invest heavily in equipment.

Anser also became a successful service because the banks demanded that it be accessible from any ordinary telephone. In 1981 more than 80 percent of the telephones in Japan were rotary dial. Even as late as 1990 more than 70 percent were still rotary. Therefore, speech recognition was an essential technology if Anser was to be successful in the marketplace.

An emerging area of telephone-based speech-processing applications is that of Intelligent Networks services. AT&T currently offers network-based services via a combination of distributed intelligence and out-of-band common channel signaling. For example, a current 800 service might begin with the prompt *Dial 1 for sales information or 2 for customer service*. Until now, caller input to Intelligent Network services required the use of Touch-Tone phones. Automatic speech recognition technology has been introduced to modernize the user interface, especially when rotary phones are used. Using the AT&T 800 Speech Recognition service, announced at the beginning of 1993, menu-driven 800 services can now respond to spoken digits in place of Touch-Tones and will provide automatic call routing.

Intelligent Network services have attracted the interest of international customers. The first AT&T deployment with speech recognition was in Spain in 1991 (Jacobs et al., 1992). The low Touch-Tone penetration rate in Spain (less than 5 percent) and the high-tech profile of Telefonica, the Spanish telephone company, were the primary motivating factors. In a sense, Telefonica is trying to leap-frog past Touch-Tone technology with more advanced technologies. The targets of this application were conservatively set to accommodate the Spanish telephone network and its unfamiliar users. The speech recognizer deployed supports speaker-independent isolated word recognition with wordspotting and barge in of the Spanish key words *uno*, *dos*, and *tres*. Figure 7 illustrates an information service based on this

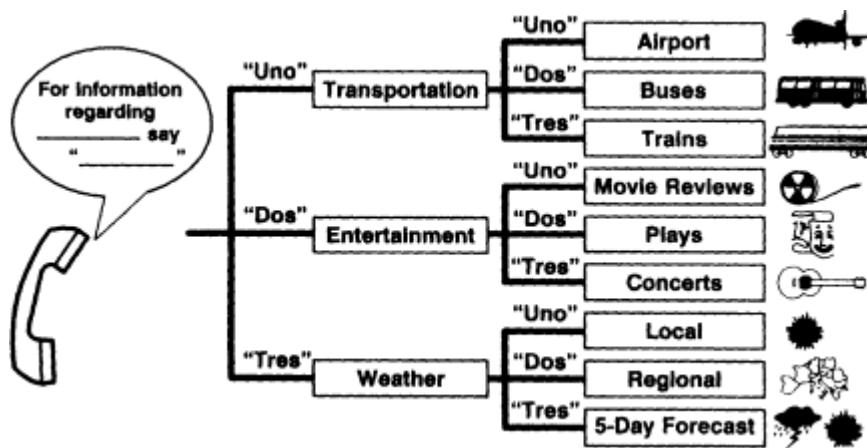


FIGURE 7 Example of an ASR-based service using AT&T's Intelligent Network.

platform. With this system a user can obtain information on any of nine topics with just two voice commands.

Particular attention was paid to recognition of key words embedded in unconstrained speech. For isolated words the recognizer correctly identifies the word 98 percent of the time. Considering all customer responses, including words embedded in extraneous speech, an overall recognition accuracy of 95 percent or better is expected. Similar systems were also tested in England, Scotland, Wales, and Ireland in 1991 (Jacobs et al., 1992).

Subword-based speech recognition has recently been applied to an information access system by BNR working with Northern Telecom (Lennig, 1992). Using the 2000 company names on the New York Stock Exchange, callers to this system can obtain the current price of a stock simply by speaking the name of the stock. Though the accuracy is not perfect, it is adequate considering the possibilities of confusion on such a large-vocabulary. The experimental system is freely accessible on an 800 number, and it has been heavily used since its inception in mid-1992. There are thousands of calls to the system each day, and evidence suggests that almost all callers are able to obtain the stock information they desire.

The general concept is voice access to information over the telephone. People obtain information from computer databases by asking for what they want using their telephone, not by talking with another person or typing commands at a computer keyboard. As this technology develops, the voice response industry will expand to include voice access to information services such as weather, traffic reports, news, and sports scores. The voice response services will begin to compete with "news radio" and computer information networks. Soon people who get information from computer databases by telephone will not think of themselves as sophisticated computer users. Ease of use is the key; people will think of themselves as having a brief conversation with a machine to get the information they need.

### Voice Dialing

One of the biggest new revenue-generating applications of speech recognition technology in telecommunications is voice dialing, or a so-called *Voice Roledex*. Currently, we must all remember hundreds of phone numbers of people and businesses that we need to call. If we want to call someone new, we either look the number up in a phone book or call directory assistance to get the number. But the phone number is only a means to an end—the end being that we

want to place a phone call. The fact of the matter is that people really do not want to keep track of phone numbers at all. We should be able to just speak the name of the party that we want to reach and have the phone call be placed totally automatically. This is one example of the *people have easier access to one another* part of our vision.

Obviously, current ASR technology is not advanced enough to handle such requests as, *Please get me the pizza place on the corner of 1st and Main, I think it's called Mom's or Tom's*. However, most requests to place telephone calls are much simpler than that—for example, *Call Diane Smith, Call home, or I'd like to call John Doe*. ASR technology, enhanced with wordspotting, can easily provide the necessary capabilities to automate much of the current dialing that occurs today.

Voice-controlled repertory dialing telephones have been on the market for several years and have achieved some level of market success. Recently, NYNEX announced the first network-based voice dialing service, whereby the user picks up his or her home phone and says the name he or she would like to call. It is a speaker-trained system supporting about 50 different names and does not have wordspotting capabilities. NYNEX believes this service will be a market winner, expecting over 10 percent of its customers to enroll.

One of the main drawbacks of the NYNEX system is that it is tied to a person's home or office telephone. Customers cannot use the service while they are away from their base phone. Currently, AT&T, Sprint, MCI, and TI are trialing expanded versions of the voice-dialing application that handles the large *away from home and office* market. These systems allow users to place phone calls using speech recognition from any phone, anywhere. In addition, these systems also plan to use speaker verification technology to provide a level of network security for the user. Users would have to enter a voice password or account number that would be used to verify their identity before allowing them to place a call. The AT&T system will also use speaker-independent, subword-based ASR instead of speaker-trained ASR for the name-dialing function of the service. This will provide an enormous reduction in data storage and will allow TTS technology to be used to read back the name that is being called. These systems will be available during 1994.

### **Voice-Interactive Phone Service**

Another example of a revenue-generating service is the Voice-Interactive Phone (VIP) service introduced by AT&T and US West in 1992. This service allows customers to access a wide range of telecommunications services by voice, with the goal of eliminating the

need for a customer to learn the different access codes for existing or new features. In addition, the service provides voice confirmation that the service requested is being turned on or off.

The procedure for using VIP is for the customer to dial an abbreviated access code (e.g., three digits) and then hear a prompt of the form:

*"Please say the name of the feature you want, or say 'HELP' for a list of the services you subscribe to, now."*

The user then speaks the name of the service and receives confirmation of connection to that service. The services available through VIP and the associated voice commands are as follows:

Service	Voice Command
Call Forwarding	Call Forwarding
Continuous Redial	Redial
Last Call Return	Return Call
Call Rejection	Call Rejection
Caller ID Blocking	Block ID
Access to Messaging Services	Messages
Temporary Deactivation of Call Waiting	Cancel Call Waiting

Based on a series of customer trials, the following results were obtained:

- Eighty-four percent of the users preferred VIP over their present method.
- Ninety-six percent of the users were comfortable with the idea of speaking to a machine.
- Most users felt that the primary benefit of VIP was not having to remember multiple codes or procedures.
- Seventy-five percent of users tried different services with VIP more often or were willing to try services they had never tried before.

### Directory Assistance Call Completion

Another application of speech recognition toward directory assistance is Directory Assistance Call Completion (DACC). The concept behind this service is that when someone calls directory assistance to obtain a telephone number, he or she will usually immediately dial that number. As mentioned above, people do not really want phone numbers, they want to be connected to someone. NYNEX (in 1992)

and AT&T and Bell Mobility (BM) (in 1993) have trialed the DACC service, which allows the number to be dialed automatically. The customer is asked a question such as, *Would like us to place the call for you? Please say yes or no.* If the customer answers *yes*, the call is placed. Ideally, the customer need not have to remember the actual telephone number.

The systems developed by AT&T and BM require an additional use of speech recognition. Directory assistance is provided by the local operating companies, the only ones that have automatic access to the specific telephone numbers in the directory. Since neither AT&T nor BM has electronic access to these numbers, connected digit recognition technology is used to recognize the voice response unit message containing the phone number, which is played back to the user. The system then dials the number.

### **Reverse Directory Assistance**

Ameritech recently announced a service called Automated Customer Name and Address (ACNA). In this service, customers are provided with name and address information associated with particular telephone numbers. After the user provides a telephone number using Touch-Tone input (currently no speech recognition technology is being used), a search is made in a reverse directory database, and text-to-speech synthesis is used to return the desired information to the user. NYNEX trialed a similar service in 1992 (Yashchin et al., 1992). For these types of voice-based information access services, where the number of responses that the system must provide the user is extremely large, it is not feasible to record each message, store it, and provide a mechanism to enter new information and change existing information. Therefore, TTS capabilities are an essential part of the service requirements.

### **Telephone Relay Service**

For AT&T's Telephone Relay Service (TRS), text-to-speech synthesis is used to help hearing-impaired individuals carry on conversations with normal-hearing individuals over telephone lines by minimizing the need for third-party intervention or eliminating the need for both parties to have TDD (Terminal Device for the Deaf) terminals. [Figure 8](#) shows a block diagram of how this service works. It is assumed that one party is hearing impaired and has a TDD terminal and the other party has no hearing impairment and no special terminal device.

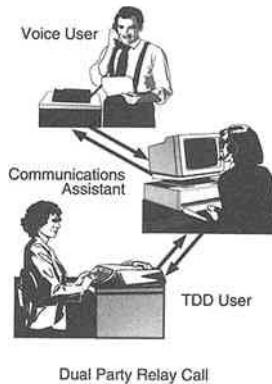


FIGURE 8 Pictorial representation of AT&T's Telephone Relay Service.

After dialing into the TRS service, an operator is assigned to the call. As the hearing party speaks, the operator transcribes the speech on a terminal. (Obviously, one would like to have a speech recognizer listening to the incoming speech. However, as stated earlier, ASR technology currently cannot support recognition of fluent spontaneous speech spoken by anyone on any topic.) The text is then transmitted to the hearing-impaired party's TDD unit. When the hearing-impaired party enters a response on his or her TDD, that text is transferred to a TTS system, which then plays out his or her response to the hearing party. This would allow anyone to communicate with a hearing-impaired person without a TDD device. It should be noted that this service has existed without TTS for several years.

The TRS service with TTS technology was trialed by AT&T in California in 1990. Fifteen operator positions were equipped with TTS equipment, and 15 positions were staffed by live operators (as a control) who would read the text to the hearing party. Over 35,000 calls were processed by TTS. Greater than 80 percent of the calls successfully used TTS for the entire duration of the call. Eighty-eight percent of TDD customers and 86 percent of hearing customers rated the TTS service as good or excellent. The worst problem was incorrect handling of spelling errors. Currently, this service is being deployed throughout Washington state.

There were many technical challenges that stood in the way of automating this service using TTS technology (Bachenko et al., 1992; J. Tschirgi, personal communication, 1993). For example, people using TDDs:

- Use only single-case type, usually uppercase. Since TTS systems generally make use of upper- and lowercase information in determining pronunciation rules, modifications had to be made in order to handle this type of text input.
- Do not use punctuation. Therefore, there are no sentence boundaries. Additionally, there is no way to disambiguate whether a sequence of letters is an abbreviation or an actual word.
- Use special abbreviations and contractions, for example *XOXOXO* for *love and kisses*; *OIC* for *Oh, I see*; *PLS* for *please*; and *Q* to indicate a question.
- Use regional abbreviations depending on where they live, for example, *LAX* for Los Angeles Airport.
- Misspell about 5 percent of words they type. Obviously, this will cause problems for any TTS system. All of these issues required extensive research and development before a successful service was deployed.

## FUTURE POSSIBILITIES

It has been observed that predictions of future technologies tend to be overly optimistic for the short term and overly pessimistic for the long haul. Such forecasts can have the unfortunate effect of creating unrealistic expectations leading to useless products, followed by premature abandonment of the effort. I have tried to counteract this tendency by carefully pointing out the limitations of current speech recognition and text-to-speech technologies while focusing on the types of applications that can be successfully deployed for mass user consumption given today's state of the art.

### Near-Term Technical Challenges

While the prospect of having a machine that humans can converse with as fluently as they do with other humans remains the Holy Grail of speech technologists and one that we may not see realized for another generation or two, there are many critical technical problems that I believe we will see overcome in the next 2 to 5 years. Solving these challenges will lead to the ubiquitous use of speech recognition and synthesis technologies within the telecommunications

industry. The only question is how these advances will be achieved. For speech recognition, these research challenges include:

- *Better handling of the varied channel and microphone conditions.* The telephone network is constantly changing, most recently moving from analog to digital circuitry and from the old-style nonlinear carbon button-type transducers to the newer linear electret type. Each of these changes affects the spectral characteristics of the speech signal. Current ASR and especially speaker verification algorithms have been shown to be not very robust to such variability. A representation of the speech signal that is invariant to network variations needs to be pursued.
- *Better noise immunity.* While advances have been made over the past few years, we are a long way away from recognition systems that work equally well from a quiet home or office to the noisy environments encountered at an airport or in a moving car.
- *Better decision criteria.* For a long time, researchers have mainly considered the speech recognition task as a two-class problem, either the recognizer is right or it is wrong, when in reality it is a three-class problem. The third possibility is that of a *nondecision*. Over the past few years, researchers have begun to study the fundamental principles that underlie most of today's algorithms with an eye toward developing the necessary metrics that will feed the creation of robust rejection criteria.
- *Better out-of-vocabulary rejection.* While current wordspotting techniques do an excellent job of rejecting much of the out-of-vocabulary signals that are seen in today's applications, they are by no means perfect. Since AT&T announced that its wordspotting technology was available for small-vocabulary applications in its products and services beginning in 1991, many other ASR vendors have realized that the ability to distinguish key word from nonkey word signals is mandatory if mass deployment and acceptance of ASR technology are to occur. Hence, more and more ASR products today are being offered with wordspotting capabilities. Additionally, as our basic research into more advanced, large-vocabulary systems progresses, better out-of-vocabulary rejection will continue to be a focusing point. With all this activity being directed to the issue, I am sure we will see a steady stream of new ideas aimed at solving this problem.
- *Better understanding and incorporation of task syntax and semantics and human interface design into speech recognition systems.* This will be essential if we are to overcome the short-term deficiencies in the basic technologies. As ASR-based applications continue to be deployed,

the industry is beginning to understand the power that task-specific constraints have on providing useful technology to its customers.

Challenges for text-to-speech synthesis research include:

- *More human-sounding speech.* While totally natural speech is decades away, improvements in prosody and speech production methodology will continue to improve the quality of the voices we hear today. One point to note: there are only a handful of laboratories currently working on TTS research; therefore, advances in TTS technology will most likely occur at a slower pace than those made in speech recognition.
- *Easy generation of new voices, dialects, and languages.* Currently, it takes many months to create new voices or to modify existing ones. As more personal telecommunications services are offered to customers, the ability to customize voices will become very important. A simple example of this might be the reading of e-mail. If I know that the e-mail was sent by a man or woman (or a child), the synthesizer should be able to read the text accordingly.

### Personal Communication Networks and Services

One of the most exciting new uses of speech technologies is in the area of Personal Communication Networks (PCNs) and Personal Communication Services (PCSs). It is quite obvious that as Personal Communication Devices (PCDs) come of age, their very nature will require the use of advanced speech technologies. As these devices become smaller and smaller, there will be no room for conventional Touch-Tone keypads or any other type of mechanical input device. What room will be available will undoubtedly be used for a video display of some kind. Moreover, the display will more than likely be too small for touch screen technologies other than those that use pen-based input. Thus, speech technologies will become necessary if we are to easily communicate with our personal communicators.

Within the next 2 to 3 years I expect to see some rudimentary speech recognition technology incorporated into PCDs. Initially, ASR will provide users with simple menus for maneuvering around the PCD, including the ability to place telephone calls across the wireless network. Voice response technology will also be included to provide audio feedback to users. This will most probably be done by using current voice coding techniques and then migrating to TTS as the technology becomes implementable on single chips and the large memory requirements of current TTS techniques can be reduced.

## Predictions

Predicting a generation in the future may be a futile exercise. It is impossible to predict when a technical revolution will occur; few people could have predicted in the 1960s the impact that VLSI would have on our society. There is also the risk of being blinded by the past when looking to the future. All we can say with assurance is that the present course of our technology will take us somewhat further; there are still engineering improvements that can be built on today's science. We can also anticipate, but cannot promise, advances in scientific knowledge that will create the basis upon which a *new* generation of speech recognizers and synthesizers will be designed.

Let me go out on a limb (a bit) and make some specific predictions:

- Algorithms for highly accurate, speaker-independent recognition of large vocabularies will soon become available. Before the year 2000, this technology will be successfully engineered into specific large-scale applications that are highly structured, even if the vocabulary is large.
- Major advances will be made in language modeling for use in conjunction with speech recognition. In contrast to the past two decades, in which advances were made in feature analysis and pattern comparison, the coming decade will be the period in which computational linguistics makes a definitive contribution to "natural" voice interactions. The first manifestations of these better language models will be in *restricted-domain* applications for which specific semantic information is available, for example, an airline reservation task (Hirschman et al., 1992; Marcus, 1992; Pallet, 1991; Proceedings of the DARPA Speech and Natural Language Workshop, 1993).
- Despite the advances in language modeling, the speech-understanding capability of computers will remain far short of human capabilities until well into the next century. Applications that depend on language understanding for *unrestricted* vocabularies and tasks will remain a formidable challenge and will not be successfully deployed for mass consumption in a telecommunications environment for several decades.
- Speech recognition over telephone lines will continue to be the most important market segment of the voice-processing industry, both in terms of the number of users of this technology and its economic impact. The ability to get information remotely, either over telephone lines or wireless personal communications systems, will drive many applications and technological advances.

- "Simple" applications of speech recognition will become commonplace. By the year 2000, more people will get remote information via voice dialogues than will by typing commands on Touch-Tone keypads to access remote databases. These information access applications will begin as highly structured dialogues and will be specific to narrow domains such as weather information or directory assistance.
- Truly human-quality text-to-speech synthesis technology will not be available for another decade. As is the case for totally unrestricted-vocabulary ASR algorithms, researchers will have to totally rethink the problem in order to achieve our vision.
- Finally, I confidently predict that at least one of the above six predictions will turn out to have been incorrect.

One thing is very clear: sooner than we might expect, applications based on speech recognition and synthesis technologies will touch the lives of every one of us.

## REFERENCES

- Acero, A., Acoustical & Environmental Robustness in Automatic Speech Recognition, Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, Pa., September 1990.
- AT&T Conversant Systems, CVIS product announcement, New York City, January 1991.
- Bachenko, J., J. Daugherty, and E. Fitzpatrick, A parser for real-time speech synthesis of conversational texts, in Proceedings of the ACL Conference on Applied NL Processing, Trente, Italy, April 1992.
- Bossemeyer, R. W., and E. C. Schwab, Automated alternate billing services at Ameritech: Speech recognition performance and the human interface, *Speech Tech. Mag.* 5(3):2430, February/March 1991.
- Church, K., Stress Assignment in Letter to Sound Rules for Speech Synthesis. in Proceedings of the ICASSP '86, Vol. 4, pp. 2423-2426, April 1986.
- Doddington, G. R., Whither speech recognition? in Trends in Speech Recognition, W. Lea, ed., Prentice-Hall, Englewood Cliffs, N.J., 1980.
- Franco, V., Automation of Operator Services at AT&T, in Proceedings of the Voice '93 Conference, San Diego, March 1993.
- Furui, S., Telephone networks in 10 years' time—technologies & services, in Proceedings of the COST-232 Speech Recognition Workshop, Rome, Italy, November 1992.
- Hermansky, H., N. Morgan, A. Buyya, and P. Kohn, Compensation for the effects of communication channel in auditory-like analysis of speech, in Proceedings of Eurospeech '91, pp. 1367-1370, September 1991.
- Hirsch, H., P. Meyer, and H. W. Ruehl, Improved speech recognition using high-pass filtering of subband envelopes, in Proceedings of Eurospeech '91, pp. 413-416, September 1991.
- Hirschberg, J., Using discourse context to guide pitch accent decisions in synthetic speech, in ESCA Workshop on Speech Synthesis, pp. 181-184, Autrans, France, ESCA, September 1990.
- Hirschman, L., et al., Multi-site DATA collection for a spoken language corpus, in

- Proceedings of the DARPA Speech and Natural Language Workshop, pp. 7-14, Harriman, N.Y., February 1992.
- Hutchins, W. J., and H. L. Somers, An Introduction to Machine Translation, Academic Press, N.Y., 1992.
- Jacobs, T. E., R. A. Sukkar, and E. R. Burke, Performance trials of the Spain and United Kingdom Intelligent Network automatic speech recognition systems, in Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, N.J., October 1992.
- Lee, C., C-H. Lin, and B-H. Juang, A study on speaker adaption of the parameters of continuous density hidden Markov models, IEEE Trans, 39(4):806-814, April 1991.
- Lennig, M., Putting speech recognition to work in the telephone network, Computer, 23(8):35-41, August 1990.
- Lennig, M., Automated bilingual directory assistance trial in Bell Canada, in Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, N.J., October 1992.
- Lennig, M., D. Sharp, P. Kenny, V. Gupta, and K. Precoda, Flexible vocabulary recognition of speech, Proc. ICSLP-92, pp. 93-96, Banff, Alberta, Canada, October 1992.
- Marcus, M., ed., Proceedings of the Fifth DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, San Mateo, Calif., 1992.
- Meisel, W., ed., Speech Recognition UPDATE, TMA Associates, Encino, Calif., 1993.
- Morimoto, T., H. Iida, A. Kurematsu, K. Shikano, and T. Aizawa, Spoken language: Towards realizing automatic telephone interpretation, in Proceedings of the Info JAPAN '90: International Conference of the Information Processing Society of Japan, pp. 553-559, 1990.
- Murveit, H., J. Butzberger, and M. Weintraub, Reduced channel dependence for speech recognition, in Proceedings of the DARPA Speech and Natural Language Workshop , pp. 280-284, Harriman, N.Y., February 1992.
- Nakatsu, R., Anser-An application of speech technology to the Japanese banking industry, Computer, 23(8):43-48, August 1990.
- Oberteuffer, J., ed., ASR News, Voice Information Associates Inc., Lexington, Mass., 1993.
- Pallet, D., Summary of DARPA Resource Management and ATIS benchmark test session, and other articles, in Speech and Natural Language, pp. 49-134, February 1991.
- Pierce, J. R., Whither speech recognition? J. Acoust. Soc. Am., 46(4):1029-1051, 1969.
- Proceedings of the DARPA Speech and Natural Language Workshop. Harriman, N.Y., January 1993.
- Rabiner, L. R., and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.
- Roe, D., and J. Wilpon, Whither speech recognition—25 years later, IEEE Trans. on Commun., November 1993, pp. 54-62.
- Roe, D. B., et al., A spoken language translator for restricted-domain context-free languages, Speech Commun., 11:311-319, 1992.
- Rohlicek, J., W. Russell, S. Roucos, and H. Gish, Continuous hidden Markov modeling for speaker-independent word spotting, in Proceedings of the ICASSP '89, pp. 627-630, March 1989.
- Rose, R., and E. Hofstetter, Techniques for task independent word spotting in continuous speech messages, in Proceedings of the ICASSP '92, March 1992.
- Rosenberg, A., and F. Soong, Evaluation of a vector quantization talker recognition system in text independent & text dependent modes, Comput. Speech Lang., 22:143157, 1987.

- Schwartz, R., Y. L. Chow, and F. Kubala, Rapid speaker adaption using a probabilistic spectral mapping, in Proceedings of the ICASSP '87, pp. 633-636, Dallas, April 1987.
- Sproat, R., J. Hirschberg, and D. Yarowsky, A corpus-based synthesizer, in Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, 1992.
- Sukkar, R., and J. Wilpon, A two pass classifier for utterance rejection in keyword spotting, in Proceedings of the ICASSP '93, volume 2, pp. 451-454, Minneapolis, Minn., April 1993.
- van Santen, J. P. H., Assignment of segmental duration in text-to-speech synthesis, *Comput. Speech Lang.*, in press.
- Waibel, A., A. Jain, A. McNair, H. Saito, A. Hauptmann, and J. Tebelskis, JANUS: A speech-to-speech translation system using connectionist & symbolic processing strategies, in Proceedings of the ICASSP '91, pp. 793-796, March 1991.
- Wilpon, J. G., L. R. Rabiner, C. H. Lee, and E. R. Goldman, Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. on Acoust., Speech, Signal Process.*, 38(11):1870-1878, November 1990.
- Wilpon, J. G., D. DeMarco, and P. R. Mikkilineni, Isolated word recognition over the DDD telephone network-Results of two extensive field studies, in Proceedings of the IEEE ICASSP '88, pp. 55-58, N.Y., April 1988.
- The Yankee Group, Voice Processing: The Second Generation of Equipment & Services, The Yankee Group, December 1991.
- Yashchin, D., S. Basson, A. Kalyanswamy, and K. Silverman, Results from automating a name & address service with speech synthesis, in Proceedings of the AVIOS, 1992.

# Speech Processing for Physical and Sensory Disabilities

*Harry Levitt*

## SUMMARY

Assistive technology involving voice communication is used primarily by people who are deaf, hard of hearing, or who have speech and/or language disabilities. It is also used to a lesser extent by people with visual or motor disabilities.

A very wide range of devices has been developed for people with hearing loss. These devices can be categorized not only by the modality of stimulation [i.e., auditory, visual, tactile, or direct electrical stimulation of the auditory nerve (auditory-neural)] but also in terms of the degree of speech processing that is used. At least four such categories can be distinguished: assistive devices (a) that are not designed specifically for speech, (b) that take the average characteristics of speech into account, (c) that process articulatory or phonetic characteristics of speech, and (d) that embody some degree of automatic speech recognition.

Assistive devices for people with speech and/or language disabilities typically involve some form of speech synthesis or symbol generation for severe forms of language disability. Speech synthesis is also used in text-to-speech systems for sightless persons. Other applications of assistive technology involving voice communication include voice control of wheelchairs and other devices for people with mobility disabilities.

## INTRODUCTION

Assistive technology is concerned with "devices or other solutions that assist people with deficits in physical, mental or emotional function" (LaPlante et al., 1992). This technology can be as simple as a walking stick or as sophisticated as a cochlear implant with advanced microelectronics embedded surgically in the ear. Recent advances in computers and biomedical engineering have greatly increased the capabilities of assistive technology for a wide range of disabilities. This paper is concerned with those forms of assistive technology that involve voice communication.

It is important in any discussion of assistive technology to distinguish between impairment, disability, and handicap. According to the *International Classification of Impairments, Disabilities, and Handicaps* (World Health Organization, 1980), an *impairment* is "any loss or abnormality of psychological, physiological or anatomical structure or function"; a *disability* is "a restriction in the ability to perform essential components of everyday living"; and a *handicap* is a "limitation on the fulfillment of a role that is normal for that individual." Whereas handicap may be the result of a disability that, in turn, may be the result of an impairment, these consequences are not necessarily contingent on each other. The fundamental aim of assistive technology is to eliminate or minimize any disability that may result from an impairment and, concomitantly, to eliminate or minimize any handicap resulting from a disability.

Figure 1 shows the extent to which different forms of assistive technology are being used in the United States, as measured by the 1990 Health Interview Survey on Assistive Devices (LaPlante et al., 1992). Of particular interest are those forms of assistive technology that involve voice communication. Assistive devices for hearing loss are the second most widely used form of assistive technology (4.0 million Americans as compared to the 6.4 million Americans using assistive mobility technology). It is interesting to note that in each of these two widely used forms of assistive technology one specific device dominates in terms of its relative use—the cane or walking stick in the case of assistive mobility technology (4.4 million users) and the hearing aid in the case of assistive hearing technology (3.8 million users). It should also be noted that only a small number of people with hearing loss who could benefit from acoustic amplification actually use hearing aids. Estimates of the number of people in the United States who should wear hearing aids range from 12 million to 20 million, or three to five times the number who actually do (Schein and Delk, 1974). A less widely used form of assistive technology

involving both speech and language processing is that of assistive devices for people with speech and/or language disabilities. Devices of this type typically involve some form of speech synthesis or symbol generation for severe forms of language disability. Speech synthesis is also used in text-to-speech systems for sightless persons. A relatively new form of assistive technology is that of voice control of wheelchairs, hospital beds, home appliances, and other such devices by people with mobility disabilities. This form of assistive technology is not widely used at present, but it is likely to grow rapidly once its advantages are realized.

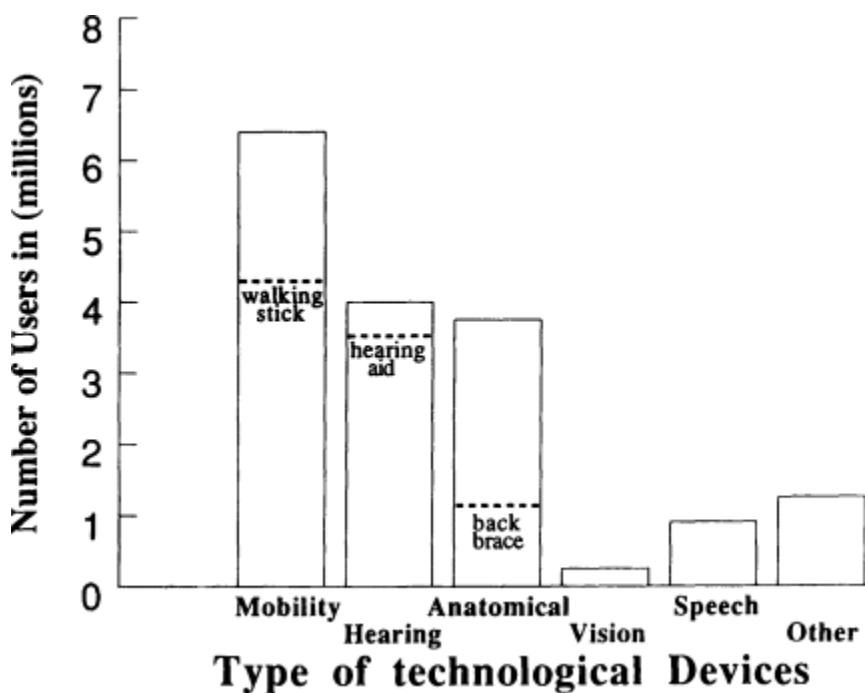


FIGURE 1 Users of assistive technology. The numbers of users (in millions) of various types of assistive technology are shown. The three most widely used types of assistive technology (mobility, hearing, and anatomical) all include a specific device that is used substantially more than any other (walking stick, hearing aid, and backbrace, respectively). The labeled dashed lines within the bars show the numbers of users for each of these specific devices. Note that eyeglasses were not considered in this survey. The diagram is based on data reported in LaPlante et al. (1992).

The demand for a given form of assistive technology has a significant effect on the research and development effort in advancing

that technology. Hearing aids and related devices represent the most widely used form of assistive technology involving voice communication. Not surprisingly, the body of research on assistive hearing technology and on hearing aids, in particular, is large, and much of the present chapter is concerned with this branch of assistive technology.

## ASSISTIVE HEARING TECHNOLOGY

### Background

Assistive technology for hearing loss includes, in addition to the hearing aid, assistive listening devices (e.g., infrared sound transmission systems for theaters), visual and tactile sensory aids, special alarms and text telephones (also known as TTYS, as in TeleTYpewriters, or TDDs, as in Telecommunication Devices for the Deaf). The use of this technology increases substantially with age, as shown in [Table 1](#). Whereas only 3.9 percent of hearing-aid users are 24 years of age or younger, over 40 percent are 75 or older. The use of TDDs/TTYS also increases with age, but the growth pattern is not as systematic (the highest percentage of users, 32.1, being in the 45- to 64-year age group). Note that this table does not specifically identify tactile sensory aids, visual displays of various kinds (e.g., captioning of television programs), or cochlear implants. Whereas the number of users of tactile aids or cochlear implants is measured in the thousands (as compared to the millions and hundreds of thousands of hearing-aid and TTYS/TDD users, respectively), the use of these two less well-known forms of assistive hearing technology is growing steadily. Visual displays

TABLE 1 Use of Hearing Technology Age by Group

Device	Percentage of Users as a Function of Age					
	Total No. of Users (millions)	24 years and under	25-44 years	45-64 years	65-74 years	75 years and over
Hearing Aid	3.78	3.9	6.0	19.7	29.1	41.3
TTY/ TDD	0.17	12.7	13.3	32.1	13.8	27.5
Special Alarms	0.08	9.2	22.3	31.5	6.6	30.2
Other	0.56	4.3	10.0	24.2	25.2	36.4

NOTE: Percentages (summed across columns) may not total 100 because of rounding.

SOURCE: LaPlante et al. (1992).

for deaf people, on the other hand, are very widely used in the form of captioned television.

An important statistic to note is the high proportion of older people who use hearing aids (about two in five). There is, in addition, a large proportion of older people who have a hearing loss and who should but do not use a hearing aid. Not only is this combined proportion very high (over three in five), but the total number of older people with significant hearing loss is growing rapidly as the proportion of older people in the population increases.

A related factor to bear in mind is that people with visual disabilities rely heavily on their hearing. Older people with visual disabilities will have special needs if hearing loss becomes a factor in their later years. This is an important concern given the prevalence of hearing loss in the older population. Similarly, people with hearing loss whose vision is beginning to deteriorate with age have special needs. The development of assistive technology for people with more than one impairment presents a compelling challenge that is beginning to receive greater attention.

Assistive devices for hearing loss have traditionally been categorized by the modality of stimulation—that is, auditory, visual, tactile, or direct electrical stimulation of the auditory nerve (auditory-neural). Another more subtle form of categorization is in terms of the degree of speech processing that is used. At least four such categories can be distinguished: assistive devices (a) that are not designed specifically for speech; (b) that take the average characteristics of speech into account, such as the long-term speech spectrum; (c) that process articulatory or phonetic characteristics of speech; and (d) that embody some degree of automatic speech recognition.

The two methods of categorization, by modality and by degree of speech processing, are independent of each other. The many different assistive devices that have been developed over the years can thus be specified in terms of a two-way matrix, as shown in [Table 2](#). This matrix provides some revealing insights with respect to the development of and the potential for future advances in assistive hearing technology.

### Hearing Aids and Assistive Listening Devices

Column 1 of [Table 2](#) identifies assistive devices using the auditory channel. Hearing aids of various kinds are covered here as well as assistive listening devices (ALDs). These include high-gain telephones and listening systems for rooms and auditoria in which the signal is transmitted by electromagnetic means to a body-worn re

ceiver. Low-power FM radio or infrared transmissions are typically used for this purpose. The primary function of most assistive listening devices is to avoid the environmental noise and reverberation that are typically picked up and amplified by a conventional hearing aid. FM transmission is typically used in classrooms and infrared transmission in theaters, auditoria, houses of worship, and other public places. Another application of this technology is to allow a hard-of-hearing person to listen to the radio or television at a relatively high level without disturbing others.

TABLE 2 Categories of Sensory Aids

		Modality			
		1	2	3	4
Type of Processing		Auditory	Visual	Tactile	Direct Electrical
1	Nonspeech specific	Early hearing aids	Envelope displays	Single-channel vibrator	Single-channel implant
2	Spectrum analysis	Modern hearing aids	Spectrographic displays	Tactile vocoder	Spectrum-based multi-channel implants
3	Feature extraction	Speech-feature hearing aids	Speech-feature displays	Speech-feature displays	Speech-feature implants
4	Speech recognition	Speech recognition-synthesis	Automated relay service and captioning	Speech-to-Braille conversion	—

Most conventional hearing aids and assistive listening devices do not make use of advanced signal-processing technology, and thus a detailed discussion of such devices is beyond the scope of this chapter. There is a substantial body of research literature in this area, and the reader is referred to Braida et al. (1979), Levitt et al. (1980), Skinner (1988), Studebaker et al. (1991), and Studebaker and Hochberg (1993).

It should be noted, however, that major changes are currently taking place in the hearing-aid industry and that these changes are

likely to result in greater use of advanced speech-processing technology in future hearing instruments. Until now the major thrust in hearing-aid development has been toward instruments of smaller and smaller size because most hearing-aid users do not wish to be seen wearing these devices. Miniature hearing aids that fit entirely in the ear canal and are barely visible are extremely popular. Even smaller hearing aids have recently been developed that occupy only the innermost section of the ear canal and are not visible unless one peers directly into the ear canal.

The development of miniature hearing aids that are virtually invisible represents an important turning point in that there is little to be gained cosmetically from further reductions in size. The process of miniaturizing hearing-aid circuits will undoubtedly continue, but future emphasis is likely to be on increasing signal-processing capability within the space available. There is evidence that this change in emphasis has already begun to take place in that several major manufacturers have recently introduced programmable hearing aids—that is, hearing aids that are controlled by a digital controller that, in turn, can be programmed by a computer.

The level of speech processing in modern hearing aids is still fairly elementary in comparison with that used in other forms of human-machine communication, but the initial steps have already been taken toward computerization of hearing aids. The potential for incorporating advanced speech-processing techniques into the coming generation of hearing instruments (digital hearing aids, new types of assistive listening devices) opens up new possibilities for the application of technologies developed for human-machine communication.

Hearing instruments that process specific speech features (row 3, column 1 of [Table 2](#)) are primarily experimental at this stage. The most successful instruments, as determined in laboratory investigations, involve either extraction of the voice fundamental frequency (FO) and/or frequency lowering of the fricative components of speech. It has been shown by Breeuwer and Plomp (1986) that the aural presentation of FO cues as a supplement to lipreading produces a significant improvement in speech intelligibility. It has also been shown by Rosen et al. (1987) that severely hearing-impaired individuals who receive no benefit from conventional hearing aids can improve their lipreading capabilities by using a special-purpose hearing aid (the SiVo) that presents FO cues in a form that is perceptible to the user by means of frequency translation and expansion of variations in FO.

A common characteristic of sensorineural hearing impairment is that the degree of impairment increases with increasing frequency. It has been suggested that speech intelligibility could be improved for

this form of hearing loss by transposing the high-frequency components of speech to the low-frequency region, where the hearing impairment is not as severe (see Levitt et al., 1980). Whereas small improvements in intelligibility (5 to 10 percent) have been obtained for moderate amounts of frequency lowering (10 to 20 percent) for moderate hearing losses (Mazor et al., 1977), the application of greatest interest is that of substantial amounts of frequency lowering for severely hearing-impaired individuals with little or no hearing above about 1000 Hz. A series of investigations on frequency-lowering schemes for hearing impairments of this type showed no significant improvements in speech intelligibility (Ling, 1969). These studies, however, did not take into account the phonetic structure of the speech signal. When frequency lowering is limited to only those speech sounds with substantially more high-frequency energy than low-frequency energy (as is the case with voiceless fricative consonants), significant improvements in speech intelligibility can be obtained (Johansson, 1966; Posen et al., 1993). Frequency lowering of fricative sounds has also proven to be useful in speech training (Guttman et al., 1970). A hearing aid with feature-dependent frequency lowering has recently been introduced for clinical use.

Hearing aids that involve speech recognition processing (row 4, column 1, [Table 2](#)) are still highly experimental. In one such application of speech recognition technology, the output of a speech recognizer was used to drive a special-purpose speech synthesizer in which important phonetic cues in the speech signal are exaggerated (Levitt et al., 1993). It has been shown that the perception of certain speech sounds can be improved significantly for people with severe hearing impairments by exaggerating relevant acoustic phonetic cues. For example, perception of voicing in consonantal sounds can be improved by increasing the duration of the preceding vowel (Revoile et al., 1986). This type of processing is extremely difficult to implement automatically using traditional methods of signal processing. In contrast, automatic implementation of vowel lengthening is a relatively simple matter for a speech recognition/synthesis system.

The major limitations of the experimental speech recognition/synthesis hearing aid evaluated by Levitt et al. (1993) were the accuracy of the speech recognition unit, the machine-like quality of the synthetic speech, the time the system took to recognize the speech, and the physical size of the system. The last-mentioned limitation does not apply if the system is to be used as a desk-mounted assistive listening device.

### Visual Sensory Aids

The development of visual sensory aids for hearing impairment follows much the same sequence as that for auditory aids. The earliest visual sensory aids were concerned primarily with making sounds visible to a deaf person (Bell, 1876). The limitations of these devices for representing speech were soon recognized, and attention then focused on more advanced methods of signal processing that took the average spectral characteristics of speech into account. The sound spectrograph, for example, is designed to make the spectral components of speech visible. (A high-frequency emphasis tailored to the shape of the average speech spectrum is needed because of the limited dynamic range of the visual display.) This device also takes into account the spectral and temporal structure of speech in the choice of bandwidths for the analyzing filters.

The visible speech translator (VST), which, in essence, is a real-time-version sound spectrograph, belongs in the category of sensory aids that take the average characteristics of the speech signal into account (Table 2, row 2). The VST was introduced in the belief that it would provide a means of communication for deaf people as well as being a useful research tool (Potter et al., 1947). Early experimental evaluations of the device supported this view. Figure 2 shows data on the number of words that two groups of hearing subjects (B and C) and one deaf subject (A) learned to recognize using the VST. The deaf subject reached a vocabulary of 800 words after 220 hours of training. The curve for this subject also shows no evidence of flattening out; that is, extrapolation of the curve suggests that this subject

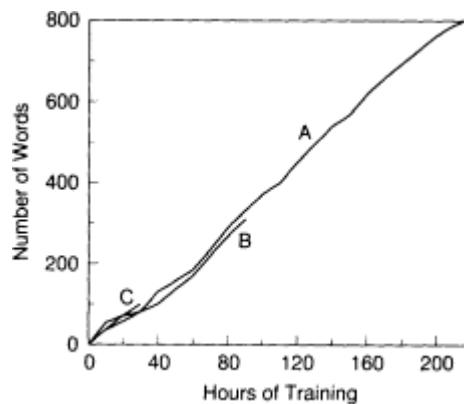


FIGURE 2 Early results obtained with the visible speech translator. The diagram shows the number of words learned by each of five subjects as a function of the amount of training (in hours). Curve A relates to a deaf engineer; curves B and C are related to two groups of two young adults each with no engineering background. Adapted from Potter et al. (1947).

would have continued to increase his vocabulary at the rate of about 3.6 words for each hour of training.

Subsequent evaluations of the VST have not been as impressive. These studies showed that the device was of limited value as an aid to speech recognition (House et al., 1968), although several subsequent studies have shown that speech spectrum displays can be of value in speech training (Stark, 1972). In the speech-training application the user need only focus on one feature of the display at a time, whereas in the case of a visual display for speech recognition several features must be processed rapidly.

Lieberman et al. (1968) have argued that speech is a complex code; that the ear is uniquely suited to interpret this code; and that, as a consequence, perception of speech by modalities other than hearing will be extremely difficult. It has since been demonstrated that it is possible for a human being to read a spectrogram without any additional information (Cole et al., 1980). The process, however, is both difficult and time consuming, and it has yet to be demonstrated that spectrogram reading is a practical means of speech communication.

A much easier task than spectrogram reading is that of interpreting visual displays in which articulatory or phonetic cues are presented in a simplified form. Visual aids of this type are categorized by row 3, column 2 of [Table 2](#). Evidence of the importance of visual articulatory cues is provided by experiments in speechreading (lipreading). Normal-hearing listeners with no previous training in speechreading have been shown to make substantial use of visual cues in face-to-face communication when the acoustic signal is masked by noise (Sumby and Pollack, 1954). Speechreading cues, unfortunately, are ambiguous, and even the most skilled speechreaders require some additional information to disambiguate these cues. A good speechreader is able to combine the limited information received auditorily with the cues obtained visually in order to understand what was said. Many of the auditory cues that are lost as a result of hearing loss are available from speechreading. The two sets of cues are complimentary to a large extent, thereby making speechreading with acoustic amplification a viable means of communication for many hearing-impaired individuals. A technique designed to eliminate the ambiguities in speechreading is that of "cued speech" (Cornett, 1967). In this technique, hand symbols are used to disambiguate the speech cues in speechreading. Nearly perfect reception of conversational speech is possible with highly trained receivers of cued speech (Nicholls and Ling, 1982; Uchanski et al., 1994).

There have been several attempts to develop sensory aids that provide visual supplements to speechreading. These include eye

glasses with tiny lights that are illuminated when specific speech features are produced (Upton, 1968). A more advanced form of these eyeglasses is shown in [Figure 3](#). A mirror image is used to place these visual cues at the location of the speaker, so that the wearer of the sensory aid does not have to change his/her focal length while looking at the speaker and simultaneously attempting to read the supplementary visual cues (Gengel, 1976).

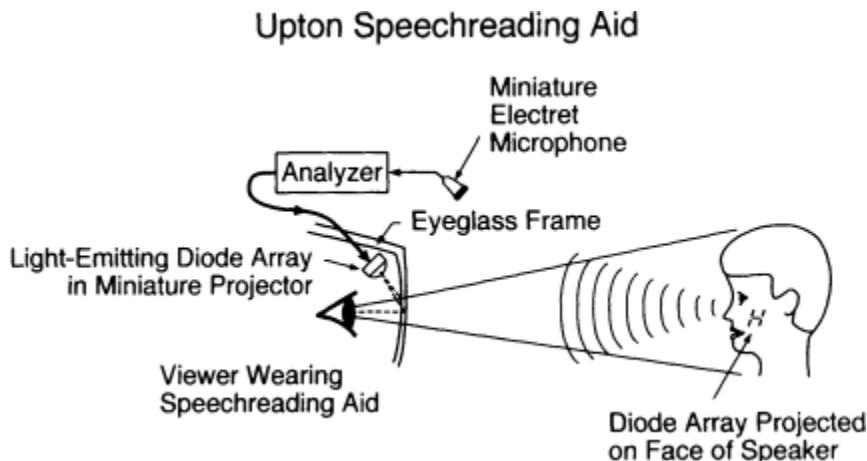


FIGURE 3 Eyeglasses conveying supplementary speech-feature cues. The speech signal is picked up by a microphone mounted on the frame of a pair of eyeglasses. The signal is analyzed, and coded bars of light indicating the occurrence of certain speech features, such as frication or the noise burst in a stop consonant, are projected onto one of the eyeglass lenses. The light beam is reflected into the eye so as to create a virtual image that is superimposed on the face of the speaker. This device, known as the eyeglass speech reader, is a more advanced version of the eyeglass speechreading aid developed by Upton (1968). (Reproduced from Pickett et al. (1974).

Experimental evaluations of visual supplements to speechreading have shown significant improvements in speech recognition provided the supplemental cues are extracted from the speech signal without error (Goldberg, 1972). In practice, the supplemental cues need to be extracted automatically from the speech signal, usually in real-time and often in the presence of background noise and reverberation. Practical systems of this type have yet to show the improvements demonstrated by the error-free systems that have been evaluated in the laboratory. Erroneous information under difficult listening conditions can be particularly damaging to speech understanding. Un

der these conditions, missing cues may be preferable to misleading cues.

In contrast to the limited improvement obtained with visual-feature displays for speech recognition, displays of this type have been found to be of great value in speech training. Visual displays of the voice fundamental frequency are widely used for teaching speech to children with hearing impairments (McGarr et al., 1992; Risberg, 1968). Further, computer-based speech-training systems have been developed that, in addition to providing a range of visual displays of articulatory features, allow for the student to work interactively with the computer, including the use of video games to maintain the student's motivation (Kewley-Port et al., 1987; Ryalls et al., 1991; Watson et al., 1989; Yamada et al., 1988).

Another very useful application of speech-processing technology in this context is that of synthesizing the speech of deaf students in order to develop a model of the underlying speech problem and investigate possible methods of intervention. In one such study, timing errors in the speech of deaf children were simulated and then removed systematically in order to determine the effect of different training schemes on speech intelligibility. It was found that prolongation of speech sounds, a common problem in the speech of deaf individuals, was not a major factor in reducing intelligibility but that distortion of the duration ratio between stressed and unstressed syllables reduced intelligibility significantly (Osberger and Levitt, 1979).

Visual displays that present speech information in the form of text are increasingly being used by deaf and severely hearing-impaired individuals. Displays of this type, as categorized by row 4 of [Table 2](#), depend on the use of speech recognition, either by machine or other human beings. In the case of a telephone relay service, the speech produced by the hearing party in a telephone conversation is recognized by a typist at a central location who types out the message, which is then transmitted to the deaf party by means of a text telephone (or TTY/TDD). The deaf party then responds, either by voice or by typing a message on his/her text telephone, which is converted to speech by either a speech synthesizer or the participating typist. Many users of the relay service do not like the participation of a third party (the typist) in their telephone conversations.

In the case of real-time captioning, a stenographer transcribes what is said on a shorthand typewriter (e.g., a Stenograph), the output of which is transmitted to a computer that translates the symbols into standard text. Displays of this type are increasingly being used at lectures and meetings involving deaf participants. These systems have also proved useful to both deaf and hearing students in the

classroom (Stuckless, 1989). A related application of real-time captioning, which is also growing rapidly, is that of captioning movies and television programs for deaf viewers.

Telephone relay services, real-time captioning, and traditional methods of captioning are expensive enterprises, and methods for automating these techniques are currently being investigated. Automatic machine recognition of speech is a possible solution to this problem provided the error rate for unconstrained continuous speech is not too high (Kanevsky et al., 1990; Karis and Dobroth, 1991). Another issue that needs to be considered is that the rate of information transmission for speech can be high and captions often need to be shortened in order to be comprehended by the audience in real-time. Synchronization of the caption to the sound track also needs to be considered as well as issues of lighting and legibility. The preparation of captions is thus not as simple a task as might first appear.

A communication aid that is rapidly growing in importance is the text telephone. Telephone communication by means of teletypewriters (TTYs) has been available to deaf people for many years, but the high cost of TTYs severely limited their use as personal communication aids. The invention of the acoustic coupler and the mass production of keyboards and other basic components of a TTY for the computer market made the text telephone (an offshoot of the TTY) both affordable and practical for deaf consumers. Modern text telephones also make use of computer techniques in storing and retrieving messages in order to facilitate the communication process.

A major limitation of text telephones is the time and effort required to generate messages in comparison with speech. A second limitation is that both parties in a telephone conversation need to have text telephones. The recently introduced telephone relay service eliminates the need for a text telephone by the hearing party, but the rate of communication is slow and there is a loss of privacy when using this service.

Most of the earlier text telephones use the Baudot code for signal transmission, which is a robust but relatively slow code in comparison with ASCII, which is used almost universally for computer communications. The use of the Baudot code makes it difficult for users of these text telephones to access electronic mail, computer notice boards, and computer information services. Methods for improving the interface between text telephones using the Baudot code and modern computer networks using the ASCII code are currently being developed (Harkins et al., 1992). Most modern text telephones are capable of communicating in either Baudot or ASCII.

The ways in which text telephones are used by the deaf commu

nity need to be studied in order to obtain a better understanding of how modern speech recognition and speech synthesis systems could be used for communicating by telephone with deaf individuals. A possible communication link between a deaf person and a hearing person involves the use of a speech recognizer to convert the speech of the hearing person to text for transmission to a text telephone and a speech synthesizer to convert the output of the deaf person's text telephone to speech. The problem of converting speech to text is well documented elsewhere in this volume (see Makhoul and Schwartz). The problem of converting the output of a text telephone to speech in real-time is not quite as great, but there are difficulties beyond those considered in modern text-to-speech systems. For state-of-the-art reviews of text-to-speech conversion see the chapters by Allen and Carlson in this volume. Most deaf users of text telephones, for example, have developed a telegraphic style of communication using nonstandard syntax. In many cases these syntactic forms are carried over from American Sign Language. The use of text generated in this way as input to a speech synthesizer may require fairly sophisticated preprocessing in order to produce synthetic speech that is intelligible to a hearing person.

Issues that need to be considered in converting the output of a text telephone to speech in real-time include:

- whether synthesis should proceed as typed, on a word-byword basis, or whether to introduce a delay and synthesize on a phrase or sentence basis, so as to control intonation appropriately and reduce pronunciation errors due to incorrect parsing resulting from incomplete information;
- rules for preprocessing that take into account the abbreviations, nonstandard syntactic forms, and sparse punctuation typically used with text telephones; and
- methods for handling typing errors or ambiguous outputs.

For example, omission of a space between words is usually easy to recognize in a visual display, but without appropriate preprocessing an automatic speech recognizer will attempt, based on the combination of letters, to produce a single word that is likely to sound quite different from the two intended words.

The above issues present an immediate and interesting challenge to researchers concerned with assistive voice technology.

### Tactile Sensory Aids

The possibility of using vibration as a means of communication was explored by Gault (1926) and Knudsen (1928). This early re

search focused on the use of single-channel devices without any preprocessing to take the characteristics of speech into account. The results showed that only a limited amount of speech information, such as the temporal envelope, could be transmitted in this way. This limited information was nevertheless found to be useful as a supplement to speechreading and for purposes of speech training. Single-channel tactile aids have since found a very useful practical application as alerting systems for deaf people. These devices fall under the category of nonspeech-specific tactile aids (column 3, row 1 of [Table 2](#)).

A tactile aid that takes the spectrum of speech into account is the tactile vocoder. An early vocoder of this type was developed by Gault and Crane (1928), in which each of the fingers was stimulated tactually by a separate vibrator that, in turn, was driven by the speech power in a different region of the frequency spectrum. The tactile vocoder provided significantly more information than a single-channel vibrator, but the amount of information was still insufficient to replace the impaired auditory system, although the device could be used as a supplement to speechreading and for speech training.

The tactile vocoder has been studied intermittently over the years (Pickett and Pickett, 1963; Sherrick, 1984). As transducer technology has improved, coupled with improved signal-processing capabilities, corresponding improvements in speech recognition using tactile aids have been obtained. In a recent study by Brooks et al. (1986), for example, a highly motivated individual with sufficient training acquired a vocabulary of about 1000 words using tactum only. This result is remarkably similar to the best results obtained thus far for a visual spectrum display (see [Figure 2](#)). As noted earlier, spectrograms are hard to read for the reasons cited by Liberman et al. (1968), and the ultimate limitation on the human perception of speech spectrum displays may be independent of whether the information is presented tactually or visually.

An important advantage of tactile sensory aids is that these devices can be worn conveniently without interfering with other important sensory inputs. In addition, a wearable sensory aid allows for long-term exposure to the sensory stimuli, thereby facilitating learning and acclimatization to the sensory aid. There are examples of deaf individuals who have worn spectrum-based tactile aids for many years and who are able to communicate effectively in face-to-face situations using their tactile aid (Cholewiak and Sherrick, 1986). Wearable multichannel tactile aids (two to seven channels) have been available commercially for several years, and several thousand of these devices are already in use, mostly with deaf children (D. Franklin, Audiological Engineering, personal communication, 1993). Experimen

tal evaluations currently in progress indicate that these wearable tactile aids have been significantly more effective in improving speech production than speech recognition skills (Robbins et al., 1992).

Relatively good results have been obtained with speech-feature tactile aids (row 3, column 3 of [Table 2](#)) for both speech production and speech reception. These devices, however, are not as widely used as spectrum-based tactile aids, possibly because they are not yet ready for commercialization.

[Figure 4](#) shows a schematic diagram of a tactile aid worn on the lower forearm that provides information on voice fundamental frequency, FO. The locus of tactile stimulation is proportional to the value of FO. When FO is low, the region near the wrist is stimulated; as FO is increased, the locus of stimulation moves away from the wrist. The use of supplemental FO cues presented tactually has been shown to produce significant improvements in lipreading ability (Boothroyd et al., 1988), although this improvement is less than that

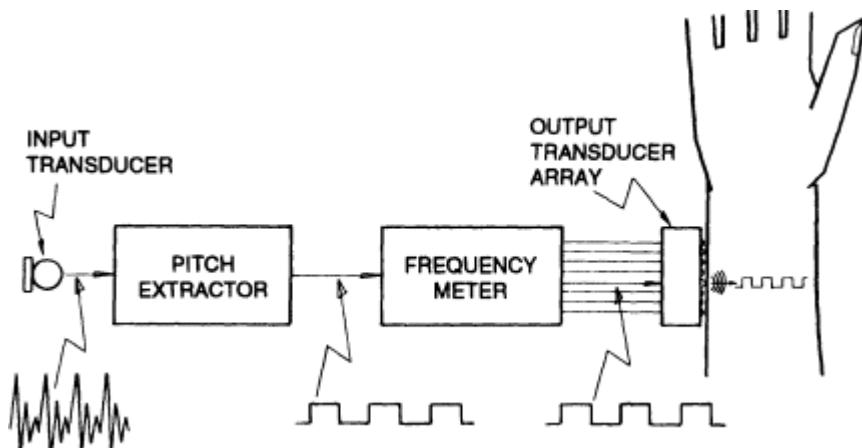


FIGURE 4 Wearable tactile display for voice fundamental frequency. Speech signals from the input transducer (an acoustic microphone or a surface-mounted accelerometer) are delivered to a pitch extractor that generates a square wave whose frequency equals one-half that of the fundamental voice frequency. A pitch period meter times the "off" period of each cycle and thereby determines which of eight output channels will be activated during the "on" period. Frequency scaling is such that the channel number is proportional to the logarithm of voice fundamental frequency. The electronic components are housed in a body-worn unit, the output of which is connected by wire to a transducer array worn on the wrist. Reproduced from Boothroyd and Hnath (1986).

obtained when the supplemental FO cues are presented auditorily (Breeuwer and Plomp, 1986). The tactile FO display has also been found to be of value as a speech-training aid (McGarr et al., 1992). A very useful feature of this aid is that it can be worn outside the classroom, thereby providing students with continuous feedback of their speech production in everyday communication.

It is important for any speech-feature sensory aid (tactile, visual, or auditory) that the speech features be extracted reliably. An experiment demonstrating the value of a tactile aid providing reliable speech-feature information has been reported by Miller et al. (1974). The key elements of the experiment are illustrated in [Figure 5](#). Several sensors are mounted on the speaker. One picks up nasal vibrations; the second picks up vocal cord vibrations; and the third, a microphone, picks up the acoustic speech signal. These signals are delivered to the subject by means of vibrators used to stimulate the subject's fingers. The subject can see the speaker's face but cannot hear what is said since the speech signal is masked by noise delivered by headphones.

The vibrators provide reliable information on three important aspects of speech production: nasalization, voicing, and whether or not speech is present. The first two of these cues are not visible in speechreading, and the third cue is sometimes visually ambiguous. Use of the tactual cues resulted in significant improvements in speechreading ability with relatively little training (Miller et al., 1974). An important implication of this experiment is that speech-feature tactile aids are of great potential value for improving articulatory cues provided by speech recognition. Cues designed to supplement those used in speechreading are delivered reliably to the user.

The central problem facing the development of practical speech-feature tactile aids is that of extracting the articulatory cues reliably. This problem is particularly difficult for wearable sensory aids because the signal picked up by the microphone on the wearable aid is contaminated by environmental noise and reverberation.

Tactile sensory aids involving speech recognition processing (row 4, column 3 of [Table 2](#)) are at the threshold of development. These devices have the potential for improving communication with deaf-blind individuals. In one such system, a discrete word recognizer is used to recognize speech produced by a hearing person. The speech is then recognized and converted to Braille (or a simpler form of raised text) for use by the deaf-blind participant in the conversation. A system of this type has already been developed for Japanese symbols by Shimizu (1989).

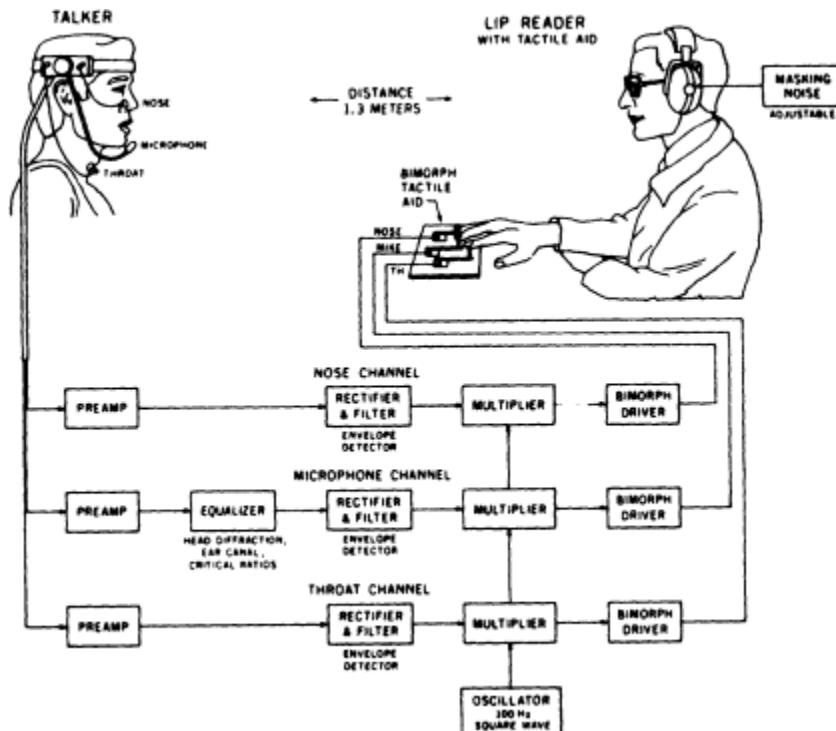


FIGURE 5 Three-channel speech-feature tactile aid. Three sensors are used to pick up nasal vibration, throat vibration, and the acoustic speech signal, respectively. The envelope of each signal is amplified and delivered to three vibrators for stimulating the subject's fingers. The vibrotactile stimuli served as supplementary speechreading cues. Masking noise is used to eliminate acoustic cues during testing. Reproduced from Miller et al. (1974).

### Direct Electrical Stimulation of the Auditory System

It has been demonstrated that a deaf person can hear sound when the cochlea is stimulated electrically (Simmons, 1966). One possible explanation for this is that the deafness may be due to damage to the hair cells in the cochlea which convert mechanical vibration to neural firings, but that the neurons connected to these hair cells may still be functional and can be triggered by the electromagnetic field generated by a cochlear implant.

Most of the early cochlear implants consisted of a single wire electrode inserted in the cochlea (House and Urban, 1973). These implants were not designed specifically for speech signals but were

very helpful in providing users with the sensation of sound and the ability to recognize common environmental sounds.

More advanced cochlear implants have been designed specifically for speech. In one such prosthesis, the Ineraid, the speech signal is filtered into four frequency bands. The signal in each band is then transmitted through the skin (using a percutaneous plug) to a set of electrodes inserted in the cochlea (Eddington, 1983). The geometry of the electrode array is such that the four frequency bands stimulate different regions of the cochlea in a manner consistent with the frequency-analyzing properties of the cochlea. This cochlear implant is categorized by row 2, column 4 of [Table 2](#).

A cochlear implant that uses speech-feature processing (the Nucleus 22-channel cochlear implant system) is shown in [Figure 6](#). This prosthesis, categorized by row 3, column 4 of [Table 2](#), was developed by Clark et al. (1983, 1990). A body-worn signal processor, shown on the left in the figure, is used to analyze the incoming acoustic signal picked up by an ear-mounted microphone. The FO and the two lowest-formant frequencies (F1 and F2) are extracted from the acoustic signal, modulated on a carrier, and then transmitted electromagnetically across the skin to a decoding unit mounted in the mastoid (a bony section of the skull behind the ear). The decoded signals take the form of pulses that are transmitted to an array of electrodes inserted in the cochlea, as shown by the dotted line in the photograph. The electrode array consists of a flexible cable-like structure ringed with 32 bands of platinum of which 10 serve as mechanical support and the remaining 22 bands serve as electrodes. Each electrode is connected to the decoder in the mastoid by means of an extremely fine flexible wire contained within the cable.

The pulses from the decoder are delivered to the electrodes so as to stimulate those regions of the cochlea that in normal speech perception would correspond roughly to the frequency regions that would be stimulated by formants 1 and 2. The rate of pulsatile stimulation is proportional to FO for voiced sounds, random stimulation being used for voiceless sounds. The pulses used to stimulate different frequency regions of the cochlea are interleaved in time so as to reduce interference between adjacent channels of stimulation. The method of stimulation thus provides important speech-feature information (FO, formant frequencies F1 and F2, and whether the speech is voiced or voiceless).

A third type of cochlear prosthesis uses an external electrode mounted on the round window of the cochlea. An advantage of this type of prosthesis, known as an extracochlear implant, is that electrodes are not inserted into the cochlea with the attendant danger of

damaging whatever residual hearing may be left. A disadvantage of this approach is that relatively high-current densities are needed, which can cause nonauditory side effects, such as facial nerve stimulation. The use of a single electrode also limits the amount of speech information that can be delivered. In the prosthesis developed by Douek et al. (1977), FO was delivered to the electrode. Significant improvements in speechreading ability were obtained using this technique.

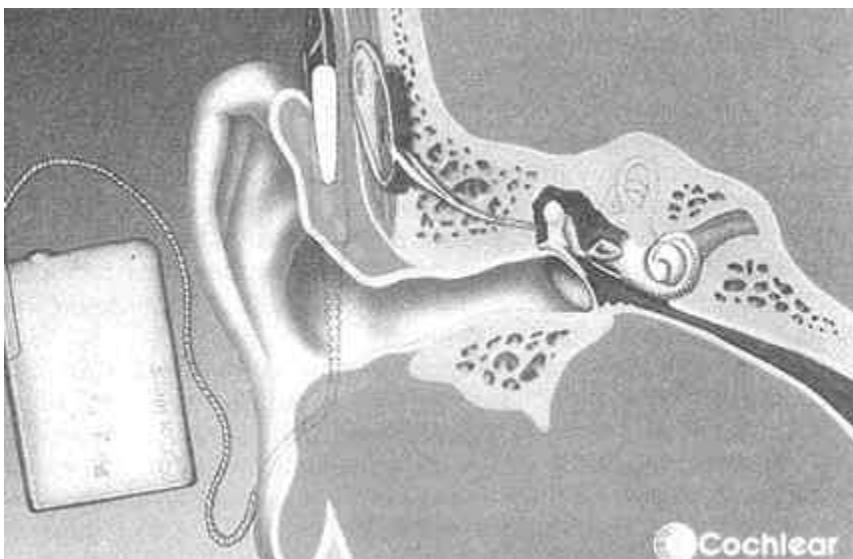


FIGURE 6 Multichannel speech-feature cochlear prosthesis (Nucleus 22-channel cochlear implant system). Speech is picked up by a microphone and delivered to a body-worn processor (shown on the left) that extracts and encodes the speech features of interest (voice fundamental frequency, formant frequencies). The encoded signals are transmitted electromagnetically across the skin. The decoding unit is mounted in the mastoid (a bony section of the skull behind the ear). The decoded signals take the form of pulses that are transmitted to an array of electrodes inserted in the cochlea, as shown by the dotted line in the photograph. Diagram reprinted courtesy of the Cochlear Corporation.

Much controversy exists regarding the relative efficacy of the various cochlear implants that have been developed. Dramatic improvements in speech understanding have been demonstrated for each of the major types of cochlear implants. For the more advanced implants, a high proportion of implantees have even demonstrated the ability to converse without the use of speechreading, as in a telephone conver

sation. These success stories have been widely publicized; in the majority of cases, however, improvements in communication ability have been significant but much less dramatic.

A major clinical trial was recently undertaken in an attempt to resolve the controversy regarding relative efficacy (Cohen et al., 1993). Three types of cochlear implants were compared: a single-channel implant, a multichannel spectrum-based implant, and a multichannel speech-feature implant. The results showed that both multichannel implants produced significantly greater improvements in communication ability than the single-channel implant. Differences between the two multichannel implants were either small or not statistically significant. Toward the end of the clinical trial, a new form of stimulus coding was introduced for the multichannel speech-feature cochlear implant. The improved processor encoded additional spectral information in three high-frequency bands by stimulating electrodes close to the high-frequency end of the cochlea. This new method of coding was found to provide a small but significant improvement in communication ability. New methods of coding using interleaved pulsatile stimuli for spectrum-based cochlear implants are currently being investigated (Wilson et al., 1993).

The one entry in [Table 2](#) that, at present, remains unfilled is that of a cochlear prosthesis using speech recognition technology (row 4, column 4). A possible future application of this technology would be a system analogous to the speech recognition/speech synthesis hearing aid in which speech is first recognized and then regenerated so as to enhance those speech features that cochlear implant users have difficulty understanding.

### Noise Reduction

A common problem with all sensory aids is the damaging effects of background noise and reverberation on speech understanding. These distortions also reduce overall sound quality for hearing aids and cochlear implants. Signal processing for noise reduction is an important but large topic, the details of which are beyond the scope of this chapter. There are, however, several aspects of the noise reduction problem that are unique to sensory aids and that need to be mentioned.

The most common method of combating noise and reverberation in practice is to place a microphone close to the speaker and transmit the resulting signal, which is relatively free of noise and reverberation, directly to the input of the sensory aid. FM radio or infrared transmission systems are commonly used for this purpose. It is not

always feasible to use a remote microphone in this way, and various alternative solutions to the problem are currently being investigated.

A method for reducing the effects of noise that is currently used in many modern hearing aids is to filter out high level frequency components that are believed to be the result of amplifying environmental noise. It has been argued that the intense components of the amplified sound, which may include noise, will mask the less intense components of the amplified sound and that these less intense components are likely to include important components of the speech signal.

Many common environmental noises have relatively powerful low-frequency components that, if amplified, would interfere with the weaker high-frequency components of speech. High-pass filters are typically used to reduce this type of environmental noise, but these filters also eliminate the speech signal in those frequency bands, thereby reducing intelligibility when only speech is present. To address this problem, adaptive filtering is used in which the cutoff frequency of the filter is adjusted automatically depending on the level of the signals to be amplified. If the low-frequency power is relatively high (e.g., the ratio of low- to high-frequency power is larger than that typical of speech), it is assumed that most of the low-frequency power is due to environmental noise, and the filter cutoff frequency is adjusted so as to attenuate the low frequencies. In this way the filter eliminates mostly noise, although some components of the speech signal will still be lost. An alternative approach to this problem is to use two-channel compression amplification and to reduce the gain in the low-frequency channel when there is excessive low-frequency power.

Experimental evaluations of the above forms of noise reduction have yielded mixed results. Some studies have reported significant gains in speech intelligibility, whereas other studies have shown decrements (Fabry, 1991; Fabry et al., 1993; Graupe et al., 1987; Van Tasell et al., 1988). Most of these studies, unfortunately, did not use appropriate control conditions and, as a consequence, these studies cannot be treated as conclusive (Dillon and Lovegrove, 1993). Substantial individual differences have also been observed in most of these studies, indicating that, depending on the nature of the hearing impairment and the circumstances under which the sensory aid is used, different forms of signal processing may be required for different people and conditions of use (Levitt, 1993).

A form of noise reduction that has yielded good results as a front end to a speech-feature cochlear implant is that of spectrum subtraction (Hochberg et al., 1992). This form of noise reduction is relatively effective in extracting the strongly voiced components of the speech

signal from a random noise background. Since the Nucleus speech-feature cochlear implant depends heavily on reliable estimation of the first two formant frequencies during voiced sounds, it is not surprising that a variation of the spectrum-subtraction method of noise reduction (the INTEL technique developed by Weiss and Aschkenasy, 1981) showed a significant improvement in speech understanding when used as a front end to this cochlear implant. The improvements obtained were equivalent, on average, to an increase in the speech-to-noise ratio of 5 dB.

Perhaps the most promising form of noise reduction for wearable sensory aids at the present time is that of a microphone array for automatically focusing in on the speech source. A directional array of this type using microphones mounted on the frame of a pair of eyeglasses can produce improvements in speech-to-noise ratio of 7 to 11 dB under conditions typical of everyday use (Bilsen et al., 1993; Soede, 1990). A relatively simple nonadaptive method of signal processing was used in the above study. Adaptive methods of noise cancellation using two or more microphones appear to be even more promising (Chabries et al., 1987; Peterson et al., 1987; Schwander and Levitt, 1987). An important practical limitation of these techniques, however, is that they will not provide any benefit in a highly reverberant acoustic environment or if both speech and noise come from the same direction.

## **OTHER FORMS OF ASSISTIVE TECHNOLOGY INVOLVING VOICE COMMUNICATION**

### **Speech Processing for Sightless People**

People with severe visual impairments are heavily dependent on the spoken word in communicating with others and in acquiring information. To address their needs, talking books have been developed. High-speed playback is a very useful tool for scanning an audio text, provided the speech is intelligible. Methods of processing speeded speech to improve intelligibility have been developed using proportional frequency lowering. When an audio recording is played back at  $k$  times its normal speed, all frequencies in that recording are increased proportionally by the ratio  $k$ . The effect of this distortion on intelligibility is reduced considerably if the playback signal is transposed downward by the same frequency ratio. Methods of proportional frequency transposition were developed some time ago (Levitt et al., 1980). These techniques work well for steady-state sounds but are subject to distortion for transient sounds. Re

cent advances in compact disc technology and computer-interactive media have eliminated much of the inconvenience in scanning recorded texts.

A second application of assistive speech technology is the use of machine-generated speech for voice output devices. These applications include reading machines for the blind (Cooper et al., 1984; Kurzweil, 1981); talking clocks and other devices with voice output; and, in particular, voice output systems for computers. The use of this technology is growing rapidly, and, with this growth, new problems are emerging.

One such problem is linked to the growing complexity of computer displays. Pictorial symbols and graphical displays are being used increasingly in modern computers. Whereas computer voice output systems are very effective in conveying alphanumeric information to sightless computer users, conveying graphical information is a far more difficult problem. Innovative methods of representing graphical information using machine-generated audio signals combined with tactile displays are being experimented with and may provide a practical solution to this problem (Fels et al., 1992).

A second emerging problem is that many sightless people who are heavily dependent on the use of machine-generated speech for both employment and leisure are gradually losing their hearing as they grow older. This is a common occurrence in the general population, as indicated by the large proportion of people who need to wear hearing aids in their later years (see [Table 1](#)). Machine-generated speech is usually more difficult to understand than natural speech. For older people with some hearing loss the difficulty in understanding machine-generated speech can be significant. For the sightless computer user whose hearing is deteriorating with age, the increased difficulty experienced in understanding machine-generated speech is a particularly troublesome problem.

Good progress is currently being made in improving both the quality and intelligibility of machine-generated speech (Bennett et al., 1993). For the special case of a sightless person with hearing loss, a possible approach for improving intelligibility is to enhance the acoustic characteristics of those information-bearing components of speech that are not easily perceived as a result of the hearing loss. The thrust of this research is similar to that underlying the development of the speech recognition/speech synthesis hearing aid (Levitt et al., 1993).

## Augmentative and Alternative Communication

People with a range of different disabilities depend on the use of augmentative and alternative methods of communication (AAC). Machine-generated speech is widely used in AAC, although the way in which this technology is employed depends on the nature of the disability. A nonvocal person with normal motor function may wish to use a speech synthesizer with a conventional keyboard, whereas a person with limited manual dexterity in addition to a severe speech impairment would probably use a keyboard with a limited set of keys or a non-keyboard device that can be controlled in various ways other than by manual key pressing. It is also possible for a person with virtually no motor function to use eye movements as a means of identifying letters or symbols to be used as input to the speech synthesizer.

A common problem with the use of synthetic speech in AAC is the time and effort required to provide the necessary input to the speech synthesizer. Few people can type at speeds corresponding to that of normal speech. It is possible for a motivated person with good motor function to learn how to use either a Stenograph or Palantype keyboard at speeds comparable to normal speech (Arnott, 1987). The output of either of these keyboards can be processed by computer so as to drive a speech synthesizer in real-time. A high-speed keyboard, however, is not practical for a nonvocal person with poor motor function. It is also likely that for this application a keyboard with a limited number of keys would be used. In a keyboard of this type, more than one letter is assigned to each key, and computer techniques are needed to disambiguate the typed message. Methods of grouping letters efficiently for reduced set keyboards have been investigated so as to allow for effective disambiguation of the typed message (Levine et al., 1987). Techniques of this type have also been used with a Touch-Tone® keypad so that a hearing person using a Touch-Tone telephone can communicate with a deaf person using a text telephone (Harkins et al., 1992). Even with these innovations, this technique has not met with much success among people who are not obliged to use a reduced set keyboard.

Other methods of speeding up the communication process is to use the redundancy of language in order to predict which letter or word should come next in a typed message (Bentrup, 1987; Damper, 1986; Hunnicutt, 1986, 1993). A variation of this approach is to use a dictionary based on the user's own vocabulary in order to improve the efficiency of the prediction process (Swiffin et al., 1987). The

dictionary adapts continuously to the user's vocabulary as the communication process proceeds.

For nonvocal people with limited language skills, methods of generating well-formed sentences from limited or ill-formed input are being investigated (McCoy et al., 1989). In some applications, symbols rather than words are used to generate messages that are then synthesized (Hunnicutt, 1993).

Computer-assisted instruction using voice communication can be particularly useful for people with disabilities. The use of computers for speech training has already been discussed briefly in the section on visual sensory aids. Similarly, computer techniques for improving human-machine communication can be of great value in developing instructional systems for people who depend on augmentative or alternative means of communication. A relatively new application of this technology is the use of computerized speech for remedial reading instruction (Wise and Olsen, 1993).

A common problem with text-to-speech systems is the quality of the synthetic voice output. Evaluations of modern speech synthesizers indicate that these systems are still far from perfect (Bennett et al., 1993). Much of the effort in the development of text-to-speech systems has been directed toward making the synthesized speech sound natural as well as being intelligible. In some applications, such as computer voice output for a hearing-impaired sightless person, naturalness of the synthesized speech may need to be sacrificed in order to improve intelligibility by exaggerating specific features of the speech signal.

The artificial larynx represents another area in which advances in speech technology have helped people with speech disabilities. The incidence of laryngeal cancer has grown over the years, resulting in a relatively large number of people who have had a laryngectomy. Many laryngectomies use an artificial larynx in order to produce intelligible speech. Recent advances in speech technology coupled with an increased understanding of the nature of speech production have resulted in significant improvements in the development of artificial larynxes (Barney et al., 1959; Sekey, 1982). Recent advances include improved control of speech prosody and the use of microprocessor-generated glottal waveforms based on recent theories of vocal cord vibration in order to produce more natural-sounding speech (Alzamora et al., 1993).

Unlike the applications of assistive technology discussed in the previous sections, the use of voice communication technology in AAC is highly individualized. Nevertheless, two general trends are apparent. The first is the limitation imposed by the relatively low speed

at which control information can be entered to the speech synthesizer by the majority of candidates for this technology. The second is the high incidence of multiple disabling conditions and the importance of taking these factors into account in the initial design of assistive devices for this population. In view of the above, an important consideration for the further development of assistive technology for AAC is that of developing flexible generalized systems that can be used for a variety of applications involving different combinations of disabling conditions (Hunnicutt, 1993).

### **Assistive Voice Control: Miscellaneous Applications**

Assistive voice control is a powerful enabling technology for people with mobility disabilities. Examples of this new technology include voice control of telephones, home appliances, powered hospital beds, motorized wheelchairs, and other such devices (Amori, 1992; Miller, 1992). In each case the set of commands to control the device is small, and reliable control using a standardized set of voice commands can be achieved using existing speech recognition technology. Several of these applications of voice control technology may also find a market among the general population (e.g., voice control of a cellular telephone while driving a car, remote control of a television set, or programming a VCR by voice). A large general market will result in mass production of the technology, thereby reducing cost for people who need this technology because of a mobility disability.

An additional consideration in developing assistive voice control technology is that a person with a neuromotor disability may also have dysarthria (i.e., unclear speech). For disabilities of this type, the speech recognition system must be capable of recognizing various forms of dysarthric speech (Goodenough-Trepagnier et al., 1992). This is not too difficult a problem for speaker-dependent automatic speech recognition for those dysarthrics whose speech production is consistent although not normal. Machine recognition of dysarthric speech can also be of value in assisting clinicians obtain an objective assessment of the speech impairment.

Voice control of a computer offers intriguing possibilities for people with motor disabilities. Relatively simple methods of voice control have already been implemented, such as remote control of a computer mouse (Miller, 1992), but a more exciting possibility that may have appeal to the general population of computer users (and, concomitantly, may reduce costs if widely used) is that of controlling or programming a computer by voice. The future in this area of assistive voice control technology appears most promising.

## ACKNOWLEDGMENT

Preparation of this paper was supported by Grant No. 5P50DC00178 from the National Institute on Deafness and Other Communication Disorders.

## REFERENCES

- Alzamora, D., D. Silage, and R. Yantorna (1993). Implementation of a software model of the human glottis on a TMS32010 DSP to drive an artificial larynx. Proceedings of the 19th Northeast Bioengineering Conference, pp. 212-214, New York: Institute of Electrical and Electronic Engineers.
- Amori, R. D. (1992). Vocomotion: An intelligent voice-control system for powered wheelchairs. Proceedings of the RESNA International '92 Conference, pp. 421423. Washington, D.C.: RESNA Press.
- Arnott, J. L. (1987). A comparison of Palantype and Stenograph keyboards in high-speed speech output systems. RESNA '87, Proceedings of the 10th Annual Conference on Rehabilitation Technology, pp. 106-108. Washington D.C.: RESNA Press.
- Barney, H., F. Haworth, and H. Dunn (1959). Experimental transistorized artificial larynx. Bell Syst. Tech. J., 38:1337-1359.
- Bell, A. G. (1876). Researches in telephony. Proceedings of the American Academy of Arts and Sciences, XII, pp. 1-10. Reprinted in Turning Points in American Electrical History, Britain, J.E., Ed., IEEE Press, New York, 1976.
- Bennett, R. W., A. K. Syrdal, and S. L. Greenspan (Eds.) (1993). Behavioral Aspects of Speech Technology. Amsterdam: Elsevier Science Publishing Co.
- Bentrup, J. A. (1987). Exploiting word frequencies and their sequential dependencies. RESNA '87, Proceedings of the 10th Annual Conference on Rehabilitation Technology, pp. 121-123. Washington, D.C.: RESNA Press.
- Bilser, F. A., W. Soede, and J. Berkhout (1993). Development and assessment of two fixed-array microphones for use with hearing aids. J. Rehabil. Res. Dev., 30(1):7381.
- Boothroyd, A., and T. Hnath (1986). Lipreading with tactile supplements. J. Rehabil. Res. Dev., 23 (1):139-146.
- Boothroyd, A., T. Hnath-Chisolm, L. Hanin, and L. Kishon-Rabin (1988). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. Ear Hear., 9:306-312.
- Braida, L. D., N. L. Durlach, R. P. Lippmann, B. L. Hicks, W. M. Rabinowitz, and C. M. Reed (1979). Hearing Aids-A Review of Past Research of Linear Amplification, Amplitude Compression and Frequency Lowering. ASHA Monograph No. 19. Rockville, Md.: American Speech-Language-Hearing Association.
- Breeuwer, M., and R. Plomp (1986). Speechreading supplemented with auditorily-presented speech parameters. J. Acoust. Soc. Am., 79:481-499.
- Brooks, P. L., B. U. Frost, J. L. Mason, and D. M. Gibson (1986). Continuing evaluation of the Queens University Tactile Vocoder, Parts I and II. J. Rehabil. Res. Dev., 23(1):119-128, 129-138.
- Chabries, D. M., R. W. Christiansen, R. H. Brey, M. S. Robinette, and R. W. Harris (1987). Application of adaptive digital signal processing to speech enhancement for the hearing impaired. J. Rehabil. Res. Dev., 24(4):65-74.
- Cholewiak, R. W., and C. E. Sherrick (1986). Tracking skill of a deaf person with long-term tactile aid experience: A case study. J. Rehabil. Res. Dev., 23(2):20-26.

- Clark, G. M., R. K. Shepherd, J. F. Patrick, R. C. Black, and Y. C. Tong (1983). Design and fabrication of the banded electrode array. *Ann. N.Y. Acad. Sci.*, 405:191-201.
- Clark, G. M., Y. C. Tong and J. F. Patrick (1990). *Cochlear Prostheses*. Edinburgh, London, Melbourne and New York: Churchill Livingstone.
- Cohen, N. L., S. B. Waltzman, and S. G. Fisher (1993). A prospective randomized cooperative study of advanced cochlear implants. *N. Engl. J. Med.*, 43:328.
- Cole, R. A., A. I. Rudnicky, V. W. Zue, and D. R. Reddy (1980). Speech as patterns on paper. In *Perception and Production of Fluent Speech*, R. A. Cole (Ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Cooper, F. S., J. H. Gaitenby, and P. W. Nye (1984). Evolution of reading machines for the blind: Haskins Laboratories' research as a case history. *J. Rehabil. Res. Dev.*, 21:51-87.
- Cornett, R. O. (1967). Cued speech. *Am. Ann. Deaf*, 112:3-13.
- Damper, R. I. (1986). Rapid message composition for large-vocabulary speech output aids: a review of possibilities. *J. Augment. Altern. Commun.*, 2:4.
- Dillon, H., and R. Lovegrove (1993). Single-microphone noise reduction systems for hearing aids: A review and an evaluation. In *Acoustical Factors Affecting Hearing Aid Performance*, 2nd Ed., G. A. Studebaker and I. Hochberg (Eds.), pp. 353-370, Needham Heights, Mass.: Allyn and Bacon.
- Douek, E., A. J. Fourcin, B. C. J. Moore, and G. P. Clarke (1977). A new approach to the cochlear implant. *Proc. R. Soc. Med.*, 70:379-383.
- Eddington, D. K. (1983). Speech recognition in deaf subjects with multichannel intracochlear electrodes. *Ann. N.Y. Acad. Sci.*, 405:241-258.
- Fabry, D. A. (1991). Programmable and automatic noise reduction in existing hearing aids. In *The Vanderbilt Hearing Aid Report II*, G. A. Studebaker, F. H. Bess, and L. B. Beck (Eds.), pp. 65-78. Parkton, Md.: York Press.
- Fabry, D. A., M. R. Leek, B. E. Walden, and M. Cord (1993). Do adaptive frequency response (AFR) hearing aids reduce upward spread of masking? *J. Rehabil. Res. Dev.*, 30(3):318-323.
- Fels, D. I., G. F. Shein, M. H. Chignell, and M. Milner (1992). See, hear and touch the GUI: Computer feedback through multiple modalities. Proceedings of the RESNA International '92 Conference, pp. 55-57, Washington, D.C.: RESNA Press.
- Gault, R. H. (1926). Touch as a substitute for hearing in the interpretation and control of speech. *Arch. Otolaryngol.*, 3:121-135.
- Gault, R. H., and G. W. Crane (1928). Tactual patterns from certain vowel qualities instrumentally communicated from a speaker to a subject's fingers. *J. Gen. Psychol.*, 1:353-359.
- Gengel, R. W. (1976). Upton's wearable eyeglass speechreading aid: History and current status. In *Hearing and Davis: Essays Honoring Hallowell Davis*, S. K. Hirsh, D. H. Eldredge, I. J. Hirsh, and R. S. Silverman (Eds.). St. Louis, Mo.: Washington University Press.
- Goldberg, A. J. (1972). A visible feature indicator for the severely hard of hearing. *IEEE Trans. Audio Electroacoust.*, AU-20:16-23.
- Goodenough-Trepagnier, C., H. S. Hochheiser, M. J. Rosen, and H-P. Chang (1992). Assessment of dysarthric speech for computer control using speech recognition: Preliminary results. Proceedings of the RESNA International '92 Conference, pp. 159-161. Washington, D.C.: RESNA Press.
- Graupe, D., J. K. Grosspietsch, and S. P. Basseas (1987). A single-microphone-based self-adaptive filter of noise from speech and its performance evaluation. *J. Rehabil. Res. Dev.*, 24 (4):119-126.

- Guttman, N., H. Levitt, and P. A. Bellefleur (1970). Articulation training of the deaf using low-frequency surrogate fricatives. *J. Speech Hear. Res.*, 13:19-29.
- Harkins, J. E., and B. M. Virvan (1989). *Speech to Text: Today and Tomorrow*. Proceedings of a Conference at Gallaudet University. GRI Monograph Series B, No. 2. Washington, D.C.: Gallaudet Research Institute, Gallaudet University.
- Harkins, J. E., H. Levitt, and K. Peltz-Strauss (1992). *Technology for Relay Service, A Report to the Iowa Utilities Board*. Washington, D.C.: Technology Assessment Program, Gallaudet Research Institute, Gallaudet University.
- Hochberg, I. H., A. S. Boothroyd, M. Weiss, and S. Hellman (1992). Effects of noise suppression on speech perception by cochlear implant users. *Ear Hear.*, 13(4):263-271.
- House, W., and J. Urban (1973). Long-term results of electrical implantation and electronic stimulation of the cochlea in man. *Ann. Otol. Rhinol. Laryngol.*, 82:504-510.
- House, A. S., D. P. Goldstein, and G. W. Hughes (1968). Perception of visual transforms of speech stimuli: Learning simple syllables. *Am. Ann. Deaf*, 113:215-221.
- Hunnicutt, S. (1986). Lexical prediction for a text-to-speech system. In *Communication and Handicap: Aspects of Psychological Compensation and Technical Aids*, E. Hjelmquist & L-G. Nilsson, (Eds.). Amsterdam: Elsevier Science Publishing Co.
- Hunnicutt, S. (1993). Development of synthetic speech technology for use in communication aids. In *Behavioral Aspects of Speech Technology*, R. W. Bennett, A. K. Syrdal, and S. L. Greenspan (Eds.). Amsterdam: Elsevier Science Publishing Co.
- Johansson, B. (1966). The use of the transposer for the management of the deaf child. *Int. Audiol.*, 5:362-373.
- Kanevsky, D., C. M. Danis, P. S. Gopalakrishnan, R. Hodgson, D. Jameson, and D. Nahamoo (1990). A communication aid for the hearing impaired based on an automatic speech recognizer. In *Signal Processing V: Theories and Applications*, L. Torres, E. Masgrau, and M. A. Lagunas (Eds.). Amsterdam: Elsevier Science Publishing Co.
- Karis, D., and K. M. Dobroth (1991). Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE J. Select. Areas Commun.*, 9(4).
- Kewley-Port, D., C. S. Watson, D. Maki, and D. Reed (1987). Speaker-dependent speech recognition as the basis for a speech training aid. *Proceedings of the 1987 IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 372-375, Dallas, Tex.: Institute of Electrical and Electronic Engineering.
- Knudsen, V. O. (1928). Hearing with the sense of touch. *J. Gen. Psychol.*, 1:320-352.
- Kurzweil, R. C. (1981). Kurzweil reading machine for the blind. *Proceedings of the Johns Hopkins First National Search for Applications of Personal Computing to Aid the Handicapped*, pp. 236-241. New York: IEEE Computer Society Press.
- LaPlante, M. P., G. E. Hendershot, and A. J. Moss (1992). Assistive technology devices and home accessibility features: Prevalence, payment, need, and trends. *Advance Data*, Number 127. Atlanta: Vital and Health Statistics of the Centers for Disease Control/National Center for Health Statistics.
- Levine, S. H., C. Goodenough-Trepagnier, C. O. Getschow, and S. L. Minneman (1987). Multi-character key text entry using computer disambiguation. *RESNA '87, Proceedings of the 10th Annual Conference on Rehabilitation Technology*, pp. 177-179. Washington, D.C.: RESNA.
- Levitt, H. (1993). Future directions in hearing aid research. *J. Speech-Lang.-Pathol. Audiol.*, Monograph Suppl. 1, pp. 107-124.

- Levitt, H., J. M. Pickett, and R. A. Houde (Eds.) (1980). *Sensory Aids for the Hearing Impaired*. New York: IEEE Press.
- Levitt, H., M. Bakke, J. Kates, A. Neuman, and M. Weiss (1993). Advanced signal processing hearing aids. In *Recent Developments in Hearing Instrument Technology, Proceedings of the 15th Danavox Symposium*, J. Beilen and G. R. Jensen (Eds.), pp. 247-254. Copenhagen: Stougaard Jensen.
- Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1968). Why are spectrograms hard to read? *Am. Ann. Deaf*, 113:127-133.
- Ling, D. (1969). Speech discrimination by profoundly deaf children using linear and coding amplifiers. *IEEE Trans. Audio Electroacoust.*, AU-17:298-303.
- Mazor, M., H. Simon, J. Scheinberg, and H. Levitt (1977). Moderate frequency compression for the moderately hearing impaired. *J. Acoust. Soc. Am.*, 62:1273-1278.
- McCoy, K., P. Demasco, Y. Gong, C. Pennington, and C. Rowe (1989). A semantic parser for understanding ill-formed input. *RESNA '89, Proceedings of the 12th Annual Conference*, pp. 145-146. Washington, D.C.: RESNA Press.
- McGarr, N. S., K. Youdelman, and J. Head (1992). *Guidebook for Voice Pitch Remediation in Hearing-Impaired Speakers*. Englewood, Colo.: Resource Point.
- Miller, G. (1992). Voice recognition as an alternative computer mouse for the disabled. *Proceedings of the RESNA International '92 Conference*, pp. 55-57, Washington D.C.: RESNA Press.
- Miller, J. D., A. M. Engebretsen, and C. L. DeFilippo (1974). Preliminary research with a three-channel vibrotactile speech-reception aid for the deaf. In *Speech Communication*, Vol. 4, *Proceedings of the Speech Communication Seminar*, G. Fant (Ed.). Stockholm: Almqvist and Wiksell.
- Nicholls, G., and D. Ling (1982). Cued speech and the reception of spoken language. *J. Speech Hear. Res.*, 25:262-269.
- Osberger, M. J., and H. Levitt (1979). The effect of timing errors on the intelligibility of deaf children's speech. *J. Acoust. Soc. Am.*, 66:1316-1324.
- Peterson, P. M., N. I. Durlach, W. M. Rabinowitz, and P. M. Zurek (1987). Multimicrophone adaptive beam forming for interference reduction in hearing aids. *J. Rehabil. Res. Dev.*, 24 (4):103-110.
- Pickett, J. M., and B. M. Pickett (1963). Communication of speech sounds by a tactual vocoder. *J. Speech Hear. Res.*, 6:207-222.
- Pickett, J. M., R. W. Gengel, and R. Quinn (1974). Research with the Upton eyeglass speechreader. In *Speech Communication*, Vol. 4, *Proceedings of the Speech Communication Seminar*, G. Fant (Ed.). Stockholm: Almqvist and Wiksell.
- Posen, M. P., C. M. Reed, and L. D. Braida (1993). The intelligibility of frequency-lowered speech produced by a channel vocoder. *J. Rehabil. Res. Dev.*, 30(1):26-38.
- Potter, R. K., A. G. Kopp, and H. C. Green (1947). *Visible Speech*. New York: van Nostrand Co.
- Revoile, S. G., L. Holden-Pitt, J. Pickett, and F. Brandt (1986). Speech cue enhancement for the hearing impaired: I. Altered vowel durations for perception of final fricative voicing. *J. Speech Hear. Res.*, 29:240-255.
- Risberg, A. (1968). Visual aids for speech correction. *Am. Ann. Deaf*, 113:178-194.
- Robbins, A. M., S. L. Todd, and M. J. Osberger (1992). Speech perception performance of pediatric multichannel tactile aid or cochlear implant users. In *Proceedings of the Second International Conference on Tactile Aids, Hearing Aids, and Cochlear Implants*, A. Risberg, S. Felicetti, G. Plant, and K. E. Spens (Eds.), pp. 247-254. Stockholm: Royal Institute of Technology (KTH).
- Rosen, S., J. R. Walliker, A. Fourcin, and V. Ball (1987). A micro-processor-based acoustic

- hearing aid for the profoundly impaired listener. *J. Rehabil. Res. Dev.*, 24(4):239-260.
- Ryalls, J., M. Cloutier, and D. Cloutier (1991). Two clinical applications of IBM's SpeechViewer: Therapy and its evaluation on the same machine. *J. Comput. User's Speech Hear.*, 7 (1):22-27.
- Schein, J. D., and M. T. Delk (1974). The Deaf Population of the United States. Silver Spring, Md.: National Association of the Deaf.
- Schwander, T. J., and H. Levitt (1987). Effect of two-microphone noise reduction on speech recognition by normal-hearing listeners. *J. Rehabil. Res. Dev.*, 24(4):87-92.
- Sekey, A. (1982). Electroacoustic Analysis and Enhancement of Alaryngeal Speech. Springfield, Ill.: Charles C. Thomas.
- Sherrick, C. E. (1984). Basic and applied research on tactile aids for deaf people: Progress and prospects. *J. Acoust. Soc. Am.*, 75:1325-1342.
- Shimizu, Y. (1989). Microprocessor-based hearing aid for the deaf. *J. Rehabil. Res. Dev.*, 26 (2):25-36.
- Simmons, R. B. (1966). Electrical stimulation of the auditory nerve in man. *Arch. Otolaryngol.*, 84:2-54.
- Skinner, M. W. (1988). Hearing Aid Evaluation. Englewood Cliffs, N.J.: Prentice-Hall.
- Soede, W. (1990). Improvement of speech intelligibility in noise: Development and evaluation of a new directional hearing instrument based on array technology. Ph.D. thesis, Delft University of Technology, The Netherlands.
- Stark, R. E. (1972). Teaching /ba/ and /pa/ to deaf children using real-time spectral displays. *Lang. Speech*, 15:14-29.
- Stuckless, E. R. (1989). Real-time captioning in education. In *Speech to Text: Today and Tomorrow. Proceedings of a Conference at Gallaudet University*, J. E. Harkins and B. M. Virvan (Eds.). GRI Monograph Series B, No. 2. Washington D.C.: Gallaudet Research Institute, Gallaudet University.
- Studebaker, G. A., and I. Hochberg, (Eds.) (1993). *Acoustical Factors Affecting Hearing Aid Performance*, Second Edition. Needham Heights, Mass.: Allyn and Bacon.
- Studebaker, G. A., F. H. Bess, and L. B. Beck (Eds.) (1991). *The Vanderbilt Hearing Aid Report II*. Parkton, Md.: York Press.
- Sumby, W. H., and I. Pollack (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, 26:212-215.
- Swiffin, A. L., J. Arnott, J. Pickering, and A. Newell (1987). Adaptive and predictive techniques in a communication prosthesis. *Augment. Altern. Commun.*, 3:181-191.
- Uchanski, R. M., L. A. Delhorne, A. K. Dix, L. D. Braida, C. M. Reed, and N. I. Durlach (1994). Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech. *J. Rehabil. Res. Dev.*, 31(1):20-41.
- Upton, H. (1968). Wearable eyeglass speechreading aid. *Am. Ann. Deaf*, 113:222-229.
- Van Tasell, D. J., S. Y. Larsen, and D. A. Fabry (1988). Effects of an adaptive filter hearing aid on speech recognition in noise by hearing-impaired subjects. *Ear Hear.*, 9:15-21.
- Watson, C. S., D. Reed, D. Kewley-Port, and D. Maki (1989). The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality. *J. Speech Hear. Res.*, 32:245-251.
- Weiss, M. (1993). Effects of noise and noise reduction processing on the operation of the Nucleus-22 cochlear implant processor. *J. Rehabil. Res. Dev.*, 30(1):117-128.
- Weiss, M., and E. Aschkenasy (1981). Wideband Speech Enhancement, Final Technical Report RADC-TR-81-53. Griffiss Air Force Base, N.Y.: Rome Air Development Center, Air Force Systems Command.

- Wilson, B. S., C. C. Finley, D. T. Lawson, R. D. Wolford, and M. Zerbi (1993). Design and evaluation of a continuous interleaved sampling (CIS) processing strategy for multichannel cochlear implants. *J. Rehabil. Res. Dev.*, 30(1):110-116.
- Wise, R., and R. Olson (1993). What computerized speech can add to remedial reading. In *Behavioral Aspects of Speech Technology*, R. W. Bennett, A. K. Syrdal, and S. L. Greenspan (Eds.). Amsterdam: Elsevier Science Publishing Co.
- World Health Organization (1980). *The International Classification of Impairments, Disabilities, and Handicaps*. Geneva: World Health Organization.
- Yamada, Y., N. Murata, and T. Oka (1988). A new speech training system for profoundly deaf children. *J. Acoust. Soc. Am.*, 84(1):43.



## **APPLICATIONS OF VOICE- PROCESSING TECHNOLOGY II**



# Commercial Applications of Speech Interface Technology: An Industry at the Threshold

*John A. Oberteuffer*

## SUMMARY

Speech interface technology, which includes automatic speech recognition, synthetic speech, and natural language processing, is beginning to have a significant impact on business and personal computer use. Today, powerful and inexpensive microprocessors and improved algorithms are driving commercial applications in computer command, consumer, data entry, speech-to-text, telephone, and voice verification. Robust speaker-independent recognition systems for command and navigation in personal computers are now available; telephone-based transaction and database inquiry systems using both speech synthesis and recognition are coming into use. Large-vocabulary speech interface systems for document creation and read-aloud proofing are expanding beyond niche markets. Today's applications represent a small preview of a rich future for speech interface technology that will eventually replace keyboards with microphones and loudspeakers to give easy accessibility to increasingly intelligent machines.

## INTRODUCTION

Speech interface technology, which encompasses automatic speech recognition, synthesized speech, and natural language processing,

comprises the areas of knowledge required for human-machine communication by voice. This paper discusses commercial applications, which are beginning to have significant impacts on business and personal use. Commercial applications of speech interface technology, which first appeared in the early 1980s, are poised now in the early 1990s at a threshold of widespread practical application. Today's applications in speech interface technology utilize speech recognition or synthesis to simply translate spoken words into commands and text or vice versa with little regard to underlying meaning. In the future as applications for human-machine communication by voice grow, the need for natural-language-processing technology to permit speech interpretation will increase. The applications and developments described below represent some very important first steps into a future that will include systems capable of understanding natural conversational speech for transcription or spoken real-time translation. Today's applications are an important bridge to that future and represent the early and productive uses of speech interface technology.

Automatic speech recognition is the ability of machines to interpret speech in order to carry out commands or generate text. An important related area is automatic speaker recognition, which is the ability of machines to identify individuals based on the characteristics of their voices. Synthetic speech, or synonymously text-to-speech, is audible speech generated by machines from standard computer-stored text. These disciplines are closely related because they both involve an analysis and understanding of human speech production and perception mechanisms. In particular, the analysis of speech into its individual components (phones) and the characterization of the acoustic waveforms of these components are common to both disciplines. Speech recognition and speech synthesis are also closely coupled at the applications level—for example, for remote database access where visual displays are not available. The use of speech recognition for input and synthetic speech for output is a powerful combination that can transform any telephone into a fully intelligent node in a computer network.

## BACKGROUND

Automatic speech recognition and text-to-speech technologies have been under development since the early days of modern electronic and computer technology in the middle part of this century. A phonemic-based text-to-speech system was demonstrated at the World's Fair in 1939 by AT&T Bell Laboratories; high-speed computers in the

early 1950s made the display of speech spectrograms and the application of pattern recognition techniques for automatic speech recognition practical. By the early 1980s, automatic speech recognition had progressed enough to make practical speech-driven data entry systems. Voice input computers are used in many industrial inspection applications where hands and eyes are busy with other tasks, allowing data to be input directly into a computer without keyboarding or transcription. In the late 1970s, a combination of optical character recognition and synthetic speech made possible the first reading machine for the blind. Although bulky and expensive initially, this device allowed blind people, for the first time, to access arbitrary text with no human intervention.

The development of automatic speech recognition and text-to-speech systems has been carried out by large and small companies and by universities. In the United States, the Defense Advanced Research Projects Agency (now the Advanced Research Projects Agency) has provided significant funding and encouragement to a number of important university and private developers of this technology. In addition, large companies such as IBM, AT&T, and Texas Instruments have provided major research and development funding for advanced speech technologies.

Beginning in about 1990 the combination of powerful inexpensive microprocessors and improved algorithms for decoding speech patterns made possible voice command systems for personal computers and telephone-based systems. More expensive, but very powerful, large-vocabulary systems for the creation of text entirely by voice also have become available. Inexpensive chip-based text-to-speech systems allow talking dictionaries and word translators to be sold as consumer products, while more expensive systems provide concept to speech from powerful databases for telephone access by remote users. With these new platforms and analytical tools available, the number of commercial PC-based applications increased significantly in the early 1990s. The success of these recent applications highlights the need for increased sensitivity to human factors in the implementation of speech technologies. Many early commercial systems were offered without much understanding of the difficulty of implementing these semihuman technologies. Recently, speech interface technology systems have sought to overcome the natural limitations of speech understanding with more user-friendly interfaces.

## TECHNOLOGY

Although automatic speech recognition and synthetic speech technology often require the use of significant computing hardware resources, the technology is essentially software-based. Digital signal processors are used by many systems, but some speech systems utilize only analog/digital converters and general-purpose computing hardware. Low-end systems may run on single chips; mid-range systems are generally PC based with digital-to-analog and analog-to-digital conversion of speech signals carried out on a separate plug-in card. In some systems this card also includes a digital signal processor for speech analysis or synthesis. In most automatic speech recognition systems the pattern search and matching algorithms run on the main microprocessor.

## THE ADVANCED SPEECH TECHNOLOGY MARKET

The automatic speech recognition market can be organized into six major segments as shown in [Table 1](#). In the early 1990s significant growth in new applications has occurred in three of these segments: computer control, speech-to-text, and telephone. In the computer control segment, a number of small and large companies introduced speech input/output products for a few hundred dollars. These products, with both automatic speech recognition and text-to-speech capability, are bundled with popular sound boards for PCs. They provide automatic speech recognition of 1000 to 2000 words for controlling the Windows® interface. Their synthetic speech capability can be used for reading text from the screen in word-processing or spreadsheet programs. The initial paucity of software applications, which could take full advantage of this speech capability, limits the popularity of these systems. This is not unlike the early days of the computer mouse

TABLE 1 Automatic Speech Recognition Market Segments

Segments	Applications
Computer Control	Disabled, CAD
Consumer	Appliances, toys
Data Entry	QA inspection, sorting
Speech-to-Text	Text generation
Telephone	Operator services, IVR
Voice Verification	Physical entry, network access

when only a few DOS applications allowed point-and-click functions and none included sophisticated operations like drag-and-drop, which the current Windows interface for PC computers allows with mouse-pointing devices.

TABLE 2 Synthetic Speech Market Segments

Segments	Applications
Assistive technology	Reading machines, voice output aids
Consumer	Pocket translator, games
Education	Talking word processors
Telephone	Database access
Voice Interface	Data entry confirmation, task instructions

In the speech-to-text segment of the market, advances in large-vocabulary systems from several companies have been introduced. Systems for voice generation of text with vocabularies of 7000 words for memo and letter generation are available for less than \$3000. Medical report generation systems with vocabularies of 50,000 words, selling for over \$20,000 each, are finding acceptance in many hospital emergency rooms. In the telephone market there are a growing number of important applications, including operator services and interactive voice response, that use automatic speech recognition. Other applications in information services use both synthetic speech and automatic speech capability for database access. Through these telephone applications, the technology is now being introduced to a very broad market and is being recognized as essential for automating human telecommunications interfaces.

The synthetic speech market is segmented into five areas, as shown in [Table 2](#). The telephone segment will see the greatest growth and the most new applications in the near term. In addition, a number of profitable applications exist for consumer products. As more advanced telephone information services are offered, the need for text-to-speech to replace or complement digitally recorded speech will grow. Talking translators and talking dictionaries with very large vocabularies are now available for a few hundred dollars.

### RECENT MARKET TRENDS

Some significant trends are apparent in the speech interface technology market. Overall, the growth of this market is being driven by

the availability of new products. But other factors, such as cost, integration and acceptance by users, vendors, and applications developers, are important determinants in the penetration of a potential market. Consolidation for the companies involved in this industry is a future trend. A number of still small companies have been in the business of speech recognition and synthetic speech since the early 1980s. Many are currently seeking major partners for business arrangements ranging from mergers to joint marketing agreements. Market expansion also is occurring: a number of large computer and telecommunications companies that have been actively involved in advanced speech technology research for a number of years are just now beginning to offer products and technology licenses commercially. Most speech interface technology vendors are working with users to understand the important human factors issues that affect product use and application. This market demands an understanding of the complexity of speech technology as well as a willingness to provide significant customer support in designing user-friendly interfaces. Overcoming attitudinal barriers is common to any new technology, but it is especially important in this human-like speech technology whose quality, while impressive in mid-1993, is still far below the level of speech understanding and pronunciation ability of even modestly educated people.

## **MARKET SIZE**

The growth of the advanced speech technology commercial market is represented in [Figure 1](#), which shows the estimated market value from 1988 through 1997. The automatic speech recognition market will grow to \$345 million by 1994 from \$92 million in 1990, showing an average compound growth rate of about 40 percent. The synthetic speech market is smaller, growing to \$148 million in 1994 from \$51 million in 1988, an approximately 30 percent compound annual growth rate. A total market for 1994 of about a half billion dollars is projected with the market by the end of the decade for speech interface technologies estimated to be above \$2 billion. These numbers represent sales of commercial products but do not include research and development dollars invested in advanced speech technologies each year by small and large companies or government programs.

## **RECENT SIGNIFICANT COMMERCIAL DEVELOPMENTS**

Developments in several segments of the automatic speech recognition market reflect trends in speech interface product growth. In

the consumer market, a voice-controlled VCR programmer/TV remote has been introduced. This well-thought-out product prompts for spoken commands like time and date to set various VCR and TV functions. It provides an easy alternative to punching in a series of instructions on a small keyboard in a darkened room. This speaker-dependent device is based on a low-cost chip. Inexpensive digital signal-processing chips for speech and other signal-computing applications have been introduced by several developers and will have a significant impact on the application of speech interface technologies in price-sensitive mass market applications. At the other end of the technology spectrum, several vendors are offering systems with vocabularies up to 20,000 to 50,000 words that recognize discrete words and permit the creation of text entirely by voice.

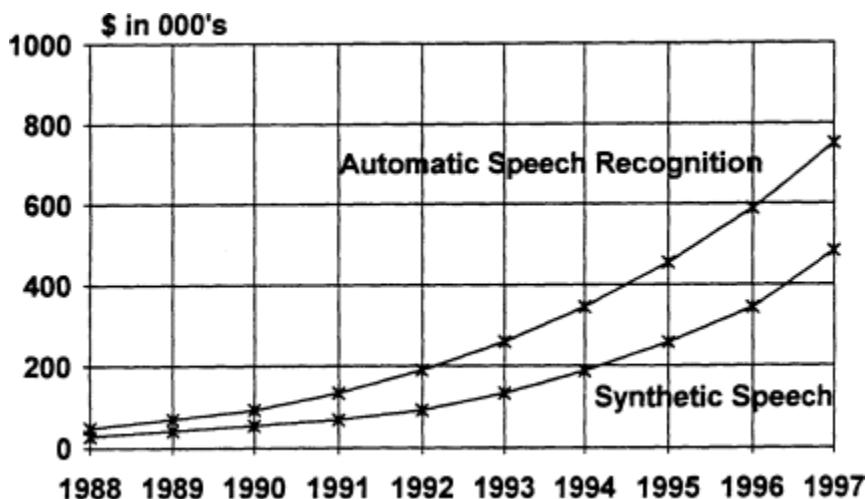


FIGURE 1 End-user market revenue for advanced speech technologies. (From VIA Information Associates.)

Major players in the personal computer market are introducing new speech technology products. Several major vendors are offering plug-in sound boards with speech recognition and synthetic speech capabilities for a few hundred dollars. These products with vocabularies of a few hundred to a few thousand words provide the ability to navigate around various Windows menus and applications by voice.

In the telecommunications segment, speech interface technologies are automating many long-distance toll calls. The driving force for automation and the use of automatic speech recognition in tele-

communications is productivity. Live operator time is estimated at \$17 million per second nationwide, so even partial automation of operator-related tasks can yield major benefits to telephone companies. Most of the regional Bell operating companies have pilot tested or introduced voice dialing and/or voice access for various services.

Advanced information services have been demonstrated by several companies. One of these services provides automated stock quotations using speaker-independent speech recognition for inquiry input and synthetic speech for output information. Callers to the service ask for one of three stock exchanges, New York, Toronto, or NASDAQ, and then ask for any one of 2000 companies on each stock exchange by name. Within a few seconds, up-to-date stock quotation information is provided by a synthetic speech system. This system represents a powerful marriage of automatic speech recognition in text-to-speech for totally automated remote database access.

An equally exciting application using text-to-speech is a system developed for the American Automobile Association in Orlando, Florida. This system allows a caller to receive driving instructions over the telephone via synthetic speech. Users can call a special number and then Touch-Tone in telephone numbers representing a starting point and a destination point in Orlando. The system uses the two telephone numbers to determine physical locations that are correlated with map information about the city of Orlando. The system calculates the best route from the starting point to the end point, taking into account both the shortest distance and the use of major highways. Having determined the route, the system then generates the text for the driving instructions, which can be faxed to the inquirer or provided by synthetic speech. The information can be accessed using a car telephone while actually driving the route. The driving directions provide real-time information to the driver. The feasibility of this service and cost would be unimaginable if live operators were involved. Even an automated system using prerecorded speech would be impossible since the combination of possible driving instructions within the city of Orlando is astronomical. It should be noted that the driving instructions that are provided do not even exist before the call is made but are created in response to a request. The instructions, which are generated by an artificial intelligence system, provide not just text-to-speech but also concept-to-speech in a system that is a powerful demonstration of computer telephone integration.

## FUTURE APPLICATIONS

In the near term, consumer products, voice input/output-capable hardware for PCs, telephone applications, and large-vocabulary text generation systems will dominate developments in speech interface technology. The decreasing cost of hardware will impact both low-end and high-end applications. At the low end, this hardware will make possible consumer applications with speech input and output for appliance control and instructions for use. At the high end, the decrease in cost of ever more powerful platforms for personal computers will make very large-vocabulary systems both less expensive and more capable. Intel's Pentium chip will provide high-end, large-vocabulary, automatic speech recognition systems with sufficient power to provide real-time, continuous speech, speaker-independent systems for text generation. As more speech recognition in text-to-speech systems become widely available for personal computers, more applications software will be written to take particular advantage of these input/output modalities. As the software becomes available and its utility is discovered by personal and business users, the demand for speech systems will increase, allowing further reductions in cost.

Before the end of the century, speech recognition and text-to-speech systems will be applied to hand-held computers. The speech interface is ideally suited to these devices because of its small space requirements and low cost. Speech input provides far more capability than pen input; speech output is competitive with small screen displays in many applications. New lap-top computers and personal digital assistants, with voice input for recording and voice annotation, will incorporate voice-powered navigation systems.

Voice dialing of telephones, whose introduction has begun on a modest scale, will become widely used in the United States, particularly with car phones and in many other applications. Limited network-based speech recognition systems have been introduced recently both in trial and actual pilot applications. In some areas it is possible to voice dial from any cellular phone using either numbers that are recognized on a speaker-independent basis or speed dialing from a personal speaker-dependent directory. This system allows voice dialing and speed dialing by voice for easy hands-free/eyes-free telephone use in a variety of situations. Telephone services using speech input and output will continue to increase. Automated directory assistance has begun in Canada using speech recognition technology.

This system allows users to complete many directory assistance calls without live operator assistance by recognizing city names and the names of common businesses.

An application that exemplifies speech interface technologies and that may find significant application in the future is the voice-controlled automated attendant. One advanced speech recognition developer is currently demonstrating such a system in its offices, which have approximately 2000 employees. The system makes it possible to ask for any one of these employees by name. The automatic speech recognition system will recognize the person's name, speak it back with the appropriate extension number using synthetic speech, and route the call. This system is fast and accurate and may be used as a telephone directory. This implementation represents another powerful integration of automatic speech recognition and synthetic speech in a system representative of those that will become commonplace by the turn of the century.

# Military and Government Applications of Human-Machine Communication by Voice\*

*Clifford J. Weinstein*

## SUMMARY

This paper describes a range of opportunities for military and government applications of human-machine communication by voice, based on visits and contacts with numerous user organizations in the United States. The applications include some that appear to be feasible by careful integration of current state-of-the-art technology and others that will require a varying mix of advances in speech technology and in integration of the technology into applications environments. Applications that are described include (1) speech recognition and synthesis for mobile command and control; (2) speech processing for a portable multifunction soldier's computer; (3) speech- and language-based technology for naval combat team tactical training; (4) speech technology for command and control on a carrier flight deck; (5) control of auxiliary systems, and alert and warning generation, in fighter aircraft and helicopters; and (6) voice check-in, report entry, and communication for law enforcement agents or special forces. A phased approach for transfer of the technology into applications is advocated, where integration of applications systems is pursued in parallel with advanced research to meet future needs.

---

\* This work was sponsored in part by the Advanced Research Projects Agency and in part by the Department of the Air Force. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government.

## INTRODUCTION

This paper describes a broad range of opportunities for military and government applications of human-machine communication by voice and discusses issues to be addressed in bringing the technology into real applications. The paper draws on many visits and contacts by the author with personnel at a variety of current and potential user organizations in the United States. The paper focuses on opportunities and on what is needed to develop real applications, because, despite the many opportunities that were identified and the high user interest, the military and government organizations contacted were generally not using human-machine communication by voice in operational systems (exceptions included an application in air traffic controller training and voice entry of zip codes by the U.S. Postal Service). Furthermore, the visits and discussions clearly identified a number of applications that today's state-of-the-art technology could support, as well as other applications that require major research advances.

Background for this paper is provided by a number of previous assessments of military applications of speech technology (Beek et al., 1977; Cupples and Beek, 1990; Flanagan et al., 1984; Makhoul et al., 1989; *Proceedings of the NATO AGARD Lecture Series*, 1990; Woodard and Cupples, 1983; Weinstein, 1991), including prior National Research Council studies (Flanagan et al., 1984; Makhoul et al., 1989) and studies conducted in association with the NATO RSG10 Speech Research Study Group (Beek et al., 1977; Cupples and Beek, 1990; *Proceedings of the NATO AGARD Lecture Series*, 1990; Weinstein, 1991). Those prior studies provide reviews of the state of the art at the time, and each outlines a number of programs in which prototype speech recognition systems were tested in application environments, including fighter aircraft, helicopters, and ship-based command centers. These efforts, as described in the references but not detailed further here, generally yielded promising technical results but have not yet been followed by operational applications. This paper focuses on users and applications in the United States, but the general trends and conclusions could apply elsewhere as well.

This paper is organized to combine reports on the military and government visits and contacts with descriptions of target applications most closely related to each organization. However, it is important to note that many of the applications pertain to a number of user organizations, as well as having dual use in the civilian and commercial areas. (Other papers in this volume describe applications of speech technology in general consumer products, telecom

munications, and aids for people with physical and sensory disabilities.) A summary relating the classes of applications to the interests of the various military and government users is provided near the end of the paper. The paper concludes with an outline of a strategy for technology transfer to bring the technology into real applications.

## TECHNOLOGY TRENDS AND NEEDS

A thorough discussion of technology trends and needs would be beyond the scope of this paper; hence, the focus here is on description of the applications. But the underlying premise is that both the performance of algorithms and the capability to implement them in real-time, in off-the-shelf or compact hardware, has advanced greatly beyond what was tested in prior prototype applications. The papers and demonstrations at a recent DARPA (Defense Advanced Research Projects Agency) Speech and Natural Language Workshops (1992) provide a good representation of the state of current technology for human-machine communication by voice. Updated overviews of the state of the art in speech recognition technology are presented elsewhere in this volume.

With respect to technological needs, military applications often place higher demand on robustness to acoustic noise and user stress than do civilian applications (Weinstein, 1991). But military applications can often be carried out in constrained task domains, where, for example, the vocabulary and grammar for speech recognition can be limited.

## SUMMARY OF VISITS AND CONTACTS

The broad range of military and government organizations that were contacted is shown in [Figure 1](#). There was broad-based interest in speech recognition technology across all these organizations. The range of interests was also deep, in the sense that most organizations were interested in applications over a range of technical difficulties, including some applications that today's state-of-the-art technology could support and others that would require major research advances. Also, many organizations had tested speech recognition systems in prototype applications but had not integrated them into operational systems. This was generally due to a perception that "the technology wasn't ready yet." But major speech recognition tests, such as the Air Force's F-16 fighter tests (Howard, 1987) and the Army's helicopter tests (Holden, 1989) were conducted a number of years ago. In general, tests such as these have not been performed with systems

that approach today's state-of-the-art recognition technology (Weinstein, 1991).

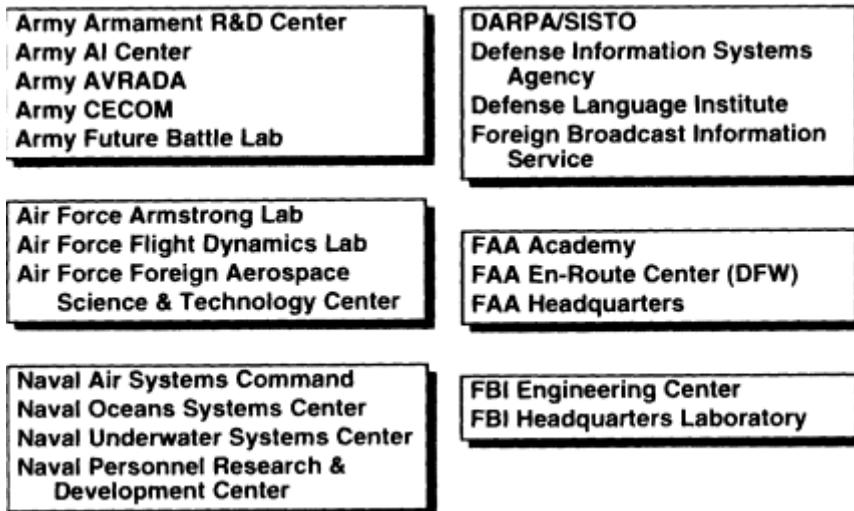


FIGURE 1 Summary of military and government organizations visited or contacted.

## ARMY APPLICATIONS

The Army visits and contacts (see [Figure 1](#)) pointed out many applications of human-machine communication by voice, of which three will be highlighted here: (1) Command and Control on the Move (C20TM); (2) the Soldier's Computer; and (3) voice control of radios and other auxiliary systems in Army helicopters. In fact, the applications for voice-actuated user interfaces are recognized by the Army to pervade its engineering development programs (E. Mettala, DARPA, unpublished presentation, Feb. 1992).

In Desert Storm the allied troops moved farther and faster, than troops in any other war in history, and extraordinary efforts were needed to make command and control resources keep pace with the troops. C20TM is an Army program aimed at ensuring the mobility of command and control for potential future needs. [Figure 2](#) illustrates some of the mobile force elements requiring C20TM and some of the potential applications for speech-based systems. Typing is often a very poor input medium for mobile users, whose eyes and hands are busy with pressing tasks. Referring to [Figure 2](#), a foot

soldier acting as a forward observer could use speech recognition to enter a stylized report that would be transmitted to command and control headquarters over a very low-rate, jam-resistant channel. Repair and maintenance in the field can be facilitated by voice access to repair information and helmet-mounted displays to show the information. In a mobile command and control vehicle, commanders need convenient access to battlefield information and convenient means for entering and updating plans. Integrated multimodal input/output (voice, text, pen, pointing, graphics) will facilitate meeting these requirements. Other applications suggested in [Figure 2](#) include simple voice translation (e.g., of forward observer reports), access to battlefield situation information, and weapons system selection.

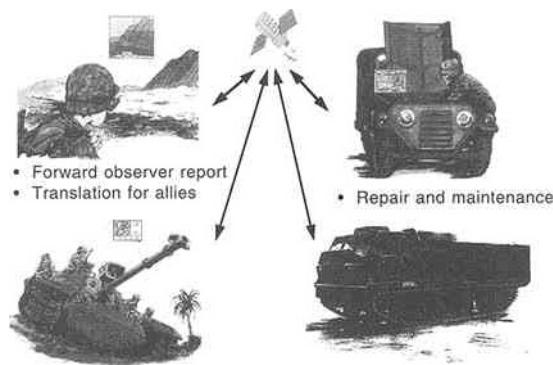


FIGURE 2 Command and Control on the Move (C20TM): force elements and example applications of human-machine communication by voice.

The Soldier's Computer is an Army Communications and Electronics Command (CECOM) program that responds to the information needs of the modern soldier. The overall system concept is shown in [Figure 3](#). Voice will be a crucial input mode, since carrying and using a keyboard would be very inconvenient for the foot sol

dier. Functions of the Soldier's Computer are similar to those mentioned above for C20TM. Technical issues include robust speech recognition in noise and smooth integration of the various input/output modes. The technology for both the Soldier's Computer and C20TM has many dual-use, peacetime applications, both for everyday use and in crises such as fires or earthquakes.

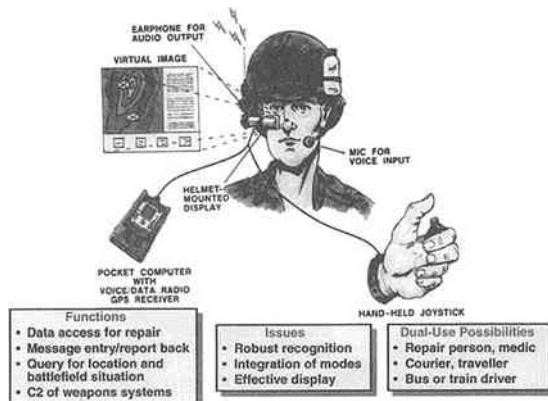


FIGURE 3 The Soldier's Computer: system concept; functions that would be assisted by human-machine communication by voice; technical issues; and possible dual-use applications.

Speech recognition for control of radios and other devices in Army helicopters is an application that has been addressed in test and evaluation programs by the Army Avionic Research and Development Activity (AVRADA) organization, as well as by groups in the United Kingdom and France. Feasibility has been demonstrated, but operational use has not been established. The Army AVRADA people I met described a tragic helicopter collision in which the fact that both pilots were tuning radios may have been the major cause of the crash. Although voice control was considered to be a viable solution, it was not established as a requirement (and therefore not implemented) because of the Army's view that speaker-independent recognition was necessary and was not yet sufficiently robust. But the state of

the art of speaker-independent recognition, particularly for small vocabularies, has advanced a great deal and is now likely to be capable of meeting the needs for control of radios and similar equipment in a military helicopter.

## NAVY APPLICATIONS

My Navy visits and contacts uncovered a wide range of important applications of speech technology, with support at very high levels in the Navy. Applications outlined here will be (1) aircraft carrier flight deck control and information management, (2) SONAR supervisor command and control, and (3) combat team tactical training.

The goal in the carrier flight deck control application is to provide speech recognition for updates to aircraft launch, recovery weapon status, and maintenance information. At the request of Vice-Admiral Jerry O. Tuttle (Assistant Chief of Operations for Space and Electronic Warfare), the Naval Oceans Systems (NOSC)<sup>1</sup> undertook to develop a demonstration system on board the *USS Ranger*. Recognition requirements included open microphone; robust, noise-resistant recognition with out-of-vocabulary word rejections; and easy integration into the PC-based onboard system. An extremely successful laboratory demonstration, using a commercially available recognizer, was performed at NOSC for Admiral Tuttle in November 1991. Subsequent tests on board the Ranger in February 1992 identified a number of problems and needed enhancements in the overall human-machine interface systems, but correction of these problems seemed to be well within the current state of the art.

The SONAR supervisor on board a surface ship needs to control displays, direct resources, and send messages while moving about the command center and looking at command and control displays. This situation creates an opportunity for application of human-machine communication by voice, and the Naval Underwater Systems Center (NUSC) has sponsored development of a system demonstrating voice activation of command and control displays at a land-based integrated test site in New London, Connecticut. The system would be used first for training of SONAR supervisors at the test site and later for shipboard applications. Initial tests with ex-supervisors from

---

<sup>1</sup> The Naval Oceans Systems Center has subsequently reorganized as the Naval Research and Development Organization.

SONAR were promising, but the supervisors expressed dissatisfaction at having to train the speaker-dependent recognizer that was used.

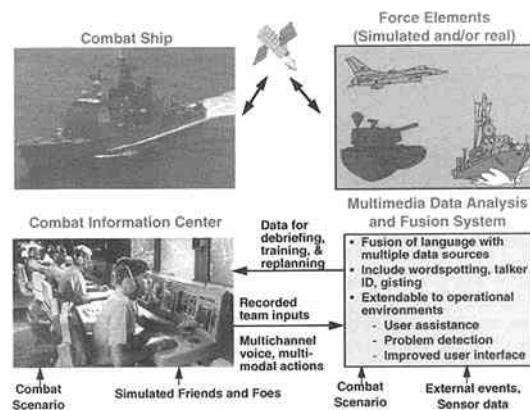


FIGURE 4 Naval combat team tactical training: system concept and applications of speech- and language-based technology.

A scenario for application of speech-and-language-based technology to Navy combat team tactical training, based on a proposal by the Navy Personnel Research and Development Center, is illustrated in Figure 4. The training scenario includes a mix of real forces and force elements simulated by using advanced simulation technology. Personnel in the Combat Information Center (either at sea or in a land-based test environment) must respond to developing combat scenarios using voice, typing, trackballs, and other modes and must communicate both with machines and with each other. As suggested in the figure, speech-based and language-based technology, and fusion of language with multiple data sources, can be used to correlate and analyze the data from a combat training exercise, to allow rapid feedback (e.g., what went wrong?) for debriefing, training, and replanning. These language-based technologies, first developed and applied in training applications where risk is not a major issue, can

later be extended to operational applications, including detection of problems and alerting users, and also to development of improved human-machine interfaces in the Combat Information Center.

The approach of first developing and using a system with human-machine communication by voice in a training application and then extending to an operational application is a very important general theme. The training application is both useful in itself and provides essential data (including, for example, language models and speech data characterizing the human-machine interaction) for developing a successful operational application.

## AIR FORCE APPLICATIONS

The Air Force continues its long-term interest in speech input/output for the cockpit and has proposed to include human-machine communication by voice in the future Multi-Role Fighter. Fighter cockpit applications, ranging from voice control of radio frequency settings to an intelligent Pilot's Associate system, have been discussed elsewhere (Weinstein, 1991; Howard, 1987) and will not be detailed further here. However, it is likely that the kinds of applications that were tested in the AFTI F-16 Program, with promising results but not complete success, would be much more successful with today's robust speech recognition technology. Voice control of radio frequencies, displays, and gauges could have significant effect on mission effectiveness and safety. A somewhat more advanced but technically feasible application is use of voice recognition in entering reconnaissance reports. Such a system is currently under development at the Defense Research Agency in the United Kingdom (Russell et al., 1990). Other potential Air Force applications include human-machine voice communication in airborne command posts, similar to Army and Navy command and control applications. In particular, entry of data and log information by voice could potentially provide significant workload reduction in a large variety of command and control center operations.

## AIR TRAFFIC CONTROL APPLICATIONS

The air traffic controller is taught to use constrained phraseology to communicate with pilots. This provides an opportunity, which is currently being exploited at the Federal Aviation Administration (FAA) Academy in Oklahoma City, at a Naval Air Technical Training Center in Orlando, Florida, and elsewhere, to apply speech recognition and synthesis to emulate pseudo-pilots in the training of air traffic

controllers. This application, illustrated in [Figure 5](#), is an excellent example of military and government application of human-machine communication by voice that is currently in regular use. Advances in speech and language technology will extend the range and effectiveness of these training applications (Weinstein, 1991). As in the Naval Combat Team Tactical training application, speech recognition technology and data fusion could be used to automate training session analysis and to provide rapid feedback to trainees.

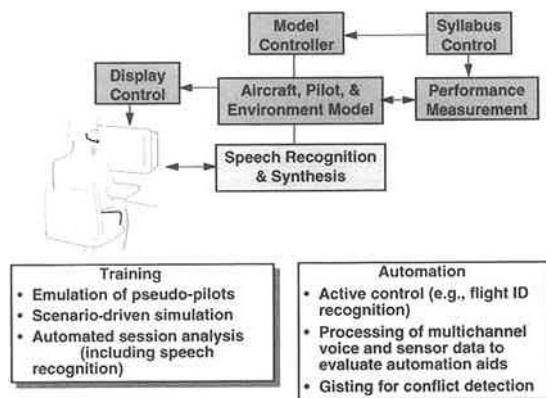


FIGURE 5 Air traffic control training and automation: system concept and applications of human-machine communication by voice.

A number of automation aid applications in air traffic control are also possible via speech technology, as indicated in [Figure 5](#). Again, the experience can be used to help build operational automation applications. An application of high current interest is on-line recognition of flight identification information from a controller's speech to quickly access information on that flight (Austin et al., 1992). More advanced potential applications include processing and fusion of multimodal data to evaluate the effectiveness of new automation aids for air traffic control and gisting (Rohlicek et al., 1992) of pilot/controller communications to detect potential air space conflicts.

## LAW ENFORCEMENT APPLICATIONS

Discussions with Federal Bureau of Investigation (FBI) personnel revealed numerous potential applications of speech and language technology in criminal investigations and law enforcement. For example, the Agent's Computer is envisioned as a portable device, with some similarity to the Soldier's Computer but specialized to the agent's needs. Functions of particular interest to agents include (1) voice check-in, (2) data or report entry, (3) rapid access to license plate or description-based data, (4) covert communication, (5) rapid access to map and direction information, and (6) simple translation of words or phrases. Fast access to and fusion of multimedia data, some language based and some image based (e.g., fingerprints and photos), together were a major need for aid in investigations. Voice-controlled database access could be used to facilitate this data access. As with the Navy and FAA training applications mentioned above, the FBI had high interest in training using simulation in combination with language-based technology for both mission execution and mission diagnosis. Criminal investigations put a major burden on agents in terms of reporting and documentation; the use of human-machine communication by voice to rapidly prepare reports ranging from structured forms to free text, was identified as an application of major interest to agents.

## SUMMARY OF USERS AND APPLICATIONS

The matrix shown in [Figure 6](#) relates the classes of applications that have been described to the interests of the various military and government users. All the applications have dual use in the civilian area. Looking across the rows, it is evident that all the users have interest in a wide range of applications with varying technical difficulty. In fact, upon showing this matrix to potential users, each user generally wanted to fill in all the boxes in his row. The most pervasive near-term application is voice data entry, which can range from entering numerical data to creating formatted military messages, to free-form report entry. The current speech recognition technology is capable of performing these functions usefully in a number of military environments, including particularly to provide operator workload reduction in command and control centers.

## TECHNOLOGY TRANSFER

A key conclusion of this study is that there is now a great opportunity for military and government applications of human-machine

communication by voice, which will have real impact both on the users and on the development of the technology. This opportunity is due both to technical advances and to very high user interest; there has been a big increase in user interest just within the past few years (i.e., since the study reported in Weinstein, 1991).

Users	Data Entry & Commun.	Data Access	Command & Control	Training	Translation
Soldier	XX	X	X	X	X
Naval CIC Officer	XX	XX	XX	XX	
Pilot	XX	X	XX		
Agent	XX	XX		X	X
Air Traffic Controller	X	XX		X	
Diplomat	X	X			XX
Joint Force Commander		XX	XX		XX

XX = primary application  
 X = additional application

FIGURE 6 Matrix relating classes of applications of human-machine communication by voice to the interests of military and government users.

The strategy of the technologists should be to select and push applications with a range of technical challenges, so that meaningful results can be shown soon, while researchers continue to advance the technology to address the harder problems. It is essential that technologists work with the users to narrow the gap between the user and the state of the art. Too often, users have tested speech recognition systems that are off the shelf but well behind the state of the art, and have been discouraged by the results.

With today's software-based recognition technology, and with the increased computing power in PCs, workstations, and digital signal-processing chips, it is now possible to develop and test applications with recognition algorithms that run in real-time, in software, on commercially available general-purpose processors and that perform very close to the state of the art. Technologists must work with users

to understand the user requirements and provide appropriate technology. For effective technology transfer, software and hardware must be portable and adaptable to new domains or to unforeseen variations in the user's needs. Eventually the user should be able to take over and continue adapting the technology to the changing needs, with little support from the technologists. Meanwhile, the technologists, having learned from each generation of operational applications, can be working to develop the research advances that will enable the next generation of operational applications.

### ACKNOWLEDGMENTS

I would like to acknowledge the contributions to this study of the following individuals: Victor Zue (MIT), Allen Sears (MITRE), Janet Baker (Dragon Systems), Charles Wayne (DARPA), Erik Mettala (DARPA), George Doddington (DARPA), Deborah Dahl (Paramax), David Ruppe (Army), Jim Schoening (Army), Christine Dean (Navy), Steve Nunn (Navy), Walter Rudolph (Navy), Jim Cupples (Air Force), Tim Anderson (Air Force), Dave Williamson (Air Force), Joe Kielman (FBI), John Hoyt (FBI), and Peter Sielman (Analysis and Technology, Inc.). A special acknowledgment goes to Victor Zue for many helpful discussions and contributions.

### REFERENCES

- Austin, S., et al. 1992. BBN Real-Time Speech Recognition Demonstrations. In Proceeding of the February 1992 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, pp. 250-251.
- Beek, B., E. P. Neuburg, and D. C. Hodge. 1977. An Assessment of the Technology of Automatic Speech Recognition for Military Applications. IEEE Trans. Acoust., Speech, Signal Process., ASSP-25:310-321.
- Cupples, E. J., and B. Beek. 1990. Applications of Audio/Speech Recognition for Military Applications. In Proceedings of the NATO/AGARD Lecture Series No. 170, Speech Analysis and Synthesis and Man-Machine Speech Communications for Air Operations, pp. 8-1-8-10.
- Flanagan, J. L., et al. 1984. Automatic Speech Recognition in Severe Environments. National Research Council, Committee on Computerized Speech Recognition Technologies. Washington, D.C.: National Academy Press.
- Holden, J. M. 1989. Testing Voice in Helicopter Cockpits. In Proceedings of the American Voice Input/Output Society (AVIOS) Conference, Sept.
- Howard, J. A. 1987. Flight Testing of the AFTI/F-16 Voice Interactive Avionics System. In Proceedings of the Military Speech Tech 1987. Arlington, Va.: Media Dimensions, pp. 76-82.
- Makhoul, J., T. H. Crystal, D. M. Green, D. Hogan, R. J. McAulay, D. B. Pisoni, R. D. Sorkin, and T. G. Stockham, Jr. 1989. Removal of Noise from Noise-Degraded

- Speech Signals. National Research Council, Committee on Hearing, Bioacoustics, and Biomechanics. Washington, D.C.: National Academy Press.
- Proceedings of the February 1992 DARPA Speech and Natural Language Workshop. 1992. Morgan Kaufmann Publishers.
- Proceedings of the NATO AGARD Lecture Series No. 170, Speech Analysis and Synthesis and Man-Machine Speech Communications for Air Operations, 1990.
- Rohlicek, J. R., et al. 1992. Gisting Conversational Speech. In Proceedings of the ICASSP'92. San Francisco, pp. 11-113-11-116.
- Russell, M. J., et al. 1990. The ARM Continuous Speech Recognition System. In Proceedings of the ICASSP'90. Albuquerque, N.Mex., April.
- Weinstein, C. J. 1991. Opportunities for Advanced Speech Processing in Military Computer-Based Systems. Proc. IEEE, 79(11):1626-1641.
- Woodard, J. P., and E. J. Cupples. 1983. Selected Military Applications of Automatic Speech Recognition. IEEE Commun. Mag., 21(9):35-44.

## TECHNOLOGY DEPLOYMENT



# Deployment of Human-Machine Dialogue Systems

*David B. Roe*

## SUMMARY

The deployment of systems for human-to-machine communication by voice requires overcoming a variety of obstacles that affect the speech-processing technologies. Problems encountered in the field might include variation in speaking style, acoustic noise, ambiguity of language, or confusion on the part of the speaker. The diversity of these practical problems encountered in the "real-world" leads to the perceived gap between laboratory and "real-world" performance.

To answer the question "What applications can speech technology support today?" the concept of the "degree of difficulty" of an application is introduced. The degree of difficulty depends not only on the demands placed on the speech recognition and speech synthesis technologies but also on the expectations of the user of the system. Experience has shown that deployment of effective speech communication systems requires an iterative process. This paper discusses general deployment principles, which are illustrated by several examples of human-machine communication systems.

## INTRODUCTION

Speech-processing technology is now at the point at which people can engage in voice dialogues with machines, at least in limited ways.

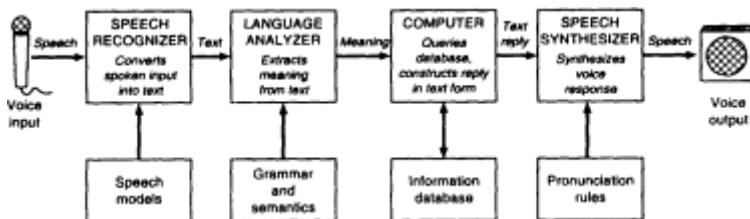


FIGURE 1 A human-machine dialogue system.

Simple voice communication with machines is now deployed in personal computers, in the automation of long-distance calls, and in voice dialing of mobile telephones. These systems have small vocabularies and strictly circumscribed task domains. In research laboratories there are advanced human-machine dialogue systems with vocabularies of thousands of words and intelligence to carry on a conversation on specific topics. Despite these successes, it is clear that the truly intelligent systems envisioned in science fiction are still far in the future, given the state of the art today.

Human-machine dialogue systems can be represented as a four-step process, as shown in Figure 1. This figure encompasses both the simple systems deployed today and the spoken language understanding we envision for the future. First, a speech recognizer transcribes sentences spoken by a person into written text (Makhoul and Schwartz, in this volume; Rabiner and Juang, 1993). Second, a language understanding module extracts the meaning from the text (Bates, in this volume; Moore, in this volume). Third, a computer (consisting of a processor and a database) performs some action based on the meaning of what was said. Fourth, the person receives feedback from the computer in the form of a voice created by a speech synthesizer (Allen, in this volume; Carlson, in this volume). The boundaries between these stages of a dialogue system may not be distinct in practice. For instance, language-understanding modules may have to cope with errors in the text from the speech recognizer, and the speech recognizer may make use of grammar and semantic constraints from the language module in order to reduce recognition errors.

In the 1993 "Colloquium on Human-Machine Communication by Voice," sponsored by the National Academy of Sciences (NAS), much of the discussion focused on practical difficulties in building and deploying systems for carrying on voice dialogues between humans and machines. Deployment of systems for human-to-machine communication by voice requires solutions to many types of problems

that affect the speech-processing technologies. Problems encountered in the field might include variation in speaking style, noise, ambiguity of language, or confusion on the part of the speaker. There was a consensus at the colloquium that a gap exists between performance in the laboratory and accuracy in the field, because conditions in real applications are more difficult. However, there was little agreement about the cause of this gap in performance or what to do about it.

A key point of discussion at the NAS colloquium concerned the factors that make a dialogue easy or difficult. Many such degrees of difficulty were mentioned in a qualitative way. To summarize the discussion in this paper it seems useful to introduce a more formal concept of the degree of difficulty of a human-machine dialogue and to list each dimension that contributes to the overall difficulty. The degree of difficulty is a useful concept, despite the fact that it is only a "fuzzy" (or qualitative) measure because of lack of precision in quantifying an overall degree of difficulty for an application.

A second point of discussion during the NAS colloquium concerned the process of deployment of human-machine dialogue systems. In several cases such systems were built and then modified substantially as the designers gained experience in what the technology could support or in user-interface issues. This paper elaborates on this iterative deployment process and contrasts it with the deployment process of more mature technologies.

### **DEGREE OF DIFFICULTY OF A VOICE DIALOGUE APPLICATION**

Whether a voice dialogue system is successful depends on the difficulty of each of the four steps in [Figure 1](#) for the particular application, as well as the technical capabilities of the computer system. There are several factors that can make each of these four steps difficult. Unfortunately, it is difficult to quantify precisely the difficulty of these factors. If the technology performs unsatisfactorily at any stage of processing of the speech dialogue, the entire dialogue will be unsatisfactory. We hope that technology will eventually improve to the point that there are no technical barriers whatsoever to complex speech-understanding systems. But until that time it is important to know what is easy, what is difficult but possible, and what is impossible, given today's technology.

What are the factors that determine the degree of difficulty of a voice dialogue system? In practice, there are several factors for each step of the voice dialogue that may make the task difficult or easy. Because these factors are qualitatively independent, they can be viewed

as independent variables in a multidimensional space. For a simple example of two dimensions of difficulty for speech recognition refer to [Figure 2](#). There are many dimensions of difficulty for a speech recognition system, two of which are shown. Eight applications are rated according to difficulty of speaking mode (vertical axis) and vocabulary size (horizontal axis). "Voice dictation" refers to commercial 30,000-word voice typewriters; "V.R.C.P." stands for voice recognition call processing, as a way to automate long-distance calling; "telephone number dialing" refers to connected digit recognition of phone numbers; "DARPA resource management" refers to the 991-word Naval Resource Management task with a constraining grammar; "A.T.I.S." stands for the DARPA (Defense Advanced Research Projects Agency) Air Travel Information System; and "natural spoken language" refers to conversational speech on any and every topic. Clearly, a task that is difficult in both the dimensions of vocabulary size and speaking style would be harder (and would have lower accuracy) than a small, isolated word recognizer, if all other factors are equal. The other factors are not equal, as discussed in the section "Dimensions of the Recognition Task." Note that telephone number dialing, despite its position on the two axes in this figure, is a difficult application because of dimensions not shown here, such as user tolerance of errors and grammar perplexity.

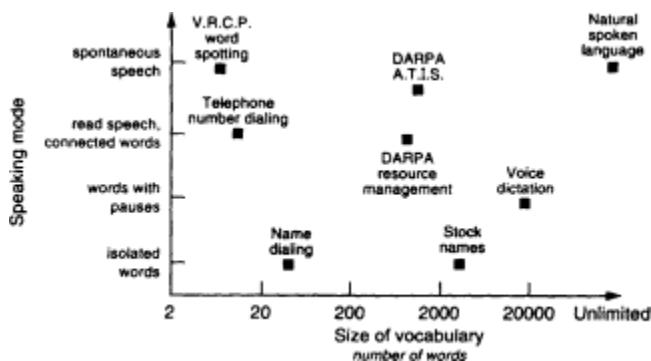


FIGURE 2 Two (of many) dimensions of difficulty for speech recognition.  
(Adapted from Atal, in this volume.)

Ideally, a potential human-machine dialogue could receive a numerical rating along each dimension of difficulty, and a cumulative degree of difficulty could be computed by summing the ratings along each separate dimension. Such a quantitative approach is overly sim

plistic. Nevertheless, it is a valuable exercise to evaluate potential applications qualitatively along each of the dimensions of difficulty.

The problems for voice dialogue systems can be separated into those of speech recognition, language understanding, and speech synthesis, as in [Figure 1](#). (For the database access stage, a conventional computer is adequate for most voice dialogue tasks. The data-processing capabilities of today's machines pose no barriers to development of human-machine communication systems.) Let us examine the steps of speech recognition, language understanding, and speech synthesis in order to analyze the specific factors, or dimensions of difficulty, that make an application easy or difficult.

### Dimensions of the Speech Recognition Task

Humans are able to understand speech so readily that they often fail to appreciate the difficulties that this task poses for machines. The exception may be the process of learning a foreign language. Indeed, there is more than a casual relationship between the problems faced by an adult listening to a foreign language and those of a machine recognizing speech.

The performance of speech recognizers is typically assessed by measuring the accuracy of the recognizer, or equivalently, its error rate. But the accuracy of any recognizer may vary widely, depending on the conditions of the experiment and the speech data. John Makhoul, in his paper in this volume, has listed some rules of thumb indicating how recognition accuracies vary. In the laboratory, speech recognizers are quite accurate in acoustic pattern matching. In real-world conditions, however, the error rate is much higher, due in part to the increased variability of speech styles encountered. Given high-quality, consistent speech samples recorded in a quiet laboratory, and given sufficient speech samples to fully train an HMM (hidden Markov model), accuracies are almost comparable to human accuracies in acoustic perception. For instance, numbers can be recognized with an error rate of less than one in 300 words (99.7 percent accuracy) (Gauvin and Lee, 1992). This result was obtained on the Texas Instruments/ National Institute of Standards and Technology speech database recorded under laboratory conditions in a soundproof booth and with a balance of dialects of native U.S. speakers. On a larger, speaker-independent task, the DARPA resource management task described below, word accuracies of about 96 percent can be achieved on a vocabulary of a thousand words (Marcus, 1992). Again, these results are obtained by using carefully spoken speech recorded in a quiet environment.

In applications the variability of speech and speaking environments is much greater, so that the same speech recognition algorithm will have error rates that are much higher than with well-controlled laboratory speech. For instance, in tests of speech recognition of credit card numbers spoken by merchants at retail stores, the error rate rises to about 2 percent per digit (from 0.3 percent) (Ramesh et al., 1992) with an algorithm similar to that described by Gauvin and Lee (1992). This increase in error rate is typical of the difference between laboratory-quality speech and the speech encountered in field conditions. In fact, speech recognition in the field is a harder task because the speech is more variable than laboratory speech.

The following are the dimensions of difficulty for speech recognition applications:

- *Speaker independence.* It is much easier to characterize an individual speaker's voice than to recognize all voice types and all dialects. Applications can be categorized, in increasing order of difficulty, as speaker-trained, speaker-adaptive, multispeaker, speaker-independent, and speaker-independent with nonnative speakers.
- *Expertise of the speaker.* People typically learn how to get good recognition results with practice. Applications in which the speakers can quickly gain experience are more likely to succeed than applications in which the majority of people use the service infrequently. As Richard Schwartz remarked at the NAS colloquium, "You are a first time user only once."
- *Vocabulary confusability.* Other things being equal, a larger vocabulary is more likely to contain confusable words or phrases that can lead to recognition errors. However, some small vocabularies may be highly confusable. The letters of the alphabet ("A, B, C, D . . .") are notoriously difficult to recognize.
- *Grammar perplexity.* An application may have a grammar in which only certain words are permitted given the preceding words in a sentence, which reduces the opportunity for errors. The perplexity of the grammar is the average number of choices at any point in the sentence.
- *Speaking mode: rate and coarticulation.* Speech sounds are strongly affected by surrounding sounds in rapid speech. Isolated words are more consistently recognized than words in fluent speech. Voice dictation systems typically require that speakers leave a slight pause between words in a sentence. Speaker variabilities, including disfluencies such as hesitation, filled pauses, and stammering, are also important. People are able to understand these normal variations in speed or loudness and to compensate for any involuntary changes caused by

stress upon the speaker. Only in the simplest cases can machines handle such conditions.

- *Channel conditions.* Speech that is distorted or obscured by noise is more difficult for machines to recognize than high-quality speech. Noise can include background speech and other acoustic noise as well as noise in the transmission channel. Variability in transmission bandwidth and in microphone characteristics also affects speech recognition accuracy.
- *User tolerance of errors.* It is important to bear in mind that voice dialogue systems, notwithstanding recent advances, remain error-prone. Given that any speech recognizer will make occasional errors, the inconvenience to the user should be minimized. This means that careful design of human factors of an application will be essential. The central questions when considering an application using a speech recognizer are: (1) What accuracy will the user of this service expect? (2) Is the speech recognizer accurate enough to meet the expectations of the user? (3) Does the benefit of using speech recognition in this application outweigh its cost, compared to alternative technologies?

Each of these dimensions of difficulty embodies some aspect of speech variability, which is the central problem of speech recognition. The more sophisticated the speech recognizer, the better it is able to cope with these practical difficulties. Increasing the robustness of speech recognizers to all types of variability is a major challenge of current speech recognition research. These sources of variability must be carefully considered when planning applications of the technology, because it is these robustness characteristics that determine whether a speech recognizer will be accurate enough to be satisfactory to the users.

### **Dimensions of the Language-Understanding Task**

The difficulties of natural language understanding are well known (Bates, in this volume; Berwick, 1987; Hirst, 1987) and will not be discussed in detail here. For speech-understanding systems the difficulties are compounded by uncertainty in interpreting the acoustic signal, which results in errors in the text (Moore, in this volume). Therefore, spoken-language-understanding systems are now limited to constrained domains in which there is little ambiguity. Furthermore, models used for speech dialogue systems tend to be simpler and less powerful than those used for text understanding. Though it is widely known that finite-state grammars are too simple to express the range of meanings of English, these elementary grammars are

typically used by speech-understanding systems. For instance, voice dictation systems by IBM (Bahl et al., 1989), Dragon Systems, and Kurzweil Applied Intelligence use simple N-gram models that estimate the probability of sequences of up to three words based on training texts. The dimensions of difficulty of language understanding are:

- *Grammar complexity and ambiguity.* The wider the range of meanings in the domain of understanding, the more complex the grammar must be to express those meanings. This complexity leads to a greater possibility of semantic ambiguity, that is, the chance that an input text sequence may have more than one possible interpretation. Finally, semantic or grammatical ambiguity may be compounded by acoustic ambiguity (words that sound similar) or speech recognition errors.
- *Language variability.* Language is very flexible. For any meaning there are many ways of expressing it. As a trivial example, there are well over 50 ways of saying "no" in English, ranging from "probably not" to "over my dead body." The degree to which the user chooses an unusual phrasing creates problems for a speech-understanding system.
- *Rejection of "off-the-subject" input.* In some applications the users may respond with a reasonable sentence that is beyond the scope of the system's language model. A collect-call system that is supposed to understand "yes" and "no" may have a difficult time coping with the response "I'll get Mommy." In cases like these the system may misrecognize the response and take some wildly incorrect action because of its misunderstanding. In some applications it is possible to train the users to avoid sentences that the system cannot understand, but this is not always practical. How can the system recognize what it does not "know"? Lynn Bates, in a comment at the NAS colloquium, has suggested building a separate language model just to catch the most frequent user responses that are out of the range of the original, more constrained language model. This second language model could be used to help prompt the user on what to say to be understood by the machine.

More powerful statistical techniques now being developed for text understanding (Marcus, in this volume) hold the promise of significantly improving the language understanding capabilities of advanced voice dialogue systems. Alex Waibel remarked at the colloquium that it is very time consuming to build special-purpose speech-understanding systems and that the long-term goal should be to create a machine that could learn the specific application with repeated

practice. With self-organizing systems such as neural networks, it might someday be possible to build this type of learning system. Recalling Richard Schwartz's comment about people becoming experts in using voice dialogue systems, it might be more practical in the short term to provide people with the feedback they need to adapt to the system rather than expect machines to adapt to people.

### Dimensions of the Speech Synthesis Task

There are two families of computer speech technologies today: digitized human speech and text-to-speech synthesis. Text-to-speech synthesis is flexible enough to pronounce any sentence but lacks the naturalness of recorded human speech. Digitized human speech is natural sounding but inflexible because only prerecorded phrases can be spoken. Text-to-speech systems are able to synthesize any text with an intelligibility almost as high as a human speaker. However, it is a major challenge to achieve naturalness in synthesized speech (Allen, in this volume; Carlson, in this volume).

For a speech output application the dimensions of difficulty relate to problems in synthesizing intelligible and pleasant-sounding computer speech, as opposed to speech understanding. The dimensions of difficulty are as follows:

- *Quantity of text.* It is impractical to record huge amounts of human speech (say, more than 100 hours) for a speech dialogue system. The vast majority of current applications in the voice response industry use recorded human speech. With digitized human speech, or waveform coding, the quality is limited only by the skill of the speaker and by the compression algorithm for the recorded speech, typically 2000 to 8000 bytes per second of speech. Recorded human speech has a major drawback, however, because every sentence must be recorded separately. Splicing together a phrase out of individually recorded words is unsatisfactory because of the "choppy" quality of the concatenated sentence. For applications in which a great variety of sentences must be spoken, or one in which the information changes frequently so that recording is impractical, text-to-speech synthesis must be used.
- *Variability of the input text.* There are applications in which the text being processed may contain abbreviations, jargon, or outright errors. Playing back electronic mail messages is one such application that must cope with error-prone input text, whereas pronouncing the names of catalog items has low variability. Also, specialized text preprocessors can be created for pronunciation of special vocabulary.

ies (such as prescription drug names); this is not practical for coverage of unrestricted English. When the text has low variability and consists of a few fixed phrases, recorded human speech can be used.

- *Length of the sentences and grammatical complexity.* Longer sentences tend to have more grammatical and semantic structure than short phrases, and current text-to-speech synthesizers provide only rudimentary linguistic analysis of the text (Allen, in this volume). Therefore, there is a tendency for longer, more complex sentences to have a poorer subjective rating than short simple phrases (Van Santen, 1993). An application in which the text is very complex, such as reading Shakespearean sonnets, would be more difficult than pronouncing words or short phrases.
- *Expectations of the listener.* Listeners are likely to be tolerant of a technology that provides them with a new and valuable service but intolerant of a system of poorer quality than what they are used to. For instance, consumers reacted positively to a service known as "Who's Calling?" in which they heard a text-to-speech synthesizer say "you have a call from the phone of John Doe" because this is a service not available before. But in other applications the quality of text-to-speech synthesis is a concern. In subjective tests of speech quality, text-to-speech synthesizers are judged significantly worse than digitized human speech (Van Santen, 1993). This remains true even when the text-to-speech synthesizer was provided with the pitch contour and the phoneme durations used by the original human speaker. Intelligibility for text-to-speech systems has been an issue in the past, but the intelligibility of modern systems is high enough for most applications. Word intelligibility measured at the word level is only slightly lower for good text-to-speech systems than for digitized human speech (Van Santen, 1993). However, intelligibility at the sentence level can be impaired when the prosody of complex sentences is so mangled that the meaning is obscured.

### **Additional Dimensions of Difficulty**

In addition to the dimensions of difficulty based on the limitations of the technologies of speech processing, there are engineering constraints on the deployment of any system: cost, size, power, and time available for deployment. In particular, cost, size, and power affect the amount of memory available for the speech processing and the power of the speech-processing chips. Hand-held devices and consumer equipment have particularly severe constraints.

- *Memory and processor requirements.* Speech recognition algo

rithms require millions of operations per second. Considerable ingenuity has been exerted to implement algorithms efficiently. In order of decreasing cost and computation power, recognizers have been programmed on parallel processors, RISC chips, floating point digital signal processors, general-purpose microprocessors, and integer digital signal processors. For speech synthesis, waveform coding systems require little processing power (a small fraction of the processing power of a digital signal processor chip), and the memory is proportional to the amount of speech to be played back. On the other hand, text to speech requires a high-speed microprocessor and between 0.5 and 5 megabytes of memory. For hand-held pronouncing dictionaries, text-to-speech synthesis is used because it is too costly to provide memory to store the waveform for each word in the dictionary.

- *System integration requirements.* As with other technologies, applications that rely on networked databases or processors, that need to access information for multiple users, that must coexist with existing equipment, or that must be compatible with future products and services will require more care during the systems engineering process than do stand-alone applications.

### Examples of Speech Applications

We have listed some of the dimensions of difficulty for a human-machine voice dialogue system. In principle, one might be able to rate an application along each dimension as "easy" or "difficult," thus arriving at an overall degree of difficulty of the application. Actual voice dialogue systems are not so easily quantified. But clearly an application that is rated "difficult" along most of the dimensions will require extraordinary effort to deploy. [Table 1](#) shows four human-machine communication systems with voice input/output.

1. VRCP (Voice Recognition Call Processing) is an AT&T service that automates the operator's role in placing long-distance calls (Wilpon, in this volume).
2. The DARPA ATIS task is an experimental system for getting information from a database of airline flights (Marcus, 1992).
3. Voice dialing refers to cellular telephones with speech recognition capability, so that calls may be made or received while driving a car without dialing by hand. Cellular telephones with voice dialing are sold by AT&T, Motorola, Nippon Electric Co., and others.
4. StockTalk (Lennig et al., 1992) is a system trialed recently by Bell Northern Research that allows people to get current prices of

TABLE 1 Degree of Difficulty of Four Voice Applications

		Application			
Dimension of Difficulty	AT&T's VRCP	DARPA ATIS	Cellular Phone with Name Dialing	Bell Northern Research's Stock Talk	
Speaker independence	Difficult: various dialects and nonnative speakers	Moderate: mostly native American English speakers	Easy: trained to one speaker	Difficult: many dialects and nonnative speakers	
Speaker expertise	Difficult: high proportion of first-time users	Moderate: speakers have time to rehearse a query	Easy: the owner is trained by the telephone	Moderate: First-time users have opportunity to practice	
Vocabulary size (acoustic confusability)	Easy: Seven dissimilar words	Difficult: unlimited vocabulary with confusable words	Moderate: user may train similar-sounding names	Moderate: over 2000 words, but most are dissimilar	
Grammar perplexity (number of choices)	Easy	Moderate: perplexity approx 50, but difficult to specify grammar	Moderate: perplexity approx 60	Difficult: no grammar; high perplexity	
Speaking mode (coarticulation and disfluencies)	Difficult: continuous extraneous speech, with barge in	Moderate: continuous speech with some disfluencies	Easy: isolated words	Easy: isolated words	
Channel variability (acoustic or electrical)	Moderate: telephone channel with handset	Easy: quiet laboratory conditions	Difficult: high noise levels, mike far from mouth	Moderate: telephone channel with handset	
User tolerance of errors	Moderate: a human operator is always available	N.A.	Moderate: user can hang up before a call is placed incorrectly	Easy: very little penalty for incorrect recognition	

Grammar complexity and ambiguity	Easy	Difficult: many complex sentences; context required	Easy	Easy: no grammar
Language variability	Moderate: synonyms for "yes" and "no"	Difficult: wide variety of sentence forms	Easy: though users forget how they trained names	Moderate: some stocks have several possible names
Rejection of extraneous speech	Moderate: must reject casual speech, answering machines	Easy: all "incorrect" queries excluded	Difficult: microphone is on at all times	N.A., no rejection
Quantity of speech to synthesize	Easy: uses prerecorded speech prompts	Difficult: TTS must synthesize millions of sentences	Easy: records user's voice for name feedback	Moderate: TTS used for company names
Variability of input text	N.A.	Moderate: text in the database can be prescreened	N.A.	Easy: pronunciations can be verified in advance
Length of sentence and grammatical complexity	N.A.	Difficult: long, complex sentences	N.A.	Easy: short, structured phrases
Listener expectations	High: recorded speech	Moderate	Moderate: should resemble user's voice	Easy: should be intelligible
Processor and memory requirements	Moderate: multichannel system	Easy: must run in close to real-time on a workstation	Difficult: low-cost, single-chip processor	Moderate: multichannel system

N.A., not applicable.

securities on several stock exchanges by speaking the name of the company.

Table 1 evaluates these four applications along each of the dimensions of difficulty for speech recognition, language understanding, and speech synthesis. The message is that each of these applications has some dimensions that are difficult and some that are easy. These are all cutting-edge systems in their own way, but the dimensions in which they excel are different.

### PROCEDURE FOR DEPLOYMENT OF SPEECH APPLICATIONS

The design and development of a voice transaction system is an iterative one. While one might think that a designer specifies the voice transaction, builds the speech input and output modules, and deploys the final system, this is not the case in practice. Because of the complexity and variety of dialogues, it is exceedingly difficult to anticipate all the factors that are critical to success. Experience has shown that developing a human-machine dialogue system is an iterative process.

A typical iterative process is shown in Figure 3. In the initial design the overall objectives of the system are set up. Development proceeds in a conventional way until the trial system is set up in field conditions. The system designers need to have an auditing system in place to determine what people say during voice transactions and what the machine's response is. Specific problems will be identified and diagnosed to correct the system's accuracy. Human-machine dia

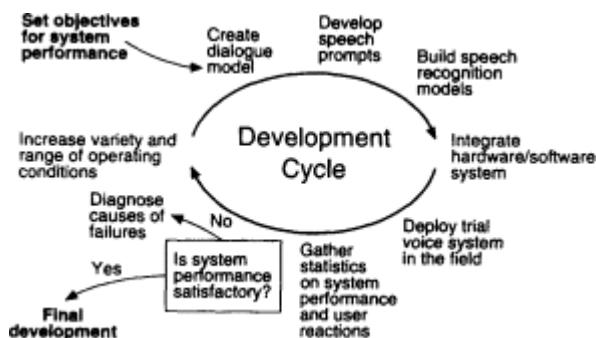


FIGURE 3 Deployment process for a voice dialogue system.

logue systems are different from more mature technologies in that they require several iterations.

With mature technologies it is possible to schedule a development timetable with a fair degree of confidence that the technology will work as planned and on schedule. Rarely if ever has this been possible with voice dialogue systems. The design of a user interface (Kamm, in this volume) is painstaking because people respond in unexpected ways. For instance, in automation of operator services, people were asked to say one of five phrases: "collect," "calling card," "operator," "person to person," and "third number." Twenty percent of the callers spoke these phrases with other words such as "please" or "uhhm." Rewording the voice prompts could reduce this problem but only with an unacceptably longer-duration prompt. The preferred solution, word spotting in speech recognition, took several years to develop and deploy. Second, it is difficult to gather speech for training speech recognizers unless you have a working system that will capture speech in exactly the environment encountered in the real service. Therefore, the speech recognition accuracy will be lower than expected based on the style of speech used for training.

The procedure outlined above for deployment of speech technology may seem ad hoc, but it is necessary given the current maturity of the technology. When an engineer designs a bridge, there is an initial design, there are calculations to check the structural soundness, the bridge is constructed, and the bridge functions as it was designed to. It is not necessary to build a "trial" bridge to find out why it is going to fall down or to design the "final" bridge by successive approximation. In this respect, speech technology is still immature. In some sense we still lack a complete set of design principles needed to guarantee the integrity of a human-machine dialogue system.

### **The Art of Human-Machine Dialogues**

Current voice dialogue practice encompasses engineering art as well as scientific knowledge. Fundamental knowledge of speech production and basic principles of pattern matching have been essential to the success of speech recognition over the past 25 years. That said, the art of successful engineering is critically important for applications of voice dialogue systems (Nakatsu and Suzuki, in this volume). There is an important element of craftsmanship in building a successful speech transaction. Often, this engineering art has been gained through trial and error. It should be emphasized that improving the engineering art is a proper and necessary topic for applied research.

The engineering art of speech recognition has improved significantly in the past few years, further opening up the range of possible applications.

*Subword units.* It is now possible to build a dictionary of models comprised of constituent phonetic (or phoneme-like) statistical models, first for small, easily distinguishable vocabularies, and later for larger vocabularies. The effort and expense of gathering speech from many speakers for each vocabulary word have been reduced.

- *Noise immunity.* Better speech enhancement algorithms and models of background noise make speech recognizers more accurate in noisy or changing environments, such as automobiles.
- *Speaker adaptation.* People can adapt quickly to dialects and accents in speech. Machines now have the beginnings of the capability to respond more accurately as they learn an individual voice.
- *Rudimentary language understanding.* The ability to spot key words in a phrase is the first step toward understanding the essence of a sentence even if some words are not recognized.
- *"Barge in."* It is sometimes desirable, when talking with a person, to be able to interrupt the conversation. In telephone-based voice response systems, it is possible to interrupt a prompt using Touch-Tones. This capability has been extended to allow users the ability to speak during a prompt and have the system recognize them.
- *Rejection.* An ability that people take for granted in conversation is the ability to detect when we do not understand. Unfortunately, this is a most difficult task for current speech recognition systems. While it is possible to determine when there are two (or more) possible words or sentences, it has been very difficult for systems to determine when people are saying something on a completely different subject. This can lead to comical, if not frustrating, results for the user. Further research is needed in detecting this type of "none of the above" response.

The design of an easy-to-use dialogue with a computer system is a significant challenge. We know from experience that is possible to design good human interfaces for computer dialogue systems. Unfortunately, it has also been verified that it is possible to design systems that aggravate people. At this time there are some general guidelines for good human interface design, but there is no "cookbook" recipe that guarantees a pleasant and easy-to-use system (Kamm, in this volume).

## CONCLUSIONS

The concept of the degree of difficulty of a human-machine voice dialogue system can be used to evaluate its feasibility. The degree of difficulty of a particular application depends on many factors. Some are obvious, but others are easy to overlook. For example, the expertise of the users has a dramatic effect on the performance of these systems. Also, the willingness of users to overlook deficiencies in the system varies widely depending on whether there are other alternatives. A comprehensive view of all the dimensions of difficulty is needed in order to assess the overall degree of difficulty.

Deployment of voice transaction services is an iterative process. Because the machine must cope with errors made by the person, and the human being must cope with errors made by the machine, the nature of the transaction is difficult if not impossible to predict in advance. Though the ultimate goal is to create a machine that can adapt to the transaction as it gains more experience, the human-machine dialogue systems of today require engineering art as well as scientific principles.

## REFERENCES

- Bahl, L. R., et al., "Large-vocabulary natural language continuous speech recognition," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 465-468, Glasgow, Scotland, May 1989.
- Berwick, R., "Intelligent natural language processing: Current trends and future prospects," pp. 156-183 in *AI in the 1980's and beyond*, W. E. Grimson and R. S. Patil, eds., MIT Press, Cambridge, Mass., 1987.
- Gauvin, J., and C. H. Lee, "Improved acoustic modeling with Bayesian learning," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 481-484, San Francisco, 1992.
- Hirst, G., *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge, England, 1987.
- Lennig, M., D. Sharp, P. Kenny, V. Gupta, and K. Precoda, "Flexible vocabulary recognition of speech," in Proceedings of the 1992 International Conference on Spoken Language Processing, pp. 93-96, Banff, Alberta, Canada, Oct. 1992.
- Marcus, M., ed., *Proceedings, Speech and Natural Language Workshop*, 1992, Harriman, New York, Morgan Kaufmann Publishers, San Mateo, Calif., Feb. 1992.
- Rabiner, L., and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993.
- Ramesh, P., et al., "Speaker-independent recognition of spontaneously spoken connected digits," *Speech Communication*, Vol. 11, pp. 229-235, 1992.
- Van Santen, J. P. H., "Perceptual experiments for diagnostic testing of text-to-speech systems," *Computer Speech and Language*, Vol. 7, No. 1, pp. 49-100, Jan. 1993.

# What Does Voice-Processing Technology Support Today?

*Ryohei Nakatsu and Yoshitake Suzuki*

## SUMMARY

This paper describes the state of the art in applications of voice-processing technologies. In the first part, technologies concerning the implementation of speech recognition and synthesis algorithms are described. Hardware technologies such as microprocessors and DSPs (digital signal processors) are discussed. Software development environment, which is a key technology in developing applications software, ranging from DSP software to support software also is described. In the second part, the state of the art of algorithms from the standpoint of applications is discussed. Several issues concerning evaluation of speech recognition/synthesis algorithms are covered, as well as issues concerning the robustness of algorithms in adverse conditions.

## INTRODUCTION

Recently, voice-processing technology has been greatly improved. There is a large gap between the present voice-processing technology and that of 10 years ago. The speech recognition and synthesis market, however, has lagged far behind technological progress. This paper describes the state of the art in voice-processing technology applications and points out several problems concerning market growth that need to be solved.

Technologies related to applications can be divided into two categories. One is system technologies and the other is speech recognition and synthesis algorithms.

Hardware and software technologies are the main topics for system development. Hardware technologies are very important because any speech algorithm is destined for implementation on hardware. Technology in this area is advancing quickly. Microprocessors with capacities of about 100 MIPS are available. Also, digital signal processors (DSPs) that have capabilities of nearly 50 MFLOPS have been developed (Dyer and Harms, 1993) for numerical calculations dedicated to voice processing. Almost all speech recognition/synthesis algorithms can be used with a microprocessor and several DSPs. With the progress of device technology and parallel architecture, hardware technology will continue to improve and will be able to cope with the huge number of calculations demanded by improved algorithms of the future.

Also, software technologies are an important factor, as algorithms and application procedures should be implemented by the use of software technology. In this paper, therefore, software technology will be treated as an application development tool. Along with the growth areas of application of voice-processing technology, various architectures and tools that support applications development have been devised. This architecture and these tools range from compilers for developing DSP firmware to software development tools that enable users to develop dedicated software from application specifications. Also, when speech processing is the application target, it is important to keep in mind the characteristics peculiar to speech. Speech communication basically is of a nature that it should work in a real-time interactive mode. Computer systems that handle speech communications with users should have an ability to cope with these operations. Several issues concerning real-time interactive communication will be described.

For algorithms there are two important issues concerning application. One is the evaluation of algorithms, and the other is the robustness of algorithms under adverse conditions. Evaluation of speech recognition and synthesis algorithms has been one of the main topics in the research area. However, to consider applications, these algorithms should be evaluated in real situations rather than laboratory situations, which is a new research trend. There are two recent improvements in algorithm evaluation. First, algorithm evaluation using large-scale speech databases, which are developed and shared by many research institutions, means that various types of algorithms can be more easily and extensively compared. The second improve

ment is that, in addition to the use of comprehensive databases for evaluation, the number of databases that include speech uttered under adverse conditions is increasing.

Also, the robustness of algorithms is a crucial issue because conditions in almost all real situations are adverse, and algorithms should therefore be robust enough for the application system to handle these conditions well. For speech recognition, robustness means that algorithms are able to cope with various kinds of variations that overlap or are embedded in speech. In addition to robustness in noisy environments, which is much studied, robustness for speech variabilities should be studied. Utterances of a particular individual usually contain a wide range of variability which makes speech recognition difficult in real conditions. Technologies that can cope with speech variabilities will be one of the key technologies of future speech recognition.

Finally, several key issues will be discussed concerning advanced technology and how its application can contribute to broadening the market for speech-processing technology.

## SYSTEM TECHNOLOGIES

When speech recognition or synthesis technology is applied to real services, the algorithms are very important factors. The system technology—how to integrate the algorithms into a system and how to develop programs for executing specific tasks—is similarly a very important factor since it affects the success of the system. In this paper we divide the system technology into hardware technology and application—or software—technology and describe the state of the art in each of these fields.

### Hardware Technology

#### Microprocessors

Whether a speech-processing system utilizes dedicated hardware, a personal computer, or a workstation, a microprocessor is necessary to control and implement the application software. Thus, microprocessor technology is an important factor in speech applications. Microprocessor architecture is categorized as CISC (Complex Instruction Set Computer) and RISC (Reduced Instruction Set Computer) (Patterson and Ditzel, 1980). The CISC market is dominated by the Intel x86 series and the Motorola 68000 series. RISC architecture was developed to improve processing performance by simplifying the instruction set and reducing the complexity of the circuitry. Recently,

RISC chips are commonly used in engineering workstations. Several common RISC chips are compared in Figure 1. Performance of nearly 300 MIPS is available. The processing speed of CISC chips has fallen behind that of the RISC in recent years, but developers have risen to the challenge to improve the CISC's processing speed, as the graph of the performance of the x86 series in Figure 2 shows. Better microprocessor performance makes it possible to carry out most speech-processing algorithms using standard hardware, whereas formerly dedicated hardware was necessary. In some applications, the complete speech-processing operation can be carried out using only a standard microprocessor.

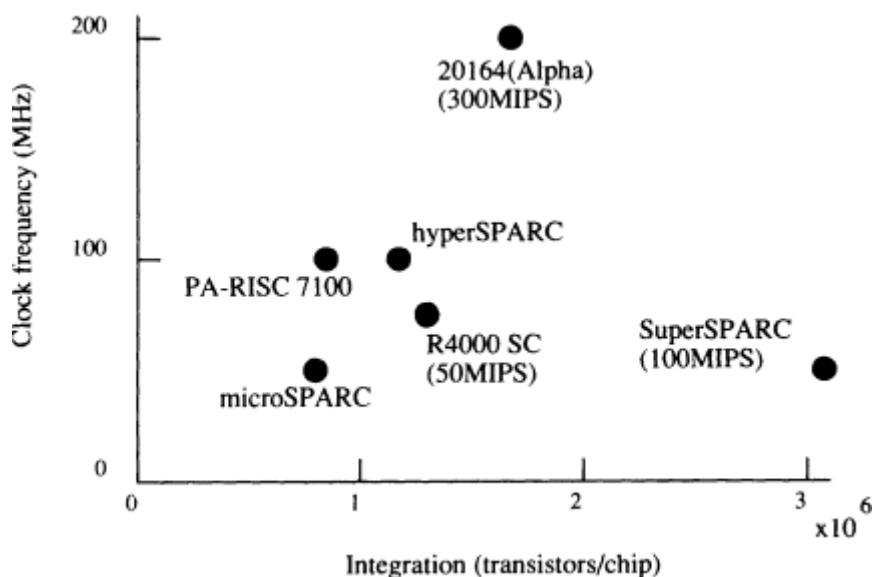


FIGURE 1 Performance of RISC chip.

### Digital Signal Processors

A DSP, or digital signal processor, is an efficient device that carries out algorithms for speech and image processing digitally. To process signals efficiently, the DSP chip uses the following mechanisms: high-speed floating point processing unit, pipeline multiplier and accumulator, and parallel architecture of arithmetic-processing units and address calculation units.

Specifications of typical current DSPs are shown in Table 1. DSPs of nearly 50 MFLOPS are commercially available. Recent complicated

speech-processing algorithms need broad dynamic range and high precision. The high-speed floating point arithmetic unit has made it possible to perform such processing in minimal instruction cycles without loss of calculation precision. Increased on-chip memory has made it possible to load the whole required amount of data and access internally for the speech analysis or pattern-matching calculations, which greatly contributes to reducing instruction cycles. Usually more cycles are needed to access external memory than are needed for accessing on-chip memory. This is a bottleneck to attaining higher throughput. The amount of memory needed for a dictionary for speech recognition or speech synthesis is too large to implement on-chip though; several hundred kilobytes or several megabytes of external memory are required to implement such a system.

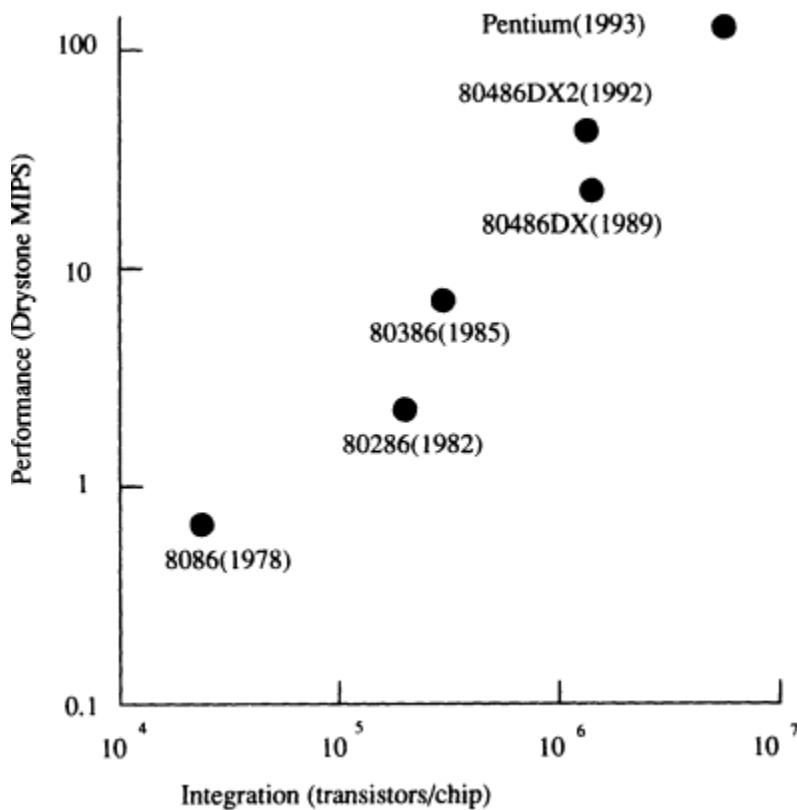


FIGURE 2 Performance improvement of x86 microprocessor.

Most traditional methods of speech analysis, as well as speech

TABLE 1 Specifications of Current Floating-Point DSPs

Developer	Name	Cycle Time	Data Width	Multiplier	On-chip RAM	Address	Technology	Electric Power Consumption
Texas Instruments, Motorola	TMS320C40 DSP96002	40 ns 50 ns	24E8 24E8	24E8 x 24E8 - 32E8 32E11 x 32E11 - 64E15 W	2K w (1K+512) W	4G w 12G w	0.8- $\mu$ m CMOS	1.5 W
AT&T	DSP3210	60 ns	24E8	24E8 x 24E8 - 32E8	2K w	1G w	1.2- $\mu$ m CMOS	1.5 W
NEC	mPD77240	90 ns	24E8	24E8 x 24E8 - 47E8	1K w	64M w	0.9- $\mu$ m CMOS	1.0 W
Fujitsu	MB86232	75 ns	24E8	24E8 x 24E8 - 24E8	512 w	1M w	0.8- $\mu$ m CMOS	1.6 W
							1.2- $\mu$ m CMOS	1.0 W

recognition algorithms based on the HMM (hidden Markov model), can be carried out in real-time by a single DSP chip and some amount of external memory. On the other hand, recent speech analysis, recognition, and synthesis algorithms have become so complex and time consuming that a single DSP cannot always complete the processing in real-time. Consequently, to calculate a complex algorithm in real-time, several DSP chips work together using parallel processing architecture. The architecture of the transputer, a typical parallel signal-processing system, is shown in [Figure 3](#). It has four serial data communication links rather than a parallel external data bus, so that the data can be sent effectively and the computation load can be distributed to (four) other transputers. The parallel programming language "Occam" has been developed for this chip.

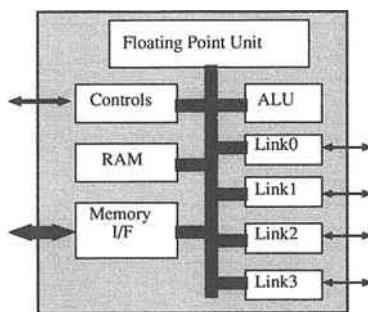


FIGURE 3 Transputer architecture (Inmos, 1989).

## Equipment and Systems

Speech-processing equipment or a speech-processing system can be developed using either microprocessors or DSPs. They can use one of the following architectures: (a) dedicated, self-contained hardware that can be controlled by a host system, (b) a board that can be plugged into a personal computer or a workstation, and (c) a dedicated system that includes complete software and other necessities for an application.

In the early days of speech recognition and synthesis, devices of type (a) were developed because the performance of microprocessors and DSPs was not adequate for real-time speech processing (and were

not so readily available). Instead, the circuits were constructed using small-scale integrated circuits and wired logic circuits. Recently, however, DSPs are replacing this type of system. A type (a) speech-processing application system can be connected to a host computer using a standard communication interface such as the RS-232C, the GPIB, or the SCSI. The host computer executes the application program and controls the speech processor as necessary. In this case the data bandwidth is limited and is applicable only to relatively simple processing operations.

As for case (b), recent improvements of digital device technology such as those shown in Figures 1 and 2, and Table 1 have made it possible to install a speech-processing board in the chassis of the increasingly popular personal computers and workstations. The advantages of this board-type implementation are that speech processing can be shared between the board and a host computer or workstation, thus reducing the cost of speech processing from that using self-contained equipment; speech application software developed for a personal computer or workstation equipped with the board operates within the MS-DOS or UNIX environment, making it easier and simpler to develop application programs; and connecting the board directly to the personal computer or workstation bus allows the data bandwidth to be greatly widened, which permits quick response—a crucial point for a service that entails frequent interaction with a user.

In some newer systems only the basic analog-to-digital and digital-to-analog conversions are performed on the board, while the rest of the processing functions are carried out by the host system.

Examples of recent speech recognition and speech synthesis systems are shown in Figure 4 and Figure 5, respectively. Figure 4 shows a transputer-based speech recognizer, which comprises one board that plugs into the VME bus interface slot of a workstation. The function of the board is speech analysis by DSPs, HMM-based speaker-independent word spotting, and recognition candidate selection using nine transputer chips (Imamura and Suzuki, 1990). The complete vocabulary dictionary is shared among the transputers, and the recognition process is carried out in parallel. The final recognition decision, control of the board, and the other processing required by the application are done in the host workstation. The application development and debugging are also done on the same host machine. The Japanese text-to-speech synthesizer shown in Figure 5 is also a plug-in type, for attachment to a personal computer. Speech synthesis units are created by the context-dependent phoneme unit using the clus

tering technique. The board consists mainly of the memory for the synthesis unit, a DSP for LSP synthesis, and a digital-to-analog converter (Sato et al., 1992).

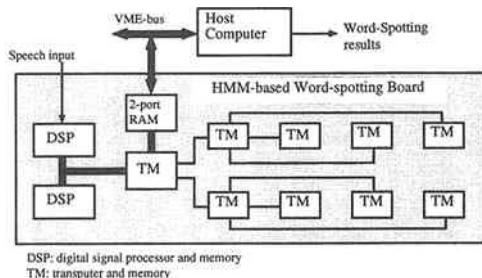


FIGURE 4 Example of a speech recognition system (Imamura and Suzuki, 1990).

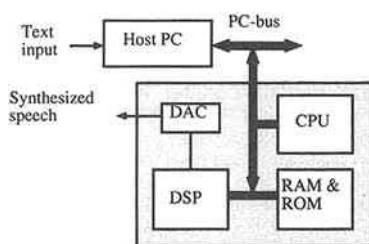


FIGURE 5 Example of a text-to-speech synthesizer (Sato et al., 1992).

Method (c) is adopted when an application is very large. The speech system and its specific application system are developed as a package, and the completed system is delivered to the user. This method is common for telecommunications uses, where the number of users is expected to be large. In Japan a voice response and speech recognition system has been offered for use in public banking services since 1981. The system is called ANSER (Automatic answer Network System for Electrical Request). At first the system had only

the voice response function for Touch-Tone telephones. Later, a speech recognition function was added for pulse, or rotary-dial, telephone users. After that, facsimile and personal computer interface capabilities were added to the system, and ANSER is now an extensive system (Nakatsu, 1990) that processes more than 30 million transactions per month, approximately 20 percent of which are calls from rotary-dial telephones. The configuration of the ANSER system is shown in [Figure 6](#).

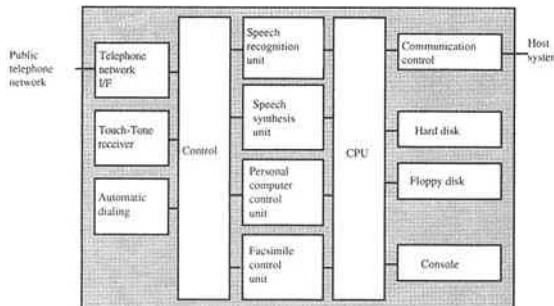


FIGURE 6 Configuration of the ANSER system (Nakatsu and Ishii, 1987).

## Application Technology Trend

### Development Environment for DSP

As mentioned above, a DSP chip is indispensable to a speech-processing system. Therefore, the development environment for a DSP system is a critical factor that affects the turnaround time of system development and eventually the system cost. In early days only an assembler language was used for DSPs, so software development required a lot of skill. Since the mid-1980s, the introduction of high level language compilers (the C cross compiler is an example) made DSP software development easier. Also, some commonly used algorithms are offered as subroutine libraries to reduce the programmer's burden. Even though the cross compilers have become more popular these days, software programming based on an assembler language may still be necessary in cases where real-time processing is critical.

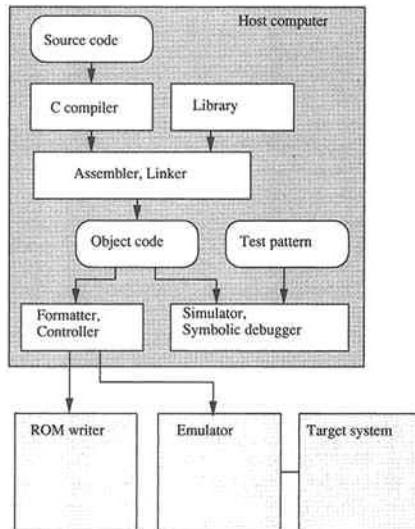


FIGURE 7 DSP development environment (Texas Instruments, 1992).

An example of a DSP development environment is shown in [Figure 7](#). A real-time operating system for DSP is available. Moreover, a parallel processing development environment is offered for parallel DSP systems. However, the critical aspect of parallel DSP programming depends substantially on the programmer's skill. A more efficient and usable DSP development environment is needed to make DSP programming easier and more reliable.

### Application Development Environment

Environments in application development take various forms, as exemplified by the varieties of DSP below:

- (a) Application software is developed in a particular language dedicated to a speech system.
- (b) Application software is developed using a high level language such as C.
- (c) Application software is developed using an "application builder," according to either application flow or application specifications.

Application software used to be written for each system using assembler language. Such cases are rare recently because of the inadequacy of a low-level programming language for developing large and complex systems. However, assembler language is still used in special applications, such as games or telephones, that require compact and high-speed software modules.

Recently, applications have usually been described by a high level language (mainly C). Speech recognition and synthesis boards are commonly plugged into personal computers, and interface subroutines with these boards can be called from an application program written in C. An application support system for Dragon Writer (Dragon Systems, Inc.) is shown in [Figure 8](#).

As an application system becomes larger, the software maintenance becomes more complicated; the user interface becomes more important; and, eventually, control of the speech system becomes more complicated, the specifications become more rigid, and improvement becomes more difficult. Therefore, the application development should be done in a higher-level language. Moreover, it is desirable for an "Application Builder" to adhere to the following principles: automatic program generation from a flowchart, automatic generation of an application's grammar, and graphical user interface for software development. Such an environment is called "Application Builder." Development of the Application Builder itself is an important and difficult current theme of automatic software generation research. So far, few systems have been implemented using the Application Builder, though some trials have been done in the field of speech-processing systems (Renner, 1992).

## Speech Input/Output Operating Systems

Availability of a speech input/output function on a personal computer or workstation is a requisite for popularizing the speech input/output function. It is adequate to implement the speech input/output function at the operating system level. There are two ways: (a) developing a dedicated operating system with speech input/output function and (b) adding a speech input/output function as a preprocessor to an existing operating system.

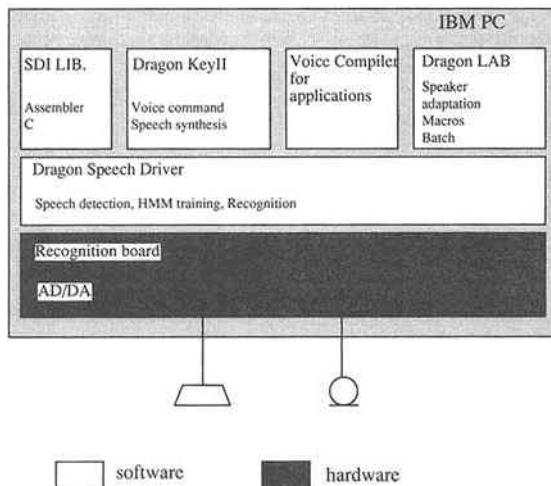


FIGURE 8 Application support of speech recognition system (Dragon Writer).

Method (a) best utilizes speech recognition and synthesis capabilities. With this method a personal computer or workstation system with a speech input/output function can be optimally designed. However, no operating system that is specific to speech input/output has yet been developed. In the field of character recognition, on the other hand, pen input specific operating systems—called "Pen OS"—have begun to appear. These operating systems are not limited to handwritten character recognition; they are also capable of pen-pointing and figure command recognition. In other words, character recognition is not the primary function but just one of the functions. This grew out of the realization that recognition of handwritten characters is not yet satisfactory and that handwriting is not necessarily faster than keyboard input. Optimum performance was achieved by combining several data input functions such as keyboard input, mouse

clicking, or pen input. This principle should be applied to the development of speech operating systems.

In method (b) there is no need to develop a new operating system. For example, a Japanese kana-kanji preprocessor, which converts keyboard inputs into 2-byte Japanese character codes, is widely used for Japanese character input. The same concept can be applied to speech input/output. An advantage of this method is that speech functions can easily be added to popular operating systems such as MS-DOS. An example of the speech preprocessor is shown in [Figure 9](#). Input can alternate between mouse and speech, and menu selection can be carried out by either mouse clicking or speech input. Input can be accomplished more efficiently by speech, especially when using a drawing application.

### Real-Time Application Support

Real-time dialogue-processing function is a future need in speech applications. The speech system must carry out an appropriate control of the speech recognizer and the speech synthesizer to realize the

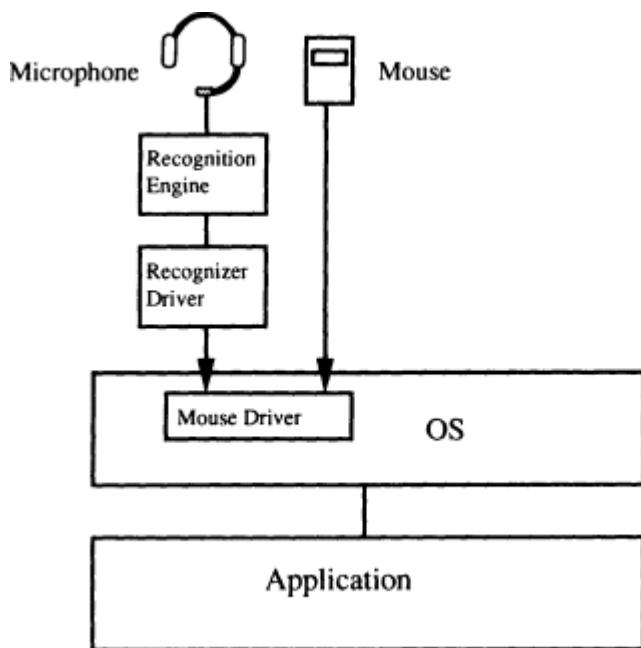


FIGURE 9 Configuration of a speech preprocessor.

real-time dialogue between humans and computers. Also, an easy method to describe this control procedure needs to be developed and opened to users so that they can develop dialogue control software easily. These points would be the main problems when developing an application system in the near future. The state of the art and the required technologies for the real-time application system are described below.

*Barge in* "Barge in" means an overlap or collision of utterances, which frequently occurs in human speech conversation. This phenomenon is actually an important factor in making the communication smooth. The time delay on an international telecommunication line makes the conversation awkward because it interferes with this barge in. It is pointed out that the same phenomenon readily occurs in dialogues between humans and computers. The following techniques allow for barge in in a speech system:

- The system starts recognition as soon as it starts sending a voice message. An echo canceller can be applied to subtract the synthesizer's voice and to maintain recognition performance.
- The system may start sending a voice message as soon as the key word is extracted even if the speaker's voice has not been completed yet.

Although the former function has been implemented in some systems, no application has yet incorporated the latter function. This function can be carried out in a small-vocabulary system, so the human-machine interface should be researched in this case.

*Key Word Extraction* Even in applications based on word recognition, humans tend to utter extra words or vocal sounds and sometimes even entire sentences. When the recognition vocabulary is limited, a simple key word extraction or spotting technique is used. (If the vocabulary size is huge or unlimited, the speech must be recognized in its entirety.) Several word-spotting algorithms have been proposed (Kimura et al., 1987; Wilpon et al., 1990) and proven efficient for extracting only key words.

*Distributed Control of DSP* Real-time processing is needed for speech recognition and speech synthesis. Currently, DSP hardware and software design is entrusted to skillful systems engineers, and this sometimes creates a bottleneck in system development. The following development environment is desirable:

- Speech-processing algorithms are transformed into programs, ignoring real-time realization problems.
- The DSP system has a flexible parallel architecture and a real-time operating system. When the program to be processed in real-time is loaded, the operating system automatically divides the job into subtasks that can be processed in real-time by single DSPs. These subtasks are distributed to several DSPs, and then the operation is executed in real-time.

## ALGORITHMS

Recognition of large spoken vocabularies and understanding of spontaneous spoken language are studied eagerly by many speech recognition researchers. Recent speech synthesis research focuses on improvement of naturalness and treatment of prosodic information. The state of the art of speech recognition/synthesis technologies is described further elsewhere in this volume in papers by Carlson and Makhoul and Schwartz. In most of these studies, rather clean and closed speech data are usually used for both training and evaluation. However, field data used in developing applications are neither clean nor closed. This leads to various problems, which are described below:

- *Databases.* A speech database that includes field data rather than laboratory data is necessary both for training and evaluating speech recognition systems.
- *Algorithm assessment.* Evaluation criteria other than those that have been used to undertake algorithm evaluation in the laboratory should be used to assess the feasibility of speech recognition/synthesis algorithms and systems to be used in real applications.
- *Robustness of algorithms.* In a real environment a broad spectrum of factors affect speech. Speech recognition/synthesis algorithms should be robust under these varied conditions. These topics will be discussed in the following sections.

## Databases

### Databases for Research

Currently, in speech recognition research, statistical methods such as the HMM are widely used. One of the main characteristics of a statistical method is that its performance generally depends on the quantity and quality of the speech database used for its training. The

amount of speech data collected is an especially important factor for determining its recognition performance. Because construction of a large database is too big a job for a single researcher or even a single research institute, collaboration among speech researchers to construct databases has been very active. In the United States, joint speech database construction is undertaken in Spoken Language programs supported by DARPA (Defense Advanced Research Projects Agency) (MADCOW, 1992; Pallet et al., 1991). In Europe, under several projects such as ESPRIT (European Strategic Program for Research and Development in Information Technology), collaborative work has been done for constructing large speech databases (Carre et al., 1984; Gauvain et al., 1990). Also, in Japan large speech databases are under construction by many researchers at various institutes (Itahashi, 1990).

For speech synthesis, on the other hand, concatenation of context-dependent speech units has recently proven efficient for producing high-quality synthesized speech. In this technology selection of the optimum unit for concatenation from a large set of speech units is essential. This means that a large speech database is necessary in speech synthesis research also, but a trend toward collaborative database construction is not yet apparent in this area. The main reason is that the amount of speech data needed for speech synthesis is far less than that for speech recognition because the aim of speech synthesis research is not to produce various kinds of synthesized speech.

## Databases for Application

The necessity of a speech database is not yet being discussed from the standpoint of application. Two reasons are as follows:

- The application area both for speech recognition and speech synthesis is still very limited. Accordingly, there is not a strong need for constructing speech databases.
- Applications are tied closely to business. This means that vendors or value-added resellers who are constructing speech databases do not want to share the information in their databases.

For telephone applications, on the other hand, disclosure of information about databases or the databases themselves is becoming common. In this area, applications are still in the early stages and competition among vendors is not strong yet. Another reason is that in telephone applications the amount of speech database necessary for training a speaker-independent speech recognition algorithm is

huge, which has led to a consensus between researchers that collaboration on database construction is necessary.

TABLE 2 Examples of Telephone Speech Database

Institute	Vocabulary	No. Speakers	Region
NTT	Sixteen words	3000 (male,: 1500; female: 1500)	Three regions Tokyo, Osaka, Kyushu
Texas Instruments	Connected-digit sentence	3500	Eleven dialects
AT&T	Three words	Training: 4000 Test: 1600	Eight regions in Spain
	Six words	Training: 22,000 Test: 4500	Eighteen regions in the United Kingdom
Oregon Graduate Institute	Name, city name	4000	
NYNEX (planning)	6500-8000 words	1000	

SOURCE: Picone, 1990; Walker and Millar, 1989.

Some examples of speech database collection through telephone lines are summarized in Table 2. In the United States, Texas Instruments is trying to construct a large telephone speech database that is designed to provide a statistically significant model of the demographics of the U.S. population (Picone, 1990). Also, AT&T is actively trying to adapt word-spotting speech recognition technology to various types of telephone services. For that purpose, the company has been collecting a large amount of speech data through telephone lines (Jacobs et al., 1992). Several other institutes are also constructing various kinds of telephone speech databases (Cole et al., 1992; Pittrelli et al., 1992). In Europe, various trial applications of speech recognition technology to telephone service are under way (Rosenbeck and Baungaard, 1992; Walker and Millar, 1989). In Japan, as discussed earlier, the ANSER speech recognition and response system has been widely used for banking services since the beginning of the 1980s. For training and evaluating the speech recognition algorithm used in the ANSER system, a large speech database consisting of utterances of more than

3000 males and females ranging in age from 20 to 70 was constructed (Nomura and Nakatsu, 1986).

As telephone service has been considered a key area for the application of speech recognition technology, various trials have been undertaken and various issues are being discussed. Several comments on these issues, based on the experiences of developing the ANSER system and participating in the operation of the banking service, are given below.

### Simulated Telephone Lines

Because of the difficulties underlying the collection of a large speech corpus through telephone lines, there has been a discussion that a synthetic telephone database, constructed by passing an existing speech database through a simulated telephone line, could be used as an alternative (Rosenbeck, 1992), and a prototype of such a simulated telephone line has been proposed (Yang, 1992).

[Table 3](#) summarizes the use of speech data to improve service performance in the ANSER system. In the first stage of the ANSER services, a speech database from simulated telephone lines was used. This database was replaced by real telephone line data because recognition based on the simulated telephone data was not reliable enough. From these experiences it was concluded that simulated telephone data are not appropriate for real telephone applications. The main reasons are as follows:

TABLE 3 Milestones of Speech Recognition in ANSER

Training Data				
		Line	No. Speakers	Region
Spring 1981	Service started in Tokyo area	Pseudo-telephone line	500	Tokyo
Autumn 1981	Service recognizer retrained	Telephone line	500	Tokyo
Spring 1982	Service area widened to Osaka and Kyushu areas	Telephone line	500	Tokyo
Autumn 1982	Speech recognizer retrained	Telephone line	1500	Tokyo, Osaka, Kyushu

- Characteristics of real telephone lines vary greatly depending on the pass selected. It is difficult, therefore, to simulate these variations using a simulated line.
- In addition to line characteristics, various noises are added to speech. Again, it is difficult to duplicate these noises on a simulated line.

To overcome the difficulty of collecting a large speech corpus through real telephone lines, NYNEX has constructed a large database, called NTIMIT (Jankowski et al., 1990), by passing recorded speech through a telephone network.

*Dialects* It is considered important for speech data to include the various dialects that will appear in real applications. This dialect factor was taken into consideration in the construction of several databases (Jacobs et al., 1992; Picone, 1990). The ANSER service used speech samples collected in the Tokyo area when the real telephone line database was introduced. As the area to which banking service is offered expanded to include the Osaka and Kyushu areas, it was pointed out that the performance was not as good in these areas as it was in the Tokyo area. In response to the complaints, telephone data were collected in both the Osaka and Kyushu areas. The speech recognizer was retrained using all of the utterances collected in all three areas, and recognition performance stabilized.

### Assessment of Algorithms

#### Assessment of Speech Recognition Algorithms

In the early days of speech recognition, when word recognition was the research target, the word recognition rate was the criterion for assessment of recognition algorithms. Along with the shift of research interest from word recognition to speech understanding, various kinds of assessment criteria, including linguistic processing assessment, have been introduced. These assessment criteria are listed in [Table 4](#). Regarding application, the following issues are to be considered.

*Assessment Criteria for Applications* In addition to the various criteria used for the assessment of recognition methods at the laboratory level, other criteria should be introduced in real applications. Some of these assessment criteria are listed in [Table 5](#). Recently, various kinds of field trials for telephone speech recognition have been undertaken in

which several kinds of assessment criteria are used (Chang, 1992; Nakatsu and Ishii, 1987; Rosenbeck and Baungaard, 1992; Sprin et al., 1992). Among these criteria, so far, the task completion rate (TCR) is considered most appropriate (Chang, 1992; Neilsen and Kristernsen, 1992). This matches the experience with the ANSER service. After several assessment criteria had been measured, TCR was determined to be the best criterion. TCR was under 90 percent, and there were many complaints from customers. As TCR exceeded 90 percent, the number of complaints gradually diminished, and complaints are rarely received now that the TCR exceeds 95 percent.

TABLE 4 Assessment Criteria for Speech Recognition

Phoneme Level	Word Level	Sentence Level
Correct segmentation rate	Isolated word: Word recognition rate	Word recognition rate (after linguistic processing)
Phoneme recognition rate	Connected word: Word recognition rate (including insertion, deletion, and substitution) Word sequence recognition rate Key-word extraction rate	Sentence recognition rate  Correct answer rate

*Assessment Using the "Wizard of Oz" Technique* When application systems using speech recognition technology are to be improved in order to raise user satisfaction, it is crucial to pinpoint the bottlenecks. The bottlenecks could be the recognition rate, rejection reliability, dialogue design, or other factors. It is nearly impossible to create a sys

TABLE 5 Assessment Criteria of Speech Recognition from User's Side

Objective Criteria	Subjective Criteria
Task completion rate	Satisfaction rating
Task completion time	Fatigue rating
Number of interactions	Preference
Number of error correction sequences	

tem to evaluate the bottlenecks totally automatically because there are too many parameters and because several parameters, such as recognition rate, are difficult to change. The "Wizard of Oz" (WOZ) technique might be an ideal alternative to this automatic assessment system. When the goal is a real application, reliable preassessment based on the WOZ technique is recommended. One important factor to consider is processing time. When using the WOZ technique, it is difficult to simulate real-time processing because humans play the role of speech recognizer in the assessment stage. Therefore, one must factor in the effect of processing time when evaluating simulation results.

Besides its use for assessment of user satisfaction, other applications of this technique include the comparison of Touch-Tone input with voice input (Fay, 1992) or of comparing human interface service based on word recognition with that based on continuous speech recognition (Honma and Nakatsu, 1987).

### **Assessment of Speech Synthesis Technology**

Speech synthesis technology has made use of assessment criteria such as phoneme intelligibility or word intelligibility (Steeneken, 1992). What is different from speech recognition technology is that almost all the assessment criteria that have been used are subjective criteria. **Table 6** summarizes these subjective and objective assessment criteria. As environments have little effect on synthesized speech, the same assessment criteria that have been used during research can be applied to real use. However, as applications of speech synthesis technology rapidly diversify, new criteria for assessment by users arise:

- a. In some information services, such as news announcements, customers have to listen to synthesized speech for lengthy periods,

TABLE 6 Assessment Criteria of Speech Synthesis

Intelligibility	Naturalness
Phoneme intelligibility	Preference score
Syllable intelligibility	MOS
Word intelligibility	
Sentence intelligibility	

- so it is important to consider customers' reactions to listening to synthesized speech for long periods.
- b. More important than the intelligibility of each word or sentence is whether the information or meaning is conveyed to the user.
  - c. In interactive services such as reservations, it is important for customers to realize from the beginning that they are interacting with computers, so for some applications synthesized speech should not only be intelligible but should also retain a synthesized quality.

Several studies have been done putting emphasis on the above points. In one study, the rate of conveyance of meaning was evaluated by asking simple questions to subjects after they had listened to several sentences (Hakoda et al., 1986). Another study assessed the effects on listener fatigue of changing various parameters such as speed, pitch, sex, and loudness during long periods of listening (Kumagami et al., 1989).

### **Robust Algorithms**

#### **Classification of Factors in Robustness**

Robustness is a very important factor in speech-processing technology, especially in speech recognition technology (Furui, 1992). Speech recognition algorithms usually include training and evaluation procedures in which speech samples from a database are used. Of course, different speech samples are used for training procedures than for evaluation procedures, but this process still contains serious flaws from the application standpoint. First, the same recording conditions are used throughout any one database. Second, the same instruction is used for all speakers in any one database.

This means that speech samples contained in a particular database have basically similar characteristics. In real situations, however, speakers' environments vary. Moreover, speech characteristics also tend to vary depending on the environments. These phenomena easily interfere with recognition. This is a fundamental problem in speech recognition that is hard to solve by algorithms alone.

Robustness in speech recognition depends on development of robust algorithms to deal with overlapping variations in speech. Factors that determine speech variations are summarized in [Table 7](#).

*Environmental Variation* Environments in which speakers input speech samples tend to vary. More specifically, variation of transmission characteristics between the speech source and the microphone such

as reflection, reverberation, distance, telephone line, and characteristics of the microphone itself are the causes of these variations.

TABLE 7 Factors Determining Speech Variation

Environment	Noise	Speaker
Reflection Reverberation	Stationary or quasi-stationary noise	Interspeaker variation
Distance to microphone	White noise Car noise	Dialect Vocal tract
Microphone characteristics	Air-conditioner noise	characteristics Speaking manner Coarticulation
	Nonstationary noise Other voices Telephone bells	Intraspeaker variation Emotion Stress Lombard effect
	Printer noise	

*Noise* In addition to various kinds of stationary noise, such as white noise, that overlap speech, real situations add various extraneous sounds such as other voices. All sound other than the speech to be recognized should be considered noise.

*Speech Variation* The human voice itself tends to vary substantially depending on the situation. In addition to interspeaker variation, which is a key topic in speech recognition, speech of the individual person varies greatly depending on the situation. This intraspeaker variation is a cardinal factor in speech recognition.

These variations, and recognition algorithms to compensate for them, are described in more detail below.

### Environmental Variation

Environmental variations that affect recognition performance are distance between the speech source and the microphone, variations of transmission characteristics caused by reflection and reverberation, and microphone characteristics. In research where only existing databases are used, these issues have not been dealt with. In real applications, speakers may speak to computers while walking about in a room and are not likely to use head-mounted microphones. The demands on speech recognition created by these needs have recently attracted researchers' attention. Measurement of characteristics, rec

ognition experiments, and algorithm improvement reveal the following facts.

*Distance Between Speaker and Microphone* As the distance between the speaker and the microphone increases, recognition performance tends to decrease because of degradation of low-frequency characteristics. Normalization of transmission characteristics by using a directional microphone or an array microphone has been found to be effective in compensating for these phenomena (Tobita et al., 1989).

*Reflection and Reverberation* It has been known that the interference between direct sound and sound reflected by a desk or a wall causes sharp dips in the frequency. Also, in a closed room the combination of various kinds of reflection causes reverberation. As application of speech recognition to conditions in a room or a car has become a key issue, these phenomena are attracting the attention of speech recognition researchers, and several studies have been done. Use of a directional microphone, adoption of an appropriate distance measure, or introduction of adaptive filtering are reported to be effective methods for preventing performance degradation (Tobita et al., 1990a; Yamada et al., 1991).

*Microphone Characteristics* Each microphone performs optimally under certain conditions. For example, the frequency of a close-range microphone flattens when the distance to the mouth is within several centimeters. Accordingly, using different microphones in the training mode and the recognition mode causes performance degradation (Acero and Stern, 1990; Tobita et al., 1990a). Several methods have been proposed to cope with this problem (Acero and Stern, 1991; Tobita et al., 1990b).

## Noise

As described earlier, all sounds other than the speech to be recognized should be considered noise. This means that there are many varieties of noise, from stationary noise, such as white noise, to environmental sounds such as telephone bells, door noise, or speech of other people. The method of compensating varies according to the phenomenon.

For high level noise such as car noise or cockpit noise, noise reduction at the input point is effective. A head-mounted noise-canceling microphone or a microphone with sharp directional characteristics is reported effective (Starks and Morgan, 1992; Viswanathan et

al., 1986). Also, several methods of using microphone arrays have been reported (Kaneda and Ohga, 1986; Silverman et al., 1992).

If an estimation of noise characteristics is possible by some means such as use of a second microphone set apart from the speaker, it is possible to reduce noise by calculating a transverse filter for the noise characteristics and applying this filter to the input signal (Nakadai and Sugamura, 1990; Powell et al., 1987).

An office is a good target for diversifying speech recognition applications. From the standpoint of robustness, however, an office does not provide satisfactory conditions for speech recognition. Offices are usually not noisy. Various kinds of sounds such as telephone bells or other voices overlap this rather calm environment, so these sounds tend to mask the speech to be recognized. Also, a desk-mounted microphone is preferable to a close-range microphone from the human interface standpoint. Use of a comb filter has been proposed for separating object speech from other speech (Nagabuchi, 1988).

## **Speaker Variation**

There are two types of speaker variation. One is an interspeaker variation caused by the differences among speakers between speech organs and differences in speaking manners. The other is intraspeaker variation. So far, various kinds of research related to interspeaker variations have been reported because this phenomena must be dealt with in order to achieve speaker-independent speech recognition. Intraspeaker variations, however, have been regarded as small noise overlapping speech and have been largely ignored. When it comes to applications, however, intraspeaker variation is an essential speech recognition factor.

Human utterances vary with the situation. Among these variations, mannerisms and the effects of tension, poor physical condition, or fatigue are difficult to control. Therefore, speech recognition systems must compensate for these variations.

There is a great practical need for recognition of speech under special conditions. Emergency systems, for example, that could distinguish words such as "Fire!" from all other voices or sounds would be very useful.

Several studies of intraspeaker variation have been undertaken. One typical intraspeaker variation is known as the "Lombard effect," which is speech variation caused by speaking under very noisy conditions (Roe, 1987). Also, in several studies utterances representing various kinds of speaking mannerisms were collected, and the HMM

was implemented to recognize these utterances (Lippmann et al., 1987; Miki et al., 1990).

## SPEECH TECHNOLOGY AND THE MARKET

As described in this paper, practical speech technologies have been developing rapidly. Applications of speech recognition and speech synthesis in the marketplace, however, have failed to keep pace with the potential. In the United States, for example, although the total voice-processing market at present is over \$1 billion, most of this is in voice-messaging services. The current size of the speech recognition market is only around \$100 million, although most market research in the 1980s forecasted that the market would soon reach \$1 billion (Nakatsu, 1989). And the situation is similar for speech synthesis. In this section the strategy for market expansion is described, with emphasis on speech recognition technology.

### Illusions About Speech Recognition Technology

In papers and surveys on speech recognition, statements such as the following have been used frequently:

"Speech is the most convenient method of communication for humans, and it is desirable to achieve oral communication between humans and computers."

"Speech recognition is now mature enough to be applied to real services."

These statements are basically correct. However, when combined, they are likely to give people the following impression:

"Speech recognition technology is mature enough to enable natural communication between computers and humans."

Of course, this is an illusion, and speech researchers or vendors should be careful not to give users this fallacious impression. The truth could be stated more precisely as follows:

- a. The capacity to communicate orally is a fundamental human capability, which is achieved through a long learning process that begins at birth. Therefore, although the technology for recognizing natural speech is advancing rapidly, there still exists a huge gap between human speech and the speech a computer is able to handle.
- b. Nevertheless, speech recognition technology has reached a level where, if applications are chosen appropriately, they can pro

vide a means for communication between computers and humans that—although maybe not natural—is at least acceptable.

### Strategy for Expanding the Market

Market studies carried out by the authors and others have identified the following keys to expansion of the speech recognition market, listed in descending order of importance (Nakatsu, 1989; Pleasant, 1989):

- *Applications and marketing.* New speech recognition applications must be discovered.
- *Performance.* Speech recognition algorithms must perform reliably even in real situations.
- *Capabilities.* Advanced recognition capabilities such as continuous speech recognition must be achieved.

Based on the above results and also on our experiences with developing and operating the ANSER system and service, the following is an outline of a strategy for widening the speech recognition market.

### Service Trials

Because practical application of speech recognition to real services is currently limited to word recognition, which is so different from how humans communicate orally, it is difficult for both users and vendors to discover appropriate new applications. Still it is necessary for vendors to try to offer various kinds of new services to users. Although many of them might fail, people would come to recognize the capabilities of speech recognition technology and would subsequently find application areas suited to speech recognition.

Telephone services might be the best, because they will offer speech recognition functions to many people and help them recognize the state of the art of speech recognition. Also, as pointed out earlier, the interesting concept of a "Speech OS" is worth trying.

### Robustness Research

It is important for speech recognition algorithms to be made more robust for real situations. Even word recognition with a small vocabulary, if it worked in the field as reliably as in the laboratory, would have great potential for application. It is delightful that re

cently the importance of robustness has attracted the attention of speech researchers and that various kinds of research are being undertaken, as described in the previous section.

One difficulty is that the research being done puts too much emphasis on stationary or quasi-stationary noise. There are tremendous variations of noise in real situations, and these real noises should be studied. Also, as stated before, intraspeaker speech variation is an important factor to which more attention should be paid.

### **Long Term Research**

At the same time, it is important to continue research of speech recognition to realize more natural human-machine communication based on natural conversation. This will be long-term research. However, as oral communication capability arises from the core of human intelligence, basic research should be continued systematically and steadily.

## **CONCLUSION**

This paper has briefly described the technologies related to speech recognition and speech synthesis from the standpoint of practical application.

First, systems technologies were described with reference to hardware technology and software technology. For hardware technology, along with the rapid progress of technology, a large amount of speech processing can be done by personal computer or workstation with or without additional hardware dedicated to speech processing. For software, on the other hand, the environment for software development has improved in recent years. Still further endeavor is necessary for vendors to pass these improvements on to end users so they can develop application software easily.

Then, several issues relating to the practical application of speech recognition and synthesis technologies were discussed. Speech databases for application and evaluation of these technologies were described. So far, because the range of applications of these technologies is limited, criteria for assessing the applications are not yet clear. Robustness of algorithms applied to field situations also was described. Various studies are being done concerning robustness.

Finally, reasons for the slow development of the speech recognition/synthesis market were discussed, and future directions for researchers and vendors to explore were proposed.

## REFERENCES

- Acero, A., and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," Proceedings of the IEEE ICASSP-90, pp. 849-852 (1990).
- Acero, A., and R. M. Stern, "Robust Speech Recognition by Normalization of the Acoustic Space," Proceedings of the IEEE ICASSP-91, pp. 893-896 (1991).
- Carre, R., et al., "The French Language Database: Defining and Recording a Large Database," Proceedings of the IEEE ICASSP-84, 42.10 (1984).
- Chang, H. M., "Field Performance Assessment for Voice Activated Dialing (VAD) Service," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IEEE (1992).
- Cole, R., et al., "A Telephone Speech Database of Spelled and Spoken Names," Proceedings of the International Conference on Spoken Language Processing, pp. 891-893 (1992).
- Dyer, S., and B. Harms, "Digital Signal Processing," Advances in Computers, Vol. 37, pp. 104-115 (1993).
- Fay, D. F., "Touch-Tone or Automatic Speech Recognition: Which Do People Prefer?," Proceedings of AVIOS '92, pp. 207-213, American Voice I/O Society (1992).
- Furui, S., "Toward Robust Speech Recognition Under Adverse Conditions," Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, pp. 3142 (1992).
- Gauvain, J. L., et al., "Design Considerations and Text Selection for BREF, a Large French Read-Speech Corpus," Proceedings of the International Conference on Spoken Language Processing, pp. 1097-1100 (1990).
- Hakoda, K., et al., "Sentence Speech Synthesis Based on CVC Speech Units and Evaluation of Its Speech Quality," Records of the Annual Meeting of the IEICE Japan, Tokyo, the Institute of Electronics Information and Communication Engineers (1986).
- Honma, S., and R. Nakatsu, "Dialogue Analysis for Continuous Speech Recognition," Record of the Annual Meeting of the Acoustical Society of Japan, pp. 105-106 (1987).
- Imamura, A., and Y. Suzuki, "Speaker-Independent Word Spotting and a Transputer-Based Implementation," Proceedings of the International Conference on Spoken Language Processing, pp. 537-540 (1990).
- Inmos, The Transputer Data Book, Inmos (1989).
- Itahashi, S., "Recent Speech Database Projects in Japan," Proceedings of the International Conference on Spoken Language Processing, pp. 1081-1084, (1990).
- Jacobs, T. E., et al., "Performance Trials of the Spain and United Kingdom Intelligent Network Automatic Speech Recognition Systems," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications IEEE (1992).
- Jankowski, C., et al., "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," Proceedings of the IEEE ICASSP-90, pp. 109-112 (1990).
- Kaneda, Y., and J. Ohga, "Adaptive Microphone-Array System for Noise Reduction," IEEE Trans. on Acoustics, Speech, and Signal Process., No. 6, pp. 1391-1400 (1986).
- Kimura, T., et al., "A Telephone Speech Recognition System Using Word Spotting Technique Based on Statistical Measure," Proceedings of the IEEE ICASSP-87, pp. 1175-1178 (1987).
- Kumagami, K., et al., "Objective Evaluation of User's Adaptation to Synthetic Speech by Rule," Technical Report SP89-68 of the Acoustical Society of Japan (1989).

- Lippmann, R., et al., "Multi-Style Training for Robust Isolated-Word Speech Recognition," Proceedings of the IEEE ICASSP-87, pp. 705-708 (1987).
- Miki, S., et al., "Speech Recognition Using Adaptation Methods to Speaking Style Variation," Technical Report SP90-19 of the Acoustical Society of Japan (1990).
- MADCOW (Multi-Site ATIS Data Collection Working Group), "Multi-Site Data Collection for a Spoken Language Corpus," Proceedings of the Speech and Natural Language Workshop, pp. 7-14, Morgan Kaufmann Publishers (1992).
- Nagabuchi, H., "Performance Improvement of Spoken Word Recognition System in Noisy Environment," Trans. of the IEICE Japan, No. 5, pp. 1100-1108, Tokyo, the Institute of Electronics, Information and Communication Engineers (1988).
- Nakadai, Y., and N. Sugamura, "A Speech Recognition Method for Noise Environments Using Dual Inputs," Proceedings of the International Conference on Spoken Language Processing, pp. 1141-1144 (1990).
- Nakatsu, R., "Speech Recognition Market-Comparison Between the US and Japan," Proceedings of the SPEECH TECH '89, pp. 4-7, New York, Media Dimensions Inc. (1989).
- Nakatsu, R., "ANSER: An Application of Speech Technology to the Japanese Banking Industry," IEEE Computer, Vol. 23, No. 8, pp. 43-48 (1990).
- Nakatsu, R., and N. Ishii, "Voice Response and Recognition System for Telephone Information Services," Proceedings of the SPEECH TECH '87, pp. 168-172, New York, Media Dimensions Inc. (1987).
- Neilsen, P. B., and G. B. Kristernsen, "Experience Gained in a Field Trial of a Speech Recognition Application over the Public Telephone Network," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IEEE (1992).
- Nomura, T., and R. Nakatsu, "Speaker-Independent Isolated Word Recognition for Telephone Voice Using Phoneme-Like Templates," Proceedings of the IEEE ICASSP86, pp. 2687-2690 (1986).
- Pallet, D., et al., "Speech Corpora Produced on CD-ROM Media by the National Institute of Standards and Technology (NIST)." Unpublished NIST document (1991).
- Patterson, D., and D. Ditzel, "The Case for the Reduced Instruction Set Computer," Computer Architecture News, Vol. 8, No. 6, pp. 25-33 (1980).
- Picone, J., "The Demographics of Speaker-Independent Digit Recognition," Proceedings of the IEEE ICASSP-90, pp. 105-108 (1990).
- Pittrelli, J. F., et al., "Development of Telephone-Speech Databases," Proceedings of the 1st Workshop on Interactive Voice Technology for Telecommunications Applications , IEEE (1992).
- Pleasant, B., "Voice Recognition Market: Hype or Hope?," Proceedings of the SPEECH TECH '89, pp. 2-3, New York, Media Dimensions Inc. (1989).
- Powell, G. A., et al., "Practical Adaptive Noise Reduction in the Aircraft Cockpit Environment," Proceedings of the IEEE ICASSP-87, pp. 173-176 (1987).
- Renner, T., "Dialogue Design and System Architecture for Voice-Controlled Telecommunication Applications," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Session IV (1992).
- Roe, D. B., "Speech Recognition with a Noise-Adapting Codebook," Proceedings of the IEEE ICASSP-87, pp. 1139-1142 (1987).
- Rosenbeck, P., "The Special Problems of Assessment and Data Collection Over the Telephone Network," Proceedings of the Workshop on Speech Recognition over the Telephone Line, European Cooperation in the Field of Scientific and Technical Research (1992).
- Rosenbeck, P., and B. Baungaard, "A Real-World Telephone Application: teleDialogue

- Experiment and Assessment," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IEEE (1992).
- Sato, H., et al., "Speech Synthesis and Recognition at Nippon Telegraph and Telephone," Speech Technology, pp. 52-58 (Feb./Mar. 1992).
- Silverman, H., et al., "Experimental Results of Baseline Speech Recognition Performance Using Input Acquired from a Linear Microphone Array," Proceedings of the Speech and Natural Language Workshop, pp. 285-290, Morgan Kaufmann Publishers (1992).
- Sprin, C., et al., "CNET Speech Recognition and Text-to-Speech in Telecommunications Applications," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IEEE (1992).
- Starks, D. R., and M. Morgan, "Integrating Speech Recognition into a Helicopter," Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, pp. 195-198 (1992).
- Steeneken, H. J. M., "Subjective and Objective Intelligibility Measures," Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, pp. 1-10 (1992).
- Texas Instruments, TMS320 Family Development Support, Texas Instruments (1992).
- Tobita, M., et al., "Effects of Acoustic Transmission Characteristics upon Word Recognition Performance," Proceedings of the IEEE Pacific Rim Conference on Communications, Computer and Signal Processing, pp. 631-634 (1989).
- Tobita, M., et al., "Effects of Reverberant Characteristics upon Word Recognition Performance," Technical Report SP90-20 of the Acoustical Society of Japan (1990a).
- Tobita, M., et al., "Improvement Methods for Effects of Acoustic Transmission Characteristics upon Word Recognition Performance," Trans. of the IEICE Japan, Vol. J73 D-II, No. 6, pp. 781-787, Tokyo, the Institute of Electronics, Information, and Communication Engineers (1990b).
- Viswanathan, V., et al., "Evaluation of Multisensor Speech Input for Speech Recognition in High Ambient Noise," Proceedings of the IEEE ICASSP-86 , pp. 85-88 (1986).
- Walker, G., and W. Millar, "Database Collection: Experience at British Telecom Research Laboratories," Proceedings of the ESCA Workshop 2, 10 (1989).
- Wilpon, J. G., et al., "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 38, No. 11, pp. 1870-1878 (1990).
- Yamada, H., et al., "Recovering of Broad Band Reverberant Speech Signal by Sub-Band MINT Method," Proceedings of the IEEE ICASSP-91, pp. 969-972 (1991).
- Yang, K. M., "A Network Simulator Design for Telephone Speech," Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IEEE (1992).

# User Interfaces for Voice Applications

*Candace Kamm*

## SUMMARY

As speech synthesis and speech recognition technologies improve, applications requiring more complex and more natural human-machine interactions are becoming feasible. A well-designed user interface is critical to the success of these applications. A carefully crafted user interface can overcome many of the limitations of current technology to produce a successful outcome from the user's point of view, even when the technology works imperfectly. With further technological improvements, the primary role of the user interface will gradually shift from a focus on adapting the user's input to fit the limitations of the technology to facilitating interactive dialogue between human and machine by recognizing and providing appropriate conversational cues. Among the factors that must be considered in designing voice interfaces are (a) the task requirements of the application, (b) the capabilities and limitations of the technology, and (c) the characteristics of the user population. This paper discusses how these factors influence user interface design and then describes components of user interfaces that can be used to facilitate efficient and effective human-machine voice-based interactions.

Voice interfaces provide an additional input and output modality for human-computer interactions, either as a component of a multimodal, multimedia system or when other input and output modalities are not available.

ties are occupied, unavailable, or not usable by the human (e.g., for users with visual or motor disabilities). One motivation for human-computer interaction by voice is that voice interfaces are considered "more natural" than other types of interfaces (e.g., keyboard, mouse, touch screen). That is, speech interfaces can provide a "look and feel" that is more like communication between humans. The underlying assumption is that by presenting this "more natural" interface to the user the system can take advantage of skills and expectations that the user has developed through everyday communicative experiences to create a more efficient and effective transfer of information between human and machine (Leiser, 1989).

A successful human-machine interaction, like a successful human-human interaction, is one that accomplishes the task at hand efficiently and easily from the human's perspective. However, current human-computer voice-based interactions do not yet match the richness, complexity, accuracy, or reliability achieved in most human-human interactions either for speech input [i.e., automatic speech recognition (ASR) or speech understanding] or for speech output (digitized or synthetic speech). This deficit is due only in part to imperfect speech technology. Equally important is the fact that, while current automated systems may contain sufficient domain knowledge about an application, they do not sufficiently incorporate other kinds of knowledge that facilitate collaborative interactions. Typically, an automated system is limited both in linguistic and conceptual knowledge. Furthermore, automated systems using voice interfaces also have an impoverished appreciation of conversational dynamics, including the use of prosodic cues to appropriately maintain turn taking and the use of confirmation protocols to establish coherence between the participants.

A well-designed voice interface can alleviate the effects of these deficiencies by structuring the interaction to maximize the probability of successfully accomplishing the task. Where technological limitations prohibit the use of natural conversational speech, the primary role of the interface is to induce the user to modify his/her behavior to fit the requirements of the technology. As voice technologies become capable of dealing with more natural input, the user interface will still be critical for facilitating the smooth flow of information between the user and the system by providing appropriate conversational cues and feedback. Well-designed user interfaces are essential to successful applications; a poor user interface can render a system unusable.

Designing an effective user interface for a voice application involves consideration of (a) the information requirements of the task,

(b) the limitations and capabilities of the voice technology, and (c) the expectations, expertise, and preferences of the user. By understanding these factors, the user interface designer can anticipate some of the difficulties and incompatibilities that will affect the success of the application and design the interaction to minimize their impact. For optimal results, user interface design must be an integral and early component in the overall design of a system. User interface design and implementation are most successful as an iterative process, with interfaces tested empirically on groups of representative users, then revised as deficiencies are detected and corrected, and then retested, until system performance is stable and satisfactory.

This paper discusses some of the aspects of task requirements, user expectations, and technological capabilities that influence the design of a voice interface and then identifies several components of user interfaces that are particularly critical in successful voice applications. Examples from several applications are provided to demonstrate how these components are used to produce effective voice interfaces.

## USER INTERFACE CONSIDERATIONS

By considering the interdependencies among the task demands of the application, the needs and expectations of the user population, and the capabilities of the technology, an interface designer can more accurately define a control flow for the application that will optimally guide the user through the interaction and handle the errors in communication in a way that facilitates task completion.

### Task Requirements

Current speech applications incorporate tasks ranging from very simple ones that require the system to recognize a user's single-word, binary-choice response to a single question (e.g., "Say yes if you will accept this collect call; say no if you won't") to more complex speech-understanding systems in which the user's input is a complex query that may address a task within a limited domain (e.g., "What times are the flights between Dallas and Boston on Monday?"). In addition, the interfaces to these applications can be highly structured, with either the computer or the user primarily responsible for controlling and directing the interaction, or more conversational, with the computer and the user frequently switching roles from the directive to the more passive participant in the dialogue. These applications differ in the amount and type of information that needs to be exchanged

during the interaction and also in the size and composition of the application vocabulary. In addition, applications may vary in the input and output modalities available to the user as well as in the costs of interaction failures.

## Information Elements

Clearly, the information requirements of the application are a central component in developing any user interface. For those applications that would otherwise occur as human-human dialogues, understanding how the dialogue typically progresses between humans can help identify the information that must be elicited from the user in order to accomplish the task. Studying the human-human interaction can also point out additional information that is likely to be provided by the user but that is either not critical to task completion or can be obtained from another source. The user interface designer can use this knowledge to determine a more natural order for eliciting the information elements and how much effort should be expended in eliciting each element (Mazor et al., 1992). The observation that users interact quite differently with machines than they do when interacting with a human agent argues against trying to model human-human dialogue precisely. However, analysis of the human-human interaction may help define the application's vocabulary, suggest appropriate wording for prompts, and identify the probable syntax of frequent responses.

Other applications do not have a human-human analog but rather were originally designed as automated systems using strictly nonvoice modalities. Although analysis of such systems can provide useful insights for designing voice interfaces, a strict conversion of the nonvoice interaction to the voice-based counterpart may not be optimal. For example, an automated bank-by-phone application using input from the telephone keypad may use an audio menu to direct the user to "For checking (account balance), press 1; for savings, press 2." Early implementations of voice interfaces for such systems directly translated this request into "For checking, say 1; for savings, say 2," rather than eliciting the more natural command words (e.g., "Do you want your account balance for checking or savings?"). Although the vocabulary choices in this application were probably the result of limitations in the availability of templates for speaker-independent ASR rather than poor interface design decisions, using vocabularies that are meaningful and memorable to the user should facilitate the efficiency of an interaction. Speech recognition systems that easily create application-specific speaker-independent vocabularies are becoming

more common, so the technological necessity for direct modeling of the vocabulary for voice-based interfaces on their key-based counterparts should soon disappear.

## Task Modalities

The set of input and output modalities available to the user affects the design of the user interface. In addition to voice, input modalities can include keyboard, mouse, touch-screen, or a telephone keypad. The availability of multiple input modalities allows the user to select the most efficient modality for entering input and also allows the system recourse to request an alternative input modality if it determines that the modality being used is not resulting in progress toward task completion. For example, if, in a telephone application, the presence of extremely high levels of background noise consistently results in incorrect recognition, the system may ask the user to switch to the telephone keypad so that the task can be completed successfully.

For output of information from the system to the user, either audio (speech) or textual modalities, or both, may be available and the appropriate modality depends on the task. For example, most voice dictation systems use textual output on a computer's monitor to provide concurrent feedback as the system recognizes words so that the user is made aware of errors or uncertainties in the system and can take corrective action. (Often, the corrective action may be taken using either voice or keyboard input, at the user's discretion.) Using voice output for feedback in this application would be highly ineffective, as it could severely tax the user's auditory memory, disrupt the user's train of thought, and possibly also interfere with reception of the user's speech by the system.

In cases where voice is the only modality available for output (as in most current telephone applications), the interface designer must take into account the temporal nature of speech output and its concomitant demands on user memory. For example, if voice menus are used, they should be constructed with only a very few options at each node, and users must be given a way to request repetition of information whenever they need it.

## Cost of Interaction Failures

Errors inevitably occur in human-computer interactions, just as they do in interactions between humans. These communication failures may be the result of the human providing unexpected or inap-

propriate input, the system misrecognizing or misinterpreting the user's speech, or a combination of the two. The costs of recognition or understanding errors can be direct (e.g., incorrect deposits in automated banking transactions or increased holding time in automated telephone services) or indirect (e.g., user dissatisfaction that leads to reduced use of a service or product). Depending on the cost of errors, the interface can be designed to be more or less conservative in its responses to input it finds ambiguous or incomplete. For example, in an information retrieval system, the cost of an error may be only the time spent making the query and hearing the incorrect answer, as long as the incorrect answer provides sufficient information for the user to determine that the original query was misunderstood. If a confirmatory subdialogue has a time cost similar to that of an incorrect response, it may be more efficient to permit the system to act on the user's voice input without first explicitly confirming that the request was accurately understood. In this case the system expects the user to be able to distinguish between inappropriate responses and correct responses and to inform the system of its error or reconstruct the query. On the other hand, if the application involves an action that, if performed incorrectly, has a high cost or greatly impedes progress toward task completion, it may be desirable to request explicit confirmation of the user's input prior to performing that action, or at least to provide feedback about the impending action and to permit the user to issue a command to cancel it. An example of this type might be a voice-activated dialing application, where a recognition error could result in charges for a phone call the user did not intend to make. For applications where the cost of errors is very high, systems should be designed to redirect the user to a nonspeech input modality or to a human agent in the event that it becomes apparent that repeated interchanges between the user and the system have not resulted in progress toward task completion.

## Technological Capabilities and Limitations

### Voice Input

Current speech applications demonstrate a wide range of voice input capabilities. The requirements of the application typically dictate whether a system uses (a) speaker-trained, speaker-adaptive, or speaker-independent recognition technology and (b) isolated-word, word-spotting, or continuous speech recognition. Prototype spoken-language-understanding systems with vocabularies of 300 to 1000 words can accept spontaneous continuous speech input without user enroll

ment (Appelt and Jackson, 1992; Pallett, 1992). These systems allow limited interactions in restricted domains, such as travel planning and providing directions, and currently have an understanding rate of about 70 percent (Cole et al., 1992). Several commercially available voice dictation systems incorporate speaker-dependent or speaker-adaptive isolated-word recognition technology for 5000- to 100,000-word vocabularies (Makhoul and Schwartz, in this volume). In contrast, most telephone applications have been restricted to isolated-word or word-spotting technology for small vocabularies in order to maintain acceptable recognition accuracy in the more adverse and variable electroacoustic environment of the public switched telephone network. If the user's input is restricted to the recognizer's vocabulary and speech style limitations, recognition performance in the latter applications can be quite high.

The capabilities of the voice input technology chosen for an application influence the priorities in the design of the user interface. As will be discussed in a later section, user expectations and needs may at times clash with the capabilities of the technology. To create a viable application, the short-term solution is to force the user to comply with the technology's requirements. For example, when the technology can accept only restricted, highly structured input, the user interface must focus primarily on guiding the user to speak in a style appropriate to the technology's limitations. The longer-term solution is to improve the technology to require less adaptation of the user's preferred behavior in accomplishing the task. As the technology allows more natural conversational input, the primary role of the user interface could shift from directing the user about how to speak toward providing more graceful resolution of ambiguities and errors.

## Voice Output

Two kinds of voice output technologies are available for voice interfaces: prerecorded (typically digitized) speech and synthetic speech. Prerecorded speech involves recording and storing natural speech for later combination and playback. Prerecorded speech has the quality of natural speech and, provided that care is taken to maintain natural prosody in combining recorded elements, it generally has more natural prosody than synthetic speech. If the voice output requirements of the application are known fully and are stable over time, prerecorded speech can be an appropriate technology for voice output.

However, if the voice output requirements are extensive and changeable, synthetic speech may be required. Speech synthesis technology

converts textual information to speech using sets of rules for converting letters to phonemes and for converting phonemes to acoustic events. As a result, messages can be constructed as they are needed. However, although several synthetic speech systems demonstrate high segmental intelligibility, they do not yet match the intelligibility or naturalness of human speech (Spiegel et al., 1988). Furthermore, there is evidence that perception of synthetic speech imposes greater processing demands on the listener than perception of natural speech (Luce et al., 1983). User interfaces that use synthetic speech for voice output should employ techniques aimed at facilitating listener comprehension. Providing messages that include syntactic and semantic context to aid comprehensibility and, where possible, using messages with relatively simple structure will reduce cognitive and memory demands on the user. In addition, applications with low-context messages (e.g., the delivery of proper names in a telephone reverse directory inquiry) should provide the user with a mechanism for requesting repetitions of the message and for the spelling of words that may be difficult for the user to understand.

### System Capabilities

Several system capabilities not directly related to speech technology per se have been shown to affect the success of voice interfaces. Some systems are incapable of "listening" to speech input while they simultaneously produce speech. With these systems it is impossible to interrupt a prompt, and the user must wait until the prompt is completed before responding. The user interface for these systems must be designed to detect and discourage violations of this protocol.

Systems that detect speech during prompts and truncate the prompt in response to user input provide more natural interfaces, but adding this capability may also lead to increased ambiguity regarding which prompt the user's input was intended to address. As a result, the user interface for a system with prompt-talk-through may need a more complex strategy for determining whether responses that interrupt prompts are delayed responses to prior prompts or anticipatory responses to the interrupted prompt.

System response time can influence user satisfaction (Shriberg et al., 1992). Interfaces to systems with slow response times for speech-and language-processing components can play interim messages (e.g., "Please wait. Your request is being processed.") to reduce perceived response time and increase user acceptance. With continuing advances in processor speed and in the efficiency of recognition search and

language-processing algorithms, near-real-time system response is becoming feasible even for complex speech-understanding tasks, so system response time may soon cease to be a significant interface issue.

In most applications the speech system is only a small component of a larger system. Constraints of this larger system can have a major impact on the user interface for an application. For example, in telephone network applications, system capabilities and resource limitations may dictate whether a customer can gain immediate access to a speech recognition subsystem when the customer picks up the telephone or whether a special code must be dialed to access the recognizer. Other system constraints that can influence the design of the user interface include whether a human agent is available in the event of recognition failure and whether relevant information and databases are available throughout the interaction or only during fixed portions.

## User Expectations and Expertise

### Conversational Speech Behaviors

Because users have so much experience with human-human speech interactions, it is not surprising that new users may expect a human-computer voice interface to allow the same conversational speech style that is used between humans. If the speech recognition technology used in an application cannot perform accurately with normal conversational speech input, the user is typically instructed to speak in the manner that the recognition system has been designed to accept. However, many speech behaviors are overlearned, and it is difficult for users to overcome these habits. Three such behaviors are (a) speaking in a continuous manner (i.e., without pausing between words), (b) anticipating responses and speaking at the same time as the other talker, and (c) interpreting pauses by the other talker as implicit exchange of turn and permission to speak.

*Continuous Speech* Speaking phrases with pauses between each word is unnatural, and the tendency to speak continuously is difficult to overcome, particularly when the user's attention is on completing a task. In an early telephone application that used isolated-word speech recognition technology, users were required to recite the seven-digit number of the telephone being used to place the call, pausing briefly after each digit. In one study 35 percent of the customers who spoke seven-digit numbers did not speak the digits in an isolated manner,

demonstrating the difficulty people have overcoming the natural behavior of speaking without interword pauses, even when they are instructed to do so (Wilpon, 1985). In another telephone study of digit recognition, only 37 percent of the customers responded with appropriate interword pauses (Yashchin et al., 1989).

In addition to the tendency to speak in a continuous manner, extraneous speech and hesitation sounds (like "uhh" or "um") are common in interactive dialogues. The inability to reliably coax humans to speak in isolation, without any nonvocabulary words or sounds, has been a driving force for the development of word spotting and continuous recognition algorithms. Word spotting allows accurate recognition of a small vocabulary in the presence of nonvocabulary utterances, even when they are modified by coarticulation effects introduced by continuous speech.

*Talk-over* In conversational settings, participants often begin to respond as soon as they understand the speaker's request. In an analysis of 50 conversations between telephone service representatives and customers, Karis and Dobroth (1991) observed that simultaneous speech occurred on 12.1 percent of the utterances of service representatives and 14.4 percent of the customers' utterances. This simultaneity rarely impairs the information flow in human-human dialogues, and, as a result, most users expect that their speech will be heard and understood even if it is spoken while the other participant in the dialogue is still talking. Many answering machine messages end with the admonition "Wait for the beep"—attesting to the difficulty people have conforming to a limited response interval.

*Conversational Dynamics* In a human-human dialogue, subtle cues carried primarily in the speech signal are used by the participants to indicate that they want to take or release control of the speaking role or to confirm that they understand or have been understood by the other party. Violating these cues can cause difficulties in user interfaces. Turn-taking permission is often conveyed by prosodic pauses. In an early implementation of an automated billing service for collect calls, customers were responding prematurely (i.e., before the system was ready to accept speech input) but consistently at a particular phrase break (i.e., brief pause) during the system's prompt (Bossemeyer and Schwab, 1990). The pause at the phrase break was interpreted by these customers as an invitation to take the floor and respond, and so they did. The pause duration in the prompt was shortened, and the resultant user interface was more successful, as users became much more likely to wait until the appropriate response window to re-

spond. Alternatively, a user interface may take advantage of the knowledge that turn taking is likely to occur during pauses, as does the interface described by Balentine and Scott (1992). In this system the recognizer is activated during pauses between items, in case the user responds immediately after hearing the menu item he/she wants. If the recognized word matches the menu item presented immediately before the user's input, the system proceeds to that selection. If the recognized word is ambiguous, the system asks for confirmation of whether the menu item presented before the interruption is the desired selection.

Cues for confirmation or signaling of potential errors are often provided by repeating the information to be confirmed, along with specific intonation patterns. For example, falling intonation can imply acknowledgment, while rising intonation of the confirmatory phrase signals uncertainty and potential error (Karis and Dobroth, 1991). Currently, such intonational cues are not recognized reliably by automated systems, but advances in understanding the contribution of intonation patterns to dialogue flow will provide a powerful enhancement to user interfaces, not only for confirmation and error recovery but also for detecting user hesitations and repairs (i.e., self-corrections a user makes within an utterance). A better understanding of how intonation is used to convey intent would also be quite useful in constructing more prosodically appropriate messages for voice output (Karis and Dobroth, 1991).

*Novice vs. Expert Users* Users have different expectations and needs depending on their experience with a system. First-time or infrequent users are likely to require instructions and/or guidance through a system to allow them to build a cognitive model of how the system works and how to interact with it. In contrast, experienced users who interact frequently with the system want ways to bypass the instructions and move through the interaction more efficiently. As humans become more experienced with the specific requirements of a human-human collaborative dialogue, they begin to increase the efficiency of the interaction by abbreviating portions of the dialogue to the minimum sufficient information necessary to accomplish the task (Leiser, 1989). For example, confirmation utterances become shorter or occur less frequently, and unnecessary function words may be omitted (e.g., "There are 48 pages in volume I, 42 in volume II, 36 in III"). Successful user interfaces for automated systems should accommodate both the needs of novices and the preferences of expert users.

## USER INTERFACE DESIGN STRATEGIES

Once the interface requirements for the application have been identified through the analysis of task demands, user characteristics, and technological factors, there are several aspects of user interfaces that have often proved effective both for alleviating system limitations and for improving an application's usability. These include dialogue flow strategies using appropriately directive and informative prompts and the use of subdialogue modules to provide access to instructions, to confirm that the user's input has been correctly recognized, and to detect errors and recover from them.

To demonstrate how far astray an interaction can go without these features, [Table 1](#) shows a transcript of a human-computer dialogue from a very early application of ASR. The application used isolated-word speaker-independent recognition technology. Employees of the company that implemented this application called the system and spoke an authorization code. The system was supposed to recognize the spoken codes, check the access privileges associated with that code, and either permit or prohibit the placement of long-distance telephone calls, based on the user's access privileges. In reviewing the transcript in [Table 1](#) (and those of other similarly unsuccessful interactions), it is apparent that the users did not know how to use the system. (In one interaction a user even asks "What am I supposed to do?") If users received written instructions to speak the digits of the authorization code with pauses between each digits, they either had not read them or did not remember them. On dialing into the system, the user heard the synthesized message "Authorization number please." The user began to comply by speaking a continuous string of numbers in response to this prompt. The system interrupted her and asked her to repeat, which she did, again speaking in a continuous manner but using a slightly louder voice. However, the problem with her initial utterance was not her speaking level but rather the fact that the system could not handle continuously spoken digits. (In addition, whether the words "sixty-one" and "hundred" were in the system's vocabulary is unclear.) In response to the user's repetition, the system spoke the number "four," presumably as a confirmation request for an erroneously recognized digit. When it became apparent to the system that the continuous input was not likely to be a digit, the system asked for another repetition. The user then responded with a reasonable question but, again, not an utterance the system was able to handle. At this point the user hung up. Other transcripts from this application demonstrated that the system had two modes of completing an interaction: one that gave the response "Thank you"

and the other issuing the word "Sorry." The latter occurred only after a nonproductive series of iterations through a loop consisting of an inappropriate input from the user, followed by another "Please repeat" prompt from the system.

TABLE 1 Transcript of an Unsuccessful Voice Interaction

System:	Authorization number, please.
User:	231 sixty-wa
System:	please repeat
User:	(speaking more loudly) 231 sixty-one hundred
System:	four (pause) please repeat
User:	how many times do you want me to repeat it? [hangs up]

What are the major problems with this user interface? First, the users seem not to be aware of what they are expected to do. This could be alleviated by using a directive prompt explicitly requesting that the digits be spoken with pauses between them. Second, when the system discovers a problem in the input (because it fails to recognize any digit with high certainty), the error message it reports ("Please repeat") does not provide the user with any indication of what the problem might be. Even though the system may not be certain about the source of the problem, it might be possible to make a reasonable guess based on the length of the input utterance. If the length of the utterance that is detected is significantly longer than would be probable if the input utterance was a single digit, the error message might instruct the user to speak the digits one at a time, or even explicitly say "Please say the next digit of the authorization number now." Third, the confirmation messages themselves were confusing, particularly when the recognized digit was not a number in the user's input utterance. It is not clear that the user realized that the word "four" was an attempt at confirmation, and it also was not apparent whether the user had any capability of alerting the system to an incorrect recognition. The confirmation prompt could have been directive and informative—for example, "Was the first digit four?", aimed at eliciting a yes or no answer. However, requesting confirmation after every digit results in a very tedious interaction. If the recognition performance is good enough, it would be better to reduce the frequency of confirmation requests by chunking confirmation messages to contain sets of three or four digits at a time. Finally, the user's only recourse to difficulties with the system seems to be to hang up. The user is apparently afforded no other control of the interaction. All of these suggestions will make the interaction lengthier, but at least the user would have a better chance of understanding how to use the system and of accomplishing the task. (See [Table 2](#) for a

sample revised interaction conforming to the technology limitations of isolated-word recognition.) The transcript in [Table 1](#) shows how a user interface that does not incorporate instructions, confirmation, and error recovery can render an application virtually unusable. The importance of these elements is discussed in more detail in the following sections, using examples that demonstrate their use in more successful voice interfaces.

TABLE 2 Suggested Interface Revisions

System:	Authorization number, please.
User:	two three one sixty-wa
System:	[detects unusually long utterance] Please speak the digits of the authorization number, pausing after each digit.
User:	two [pause] three [pause] one [pause] six [pause] one [pause] oh [pause] oh [pause]
System:	I heard two three nine six one oh oh. Is that correct?
User:	No, two three <i>one</i> .
System:	[detects unusually long utterance, assumes negative response] Please answer yes or no. Is the first digit two?
User:	Yes.
System:	Is the second digit three?
User:	Yes.
System:	Is the third digit nine?
User:	No.
System:	Please say the third digit again.
User:	One.
System:	Is the third digit one?
User:	Yes.
System:	[even if the recognizer still misrecognizes the input as "nine," the system does not accept nine as a candidate and selects the next best choice.] Is the authorization number two three one six one oh oh?
User:	Yes.
System:	Checking authorization number two three one six one oh oh.
System:	Authorization number accepted. You may place your call now.

### Dialogue Flow

Underlying every user interface is a plan for obtaining the information elements necessary to complete the task. In more structured interactions the order of acquiring these elements may be explicitly defined, whereas in more conversational interfaces the order in which the information is gathered may be controlled more by the user. Directive prompts that explicitly instruct a user are useful for guiding an interaction. They are especially critical when the technology requires input that is restricted (e.g., to isolated-word input, to a small vocabulary, or to a circumscribed time interval), but they are also

useful for resolving ambiguities in systems where more natural input is permitted.

TABLE 3 Effect of Prompt Wording on Customer Responses

Customer Response	Proportion of Responses (%)	
	Prompt: Will you accept the call?	Prompt: Say yes if you accept the call; otherwise, say no.
Isolated yes or no	54.5	80.8
Multiword yes or no	24.2	5.7
Other affirmative or negative	10.7	3.4
Inappropriate response (e.g., "She's not here," "It's for you")	10.4	10.2

Directive prompts can significantly increase user compliance to system restrictions. As shown in Table 3, a Bellcore study of customer responses to alternate billing prompts demonstrated that the percentage of acceptable inputs to an isolated-word recognition system with a two-word vocabulary (the words "yes" and "no") increased from about 55 percent to 81 percent when the prompt was changed from "Will you accept the charges (for a collect call)?" to "Say yes if you will accept the call; otherwise, say no." Other studies of similar applications have demonstrated similar prompt-related effects (Bossemeyer and Schwab, 1990; Spitz et al., 1991).

Directive prompts can also be used after ambiguous or error conditions to attempt to tell the user how to modify the next input to improve the chances of a successful input. For example, if the system (a) does not allow prompt talk-through, (b) detects high energy at the beginning of the input to the recognizer, and (c) has low certainty about the accuracy of the recognized word or phrase, the system might assume that the user began speaking before the prompt was over and would reprompt the user for the input, augmenting the reprompt with "Please speak *after* the beep" (with stress on the word "*after*").<sup>1</sup>

<sup>1</sup> In fact, this strategy has been implemented in a voice-activated dialing application developed by NYNEX.

### Feedback and Confirmation

One way to limit erroneous actions by the system is to give the user feedback about the application's state and to request confirmation that the system's interpretation of the input or impending action on that input is what the user intended. However, providing feedback and eliciting confirmation on each piece of information exchanged between the user and the system often results in a lengthy, tedious, and inefficient interaction. As mentioned above, if the cost of errors is minimal, and if the user is provided sufficient information to ascertain that the system's response was appropriate, it may be reasonable to forgo many of the confirmation interchanges for some applications. For example, in a research prototype stock quotation service developed at Bell Northern Research (Lennig et al., 1992), the system immediately searches the stock database and provides the user with a stock quotation for any input utterance (except utterances that are recognized as command phrases in the application). If the system retrieves a different stock than the user requested, the user can say "Incorrect stock" to retrieve stock market information for the stock that was the next best match to the input utterance, or, if the user is not aware of that option, the user can repeat the stock name. This strategy provides a way for the user to take control to indicate errors and backtrack to a stable point in the dialogue.

Many applications issue requests for confirmation only when there is some level of ambiguity about the input (e.g., when the most likely recognition candidate has a relatively low likelihood or certainty or when an input utterance to a spoken language system cannot be parsed). In these cases, whenever possible, the confirmation request should not be a simple request for repetition of the input utterance, since that utterance has a history of not being handled successfully. Rather, the confirmation request should be a directive prompt that attempts to resolve the ambiguity and progress toward task completion. One strategy is to request confirmation by phrasing the request as a yes/no or forced-choice question about the accuracy of the best guess of the recognized utterance. Katai et al. (1991) observed that, even when the yes/no confirmation strategy is used, users often respond with a repeat of the command, suggesting that it might be useful to extend the response set for the confirmation question to include the particular word or phrase that is being confirmed.

Providing feedback for confirmation is also critical in interfaces that do not provide voice output. For example, some voice dictation systems provide the user with a short list of the most likely word candidates, so that the user can indicate the correct alternative by

identifying its position in the list or by spelling it. In any application the user interface must allow the user to initiate corrective actions.

### Instructions

To use a system efficiently and effectively, the user must know what is expected by the system and what information the system can provide. For simple tasks the instructions to the user may consist of a single prompt. For more complicated interactions with a variety of options, the instructions may be quite lengthy and detailed. Novice users will need more help, and experienced users may not need any guidance to use the system (unless an error occurs). An optimal user interface should provide the option of accessing instructions or may automatically offer instructions if the user's behavior suggests that they are needed. An example of the latter strategy is implemented in a prototype voice-activated directory service used at Bolt, Baranek and Newman. In this system no initial prompt is given when the user is connected to the system. Expert users know to say the first and last names of the person they want to reach. However, if the system does not detect the onset of speech within 2 seconds, it provides a brief prompt instructing the user (J. Makhoul, BBN, personal communication, 1992).

The Bell Northern Research stock quotation application mentioned above (Lennig et al. 1992) has extensive instructions describing the capabilities of the system. The brief welcoming prompt to this service tells the user to say the name of the stock or "Instructions" to hear the instructions. Thus, the expert user can bypass the instructions and progress directly to the primary task. New users, however, are provided with the opportunity to access the instructions. At the beginning of the instructions, the user is told how to exit the instructions at any time and return to the initial prompt. This informs the user that he/she can gain control of the interaction at any time and how that can be accomplished.

### Error Recovery

Since errors are inevitable, error recovery procedures are also an inevitable requirement in a user interface. The goal of error recovery procedures is to prevent the complete breakdown of the system into an unstable or repetitive state that precludes making progress toward task completion. Obviously, error recovery requires the cooperation of the user, and both the system and the user must be able to

initiate error recovery sequences. The feedback and confirmation dialogues described previously are the first step at detecting errors.

In the Bell Northern Research stock quotation application (Lennig et al., 1992), the user initiates error correction by saying "Incorrect stock." The system then provides the stock information for the next best match to the original input and asks the user if that is the correct stock. The system will propose a few alternate candidates and, if the user confirms none of them, will ask the user to repeat the stock name. If no new candidates are highly likely, the system suggests that the user check to be sure the requested stock is really offered on that stock exchange.

In other applications the system flags the error or uncertainty and initiates error recovery. For example, in the voice dictation task, the system signals recognition uncertainty by taking control of the interaction and suggesting a set of possible alternatives for the user to confirm. By requiring the user to indicate the correct alternative by providing its item number on the list or by spelling it, the error recovery module uses a much more restricted vocabulary than the general dictation task to increase the likelihood of correct recognition of the user's corrective response.

When provided with error messages, users do modify their subsequent behavior. In a Wizard of Oz study where a human listener entered accurate transcripts of the user's speech to the natural language component of a spoken-language-understanding system for travel planning, Hunnicutt et al. (1992) observed that users were able to make use of the system's error messages to rephrase their queries and reliably recovered from errors about 80 percent of the time. Additional studies by the same authors showed that, in a fully automated system that occasionally hypothesized incorrect utterances, users were generally able to identify misleading responses from the system and successfully recovered from them, even when system error messages were not very specific. These results suggest that, if a system can make a reasonable assumption about the probable cause of the error, directive error messages (e.g., "I didn't hear your response. Please speak louder" after a system time-out) are useful. In addition, even if the system's assumption is wrong (e.g., the user didn't speak too softly but rather not at all), the user generally has sufficient information to be able to disregard inappropriate messages arising from the system's incorrect hypotheses.

Finally, some voice interactions will not succeed, no matter how clever the error detection and recovery strategies. In these cases the user interface should inform the user of the problem and its probable cause and direct the user to an alternative solution—for example, a

human agent, if there is one available, or an alternate input modality that may be more reliable. As a general principle, user interfaces should also allow the user to initiate a graceful exit from the application at any stage in the interaction.

## EVALUATING TECHNOLOGY READINESS

This paper has stressed the interdependencies among application requirements, user needs, and system capabilities in designing a successful voice interface. Often, voice interface developers are asked whether speech technology is "good enough" or "ready" to accomplish a particular task. Although the query is stated simply, a thorough answer must extend beyond the focus on the technology alone to address other questions that include:

1. Would a typical user judge the automated interaction to be satisfactory—that is, is the user successful at completing the task, and is the interaction easy and efficient from the user's viewpoint?
2. Is the degree of successful automation sufficient to make its use financially attractive for the application provider? Do the benefits and savings accrued from the automation effort offset the costs of implementation?

The "readiness" question should encompass not only the technological capabilities but also the context of the specific application, its users, and the entire system within which the application is instantiated. The optimal way to estimate "readiness" is in field trials of the application, using prototype systems and users who represent the population that the application is intended to serve. Performance measures for assessing the success of such field trials would include task completion rates and user reactions. If a field trial demonstrates that the application is successful for the user, a cost/benefit analysis can then be applied to determine whether the application is "ready" for deployment (i.e., is economically justifiable).

## CONCLUSION

Whether the voice application is relatively simple (e.g., recognizing only a limited set of key words) or quite complex (e.g., speech-understanding systems that permit more natural conversational interactions), a well-designed user interface will always be an essential element in successful human-machine communication by voice. The designer of a voice interface must consider the interrelationships among the task requirements, technological capabilities, and user expecta

tions in implementing an interface that will facilitate "ease of use" and naturalness of a human-machine interaction. Voice interfaces can be made less brittle by incorporating directive prompts in the main task dialogue as well as in confirmation requests and by providing user control to access instructions, to indicate and correct errors, and to exit the system as necessary. Iterative testing with representative users is necessary in developing a robust voice interface. With improved understanding of how prosody and message content cue conversational dynamics for confirmation, error recovery, hesitations, and repair, and with the development of reliable techniques for automatically recognizing and providing these cues, voice interfaces can begin to realize the promise of a "more natural" interface for human-computer interactions.

### ACKNOWLEDGMENTS

The author would like to thank Mary Jo Altom, Susan Dumais, Dan Kahn, and Sharad Singhal for helpful comments on early drafts of this paper.

### REFERENCES

- Appelt, D., and E. Jackson. "SRI International February 1992 ATIS benchmark test results." *Proceedings of the DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, San Mateo, Calif., 1992.
- Balentine, B. E., and B. L. Scott. "Goal orientation and adaptivity in a spoken human interface." *Journal of the American Voice I/O Society*, 11:46-60, 1992.
- Bossemeyer, R. W., and E. C. Schwab, Automated alternate billing services at Ameritech. *Journal of the American Voice I/O Society*, 7:47-53, 1990.
- Cole, R., L. Hirschman, et al. Workshop on Spoken Language Understanding. Oregon Graduate Institute Technical Report No. CS/E 92-014, Oregon Graduate Institute, Beaverton, Oregon, 1992.
- Hunnicutt, S., L. Hirschman, J. Polifroni, and S. Seneff. "Analysis of the effectiveness of system error messages in a human-machine travel planning task." *Proceedings of the International Conference on Spoken Language Processing '92*, pp. 196-200, University of Alberta, Edmonton, Alberta, Canada, 1992.
- Karis, D., and K. M. Dobroth. "Automating services with speech recognition over the public switched telephone network: Human factors considerations." *IEEE Journal on Selected Areas in Communications*, 9:574-585, 1991.
- Katai, M., A. Imamura, and Y. Suzuki. "Voice activated interaction system based on HMM-based speaker-independent word-spotting." Paper presented at American Voice I/O Society '91, Atlanta, Ga., 1991.
- Leiser, R. G. "Improving natural language and speech interfaces by the use of metalinguistic phenomena." *Applied Ergonomics*, 20:168-173, 1989.
- Lennig, M., D. Sharp, P. Kenny, V. Gupta, and K. Precoda. "Flexible vocabulary recognition of speech." *Proceedings of the International Conference on Spoken Lan*

- guage Processing, University of Alberta, Edmonton, Alberta, Canada, pp. 93-96, 1992.
- Luce, P., T. Fuestel, and D. Pisoni. "Capacity demands in short term memory for synthetic and natural speech." *Human Factors*, 25:17-32, 1983.
- Mazor, B., B. Zeigler, and J. Braun. "Automation of conversational interfaces." *Proceedings of the American Voice I/O Society*, San Jose, Calif., 1992.
- Pallett, D. "ATIS benchmarks." *Proceedings of the DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, San Mateo, Calif., 1992.
- Shriberg, E., E. Wade, and P. Price. "Human-machine problem solving using Spoken Language Systems (SLS): Factors affecting performance and user satisfaction." *Proceedings of the DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, San Mateo, 1992.
- Spiegel, M. F., M. J. Altom, M. J. Macchi, and K. L. Wallace. "Using a monosyllabic test corpus to evaluation the intelligibility of synthesized and natural speech." *Proceedings of the American Voice I/O Society*, San Jose, Calif., 1988.
- Spitz, J. and the Artificial Intelligence Speech Technology Group. "Collection and analysis of data from real users: Implications for speech recognition/understanding systems." *Proceedings of the DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, San Mateo, Calif., 1991.
- Wilpon, J. G. "A study on the ability to automatically recognize telephone-quality speech from large customer populations." *AT&T Technical Journal*, 64:423-451, 1985.
- Yashchin, D., S. Basson, N. Lauritzen, S. Levas, A. Loring, and J. Rubin-Spitz. "Performance of speech recognition devices: Evaluating speech produced over the telephone network." *Proc. of the IEEE ICASSP*, Piscataway, N.J., pp. 552-555, 1989.

## **TECHNOLOGY IN 2001**



# Speech Technology in the Year 2001

*Stephen E. Levinson and Frank Fallside*

## SUMMARY

This paper introduces the session "Technology in the Year 2001" and is the first of four papers dealing with the future of human-machine communication by voice. In looking to the future it is important to recognize both the difficulties of technological forecasting and the frailties of the technology as it exists today—frailties that are manifestations of our limited scientific understanding of human cognition. The technology to realize truly advanced applications does not yet exist and cannot be supported by our presently incomplete science of speech. To achieve this long-term goal, the authors advocate a fundamental research program using a cybernetic approach substantially different from more conventional synthetic approaches. In a cybernetic approach, feedback control systems will allow a machine to adapt to a linguistically rich environment using reinforcement learning.

## INTRODUCTION

The title of this session is "Technology in the Year 2001." This colloquium has discussed a number of the state-of-the-art issues: the scientific bases of human-machine communication by voice; the three

technologies, recognition, synthesis, and natural language understanding; and, finally, the applications of this technology.

When the blueprint for this session was fitted together this session was called "Future Technology." The organizers felt that we should think really about it in a very "blue sky" sort of way. I was alarmed by the project altogether at that stage, rushed back home, and started reading about Leonardo da Vinci, H. G. Wells, and dreamed up a few impossible applications for speech recognition. During these ruminations, I thought, there are many interesting things we could discover—how to navigate the oceans of the world safely or, possibly, information about the location of treasure ships lost by the Spanish many years ago. I am sure that squids and other marine animals could tell us a great deal about that. There is also the question of HAL or Blade Runner, Ed Newbard, and old Napoleon Solo who used to ask for channel D. However, after some discussion with the speakers today, they indicated they did not want this sort of stuff at all.

It was decided that we should talk about evolutionary technology—rather than revolutionary technology. So we are talking about what is likely to be possible in the year 2001. In passing, we might note that the ideas of some of our predictions are not all that far away. We have rough models of HAL right now; of Blade Runner, I'm less certain.

However, we have put together a very interesting program for this last session. Certainly, the three speakers are eminently suited to this. They have all made significant contributions to the state of the art in several areas. One of the things we decided to do was to change the order slightly so that Sadaoki Furui will talk first about ultimate synthesis/recognition systems to give us a flavor of his view of the systems that are likely to be available. And then our two other experts will discuss research directions—B. Atal, in the area of speech, and M. Marcus in the area of natural language.

The paragraphs above are a slightly edited version of an audio recording of Frank Fallside's introduction of this session of the colloquium. They are included here for two reasons. First, they capture rather well Frank's persona. As I read them, I can hear his enunciation of the words in his marvelous accent and diction, which ever so slightly betrayed the intended intellectual mischief. Second, of course, is the intellectual mischief itself. What Frank was saying was that predicting the future of technology is fraught with danger and is thus best approached with a bit of self-deprecating humor.

Before exploring that idea further, it is worthwhile to make a few observations about the views of the speakers in this session. There is no need to summarize the material as the papers are presented in

this volume in their entirety. It is interesting, however, to note some common themes.

First, the three speakers recognize the difficulty of technological forecasting and thus do not fix any of their predictions or research programs to any specific date, not even the year 2001 of the session title. Both Atal and Furui use human performance as an important benchmark for assessing progress. The importance of this measure is discussed in Session III, Speech Recognition Technology (Levinson, in this volume). Atal lists specific problems to which the present lack of a solution is an indication of gaps in our scientific understanding of spoken language. Included in his list are learning, adaptation, synthetic voice quality, and semantics. He suggests that some of these problems might be addressed by finding new, more faithful mathematical representations of the acoustic signal.

Furui points to other inadequacies such as poor multivoice and multilingual capabilities as indicative of a fundamental lack of understanding of speech. He suggests that combining recognition and synthesis in applications might be of help. As we shall note later, the closed recognition/synthesis loop is a very powerful tool that is central to Fallside's ambitious research program.

The presentation by Marcus is somewhat different from the two preceding it in the sense that it deals with the specific technical problem of statistical/structural models of language. However, indirectly, he addresses two of the same problems discussed by Atal and Furui. First, his statistical approach aims at the problem of meaning because it is a syntactico-semantic theory in which the semantics derives from lexical cooccurrences in specific syntactic structures. It also bears on the problem of learning in the sense that these complex models must be trained on (i.e., must be learned from) large linguistic corpora.

Thus, although never explicitly stated, the thrust of all three presentations is a clear call for fundamental research to resolve some of the critical questions surrounding speech communication. As such, these papers stand in direct opposition to the sentiments expressed in the session on speech technology to the effect that there are no fundamental impediments to the application of speech technology. To some extent, Atal, and to a greater extent Furui, envision beneficial applications of a mature speech technology. But their call for fundamental research is an admission that the technology to realize these applications does not yet exist and cannot be supported by a presently incomplete science of speech.

After the three aforementioned presentations, session chairman Frank Fallside opened the session for general discussion. There was an enthusiastic response from the attendees mostly in the form of

technical comments related to the subject matter of the presentations rather than their long-term implications. The chairman did not try to steer the discussion toward the more philosophical aspects of the presentations even though his opening remarks were of a decidedly philosophical tone. Nor did he choose to appropriate any of the discussion period to report on his own research program even though it is aimed squarely at solving some of the fundamental problems raised by the session's speakers. In retrospect it is a pity he did not do so, although such action on his part would have been out of character, because he died shortly after the colloquium having deliberately relinquished an opportunity to make his ideas more widely known.

However, Fallside's approach to speech communication is clearly set forth, if only in conceptual form, in his keynote lecture at the 1991 Eurospeech Conference (Fallside, 1991). The insight upon which his research program is based is that speech communication in humans is an acquired skill involving the simultaneous learning of both perception and generation. Therefore, he argues, a mechanical system should do likewise by forming a closed loop system of analysis and synthesis components and allowing it to adapt to a linguistic environment.

Fallside treats only the linguistic aspects of speech communication. Whereas in a similar spirit but quite independently, Levinson (1989) argues that the entire sensory-motor periphery is required for humans to fully develop their cognitive function. As did Fallside, Levinson suggests that this behavior can be simulated with a feedback-controlled robot that interacts with a natural environment in the presence of a cooperative teacher. This idea has been explored experimentally by Gorin et al. (1991) and Sankar and Gorin (1993).

Whether or not these two hypotheses have any value remains to be seen. They do, however, share two important features. First, they are cybernetic rather than synthetic approaches, and second, they are unconventional, highly speculative, and not presently feasible.

All present approaches to speech communication are synthetic—that is, they advocate that we should first figure out, by any means available, how spoken language works. We should then capture that process in a mathematical model and finally implement the model in a computer program. By contrast, the cybernetic approach says we should use feedback control systems to allow a machine to adapt to a linguistically rich environment using reinforcement learning. This approach requires only limited *a priori* understanding of the linguistic phenomena under study.

The boldness (many would say foolishness) of cybernetic organic approaches is actually appropriate to the magnitude of the task we

have set for ourselves. It must be realized that the quest to build a machine with human-like linguistic abilities is tantamount to simulating the human mind. This is, of course, an age-old philosophical quest, the rationality of which has been debated by thinkers of every generation. If the problem of simulating the mind is intractable, we shall develop a speech technology that is little more than a curiosity with some limited commercial value. If, however, the problem admits of a solution, as I believe it does, the resulting technology will be of historic proportions.

Frank Fallside did not live to see his research program carried out. That program might well turn out to be an important component in the accomplishment of the ultimate goal of speech research, to build a machine that is indistinguishable from a human in its ability to communicate in natural spoken language. Frank Fallside will never see such a machine. Sadly, the same is most likely true for this colloquium's participants. However, I believe the ultimate goal can be accomplished. I only hope that our intellectual descendants who finally solve the problem do not wonder why we were so conservative in our thinking, thus leaving the breakthrough to be made by a much later generation.

## REFERENCES

- Fallside, F., "On the Acquisition of Speech by Machines, ASM," Proc. Eurospeech 91, Genoa, Italy, 1991.
- Gorin, A. L., et. al., "Adaptive Acquisition of Language," Computer Speech and Language 5 (2):101-132, 1991.
- Levinson, S. E., "Implication of an Early Experiment in Speech Understanding," Proceedings of the AI Symposium, pp. 36-37, Stanford, Calif., 1989.
- Sankar, A., and A. L. Gorin, "Visual Focus of Attention in Adaptive Language Acquisition," Neural Networks for Speech and Vision Applications, R. Mammone, Ed., Chapman and Hall, 1993.

# Toward the Ultimate Synthesis/Recognition System

*Sadaoki Furui*

## SUMMARY

This paper predicts speech synthesis, speech recognition, and speaker recognition technology for the year 2001, and it describes the most important research problems to be solved in order to arrive at these ultimate synthesis and recognition systems. The problems for speech synthesis include natural and intelligible voice production, prosody control based on meaning, capability of controlling synthesized voice quality and choosing individual speaking style, multilingual and multidialectal synthesis, choice of application-oriented speaking styles, capability of adding emotion, and synthesis from concepts. The problems for speech recognition include robust recognition against speech variations, adaptation/normalization to variations due to environmental conditions and speakers, automatic knowledge acquisition for acoustic and linguistic modeling, spontaneous speech recognition, naturalness and ease of human-machine interaction, and recognition of emotion. The problems for speaker recognition are similar to those for speech recognition. The research topics related to all these techniques include the use of articulatory and perceptual constraints and evaluation methods for measuring the quality of technology and systems.

## VISION OF THE FUTURE

For the majority of humankind, speech production and understanding are quite natural and unconsciously acquired processes performed quickly and effectively throughout our daily lives. By the year 2001, speech synthesis and recognition systems are expected to play important roles in advanced user-friendly human-machine interfaces (Wilpon, in this volume). Speech recognition systems include not only those that recognize messages but also those that recognize the identity of the speaker. Services using these systems will include database access and management, various order-made services, dictation and editing, electronic secretarial assistance, robots (e.g., the computer HAL in *2001—A Space Odyssey*), automatic interpreting (translating) telephony, security control, and aids for the handicapped (e.g., reading aids for the blind and speaking aids for the vocally handicapped) (Levitt, in this volume). Today, many people in developed countries are employed to sit at computer terminals wearing telephone headsets and transfer information from callers to computer systems (databases) and vice versa (information and transaction services). According to the basic idea that boring and repetitive tasks done by human beings should be taken over by machines, these information-transfer workers should be replaced by speech recognition and synthesis machines. Dictation or voice typewriting is expected to increase the speed of input to computers and to allow many operations to be carried out without hand or eye movements that distract attention from the task on the display.

Figure 1 shows a typical structure for task-specific voice control and dialogue systems. Although the speech recognizer, which converts spoken input into text, and the language analyzer, which extracts meaning from text, are separated into two boxes in the figure, it is desirable that they perform with tight mutual connection, since it is necessary to use semantic information efficiently in the recognizer to obtain correct texts. How to combine these two functions is a most important issue, especially in conversational speech recognition (understanding). Then, the meanings extracted by the language analyzer are used to drive an expert system to select the desired action, to issue commands to various systems, and to receive data from these systems. Replies from the expert system are transferred to a text generator that constructs reply texts. Finally, the text replies are converted into speech by a text-to-speech synthesizer. "Synthesis from concepts" is performed by the combination of the text generator and the text-to-speech synthesizer.

Figure 2 shows hierarchical relationships among the various types

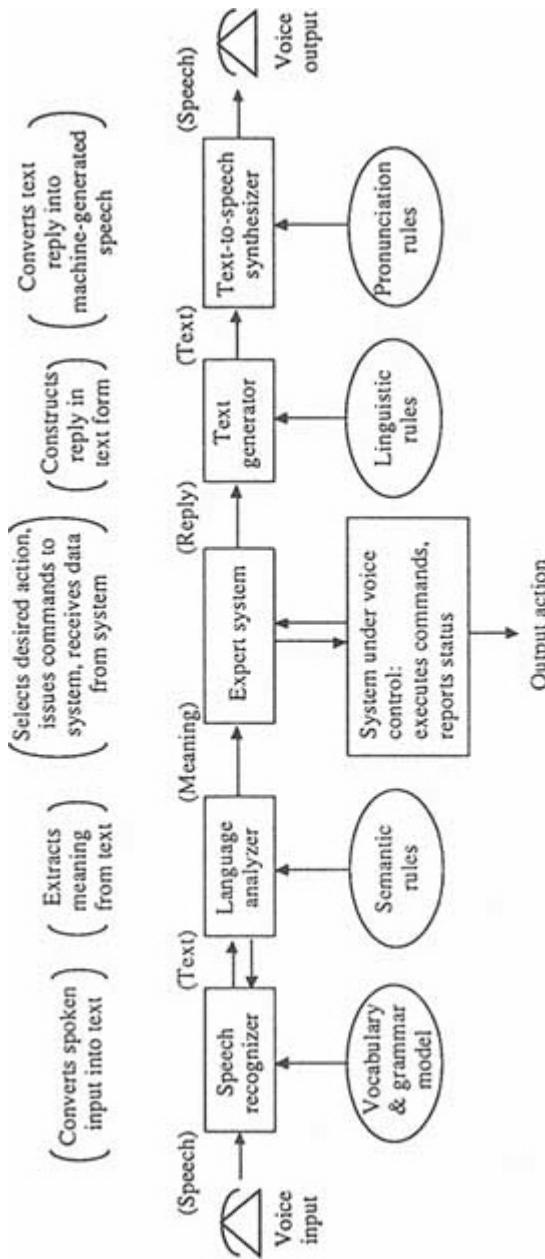


FIGURE 1 Typical structure for task-specific voice control and dialogue systems. (Modified from Flanagan, 1991.)

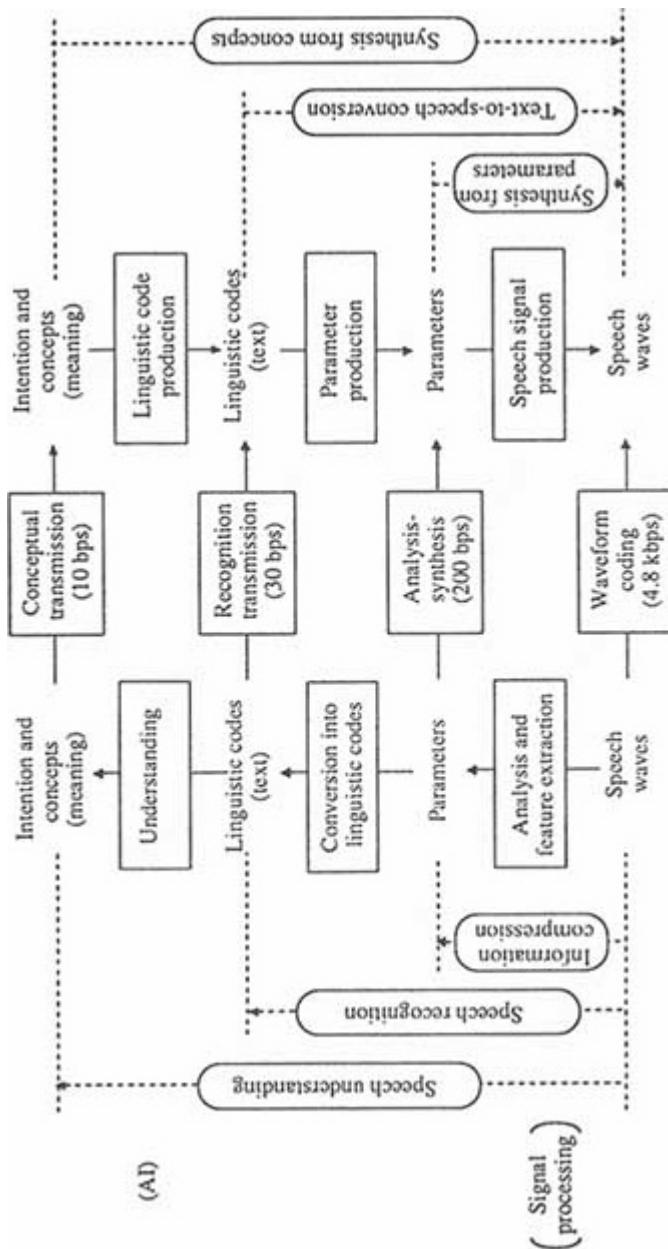


FIGURE 2 Principal speech information-processing technologies and their relationships. (From Furui, 1989).

of speech recognition, understanding, synthesis, and coding technologies. The higher the level, the more abstract the information. This figure is closely related to [Figure 1](#); speech recognition/understanding is the process progressing upward from the bottom to one of the higher levels of [Figure 2](#), and speech synthesis is the process progressing downward from one of the higher levels to the bottom. Historically, speech technology originated from the bottom and has developed toward the extraction and handling of higher-level information. Some of the technology indicated in the figure remains to be investigated. Ultimate speech synthesis/recognition systems that are really useful and comfortable for users should match or exceed human capability. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to use. For this purpose the ultimate systems must be able to handle conceptual information, the highest level of information in [Figure 2](#).

It is, however, neither necessary nor useful to try to use speech for every kind of input and output in computerized systems. Although speech is the fastest and easiest input and output means for simple exchange of information with computers, it is inferior to other means in conveying complex information. It is important to have an optimal division of roles and cooperation in a multimedia environment that includes images, characters, tactile signals, handwriting, etc. (Cohen, in this volume). HuMaNet, built by AT&T Bell Laboratories, is one such advanced experimental multimedia communication system (Flanagan, 1991).

From the human interface point of view, future computerized systems should be able to automatically acquire new knowledge about the thinking process of individual users, automatically correct user errors, and understand the intention of users by accepting rough instructions and inferring details. A hierarchical interface that initially uses figures and images (including icons) to express global information and then uses linguistic expression, such as spoken and written languages, for details would be a good interface that matches the human thinking process.

Ultimate communication systems are expected to use "virtual reality" technology. [Figure 3](#) shows the developing stages of the video and audio interfaces in teleconferencing. A teleconferencing system that uses virtual reality will become possible as these systems evolve from the present concentration type to the projection type and then to the three-dimensional type. In virtual reality systems the participants are not necessarily real human beings. They can be robots or electronic secretaries incorporating speech recognizers, synthesizers,

and expert systems. It is interesting to consider the roles of speech synthesis and recognition technologies in these systems.

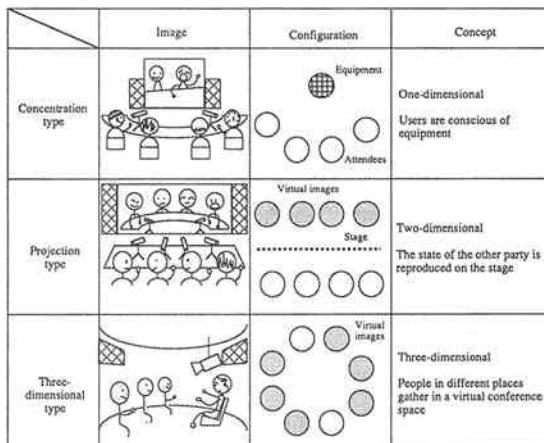


FIGURE 3 Evolution in teleconferencing. (From Koizumi, 1991.)

## FUTURE SPEECH SYNTHESIZERS

Future speech synthesizers should have the following features:

- Highly intelligible (even under noisy and reverberant conditions and when transmitted over telephone networks)
- Natural voice sound
- Prosody control based on meaning
- Capable of controlling synthesized voice quality and choosing individual speaking style (voice conversion from one person's voice to another, etc.)
- Multilingual, multidialectal
- Choice of application-oriented speaking styles, including rhythm and intonation (e.g., announcements, database access, newspaper reading, spoken e-mail, conversation)

- Able to add emotion
- Synthesis of voice from concepts

In present commercial speech synthesizers, voice quality can be selected from male, female, and children's voices. No system has been constructed, however, that can precisely select or control the synthesized voice quality. Research into the mechanism underlying voice quality, inclusive of voice individuality, is thus needed to ensure that synthesized voice is capable of imitating a desired speaker's voice or to select any voice quality such as harshness or softness.

For smooth and successful conversation between computerized systems and people by means of speech recognizers and synthesizers, how to prompt the users by synthesized commands or questions is crucially important. It has been reported that, when users are expected to respond to the machine with isolated words, the percentage of isolated-word responses depended strongly on the prompts made by the machine (Basson, 1992). It was also reported that the intonation of prompted speech as well as the content could strongly influence user responses.

## FUTURE SPEECH RECOGNIZERS

Future speech recognition technology should have the following features:

- Few restrictions on tasks, vocabulary, speakers, speaking styles, environmental noise, microphones, and telephones
- Robustness against speech variations
- Adaptation and normalization to variations due to environmental conditions and speakers
- Automatic knowledge acquisition for phonemes, syllables, words, syntax, semantics, and concepts
- The ability to process discourse in conversational speech (e.g., to analyze context and accept ungrammatical sentences)
- Naturalness and ease of human-machine interaction.
- Recognition of emotion (prosodic information)

Extraction and normalization of (adaption to) voice individuality is one of the most important issues (Furui, 1992a). A small fraction of people occasionally produce exceptionally low recognition rates (the so-called sheep and goats phenomenon). Speaker normalization (adaption) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Experiments have shown that people can adapt to a new speaker's voice

after hearing just a few syllables, irrespective of the phonetic content of the syllables (Kato and Furui, 1985).

TABLE 1 Projections for Speech Recognition

Year	Recognition Capability	Vocabulary Size	Applications
1990	Isolated/connected words Whole-word models, word spotting, finite-state grammars, constrained tasks	10-30	Voice dailing, credit card entry, catalog ordering, inventory inquiry, transaction inquiry
1995	Continuous speech Subword recognition elements, stochastic language models	100-1000	Transaction processing, robot control, resource management
2000	Continuous speech Subword recognition elements, language models representative of natural language, task- specific semantics	5000-20000	Dictation machines, computer-based secretarial assistants, database access
2000+	Continous speech Spontaneous speech grammar, syntax, semantics; adaptation, learning	Unrestricted	Spontaneous speech interaction, translating, telephony

SOURCE: Modified from Rabiner and Juang (1993).

Automatic knowledge acquisition is very important in achieving systems that can automatically follow variations in tasks, including the topics of a conversation. Not only the linguistic structures but also the acoustic characteristics of a speech vary according to the task. Since it is impossible to collect a large database for every kind of task, the recognizers should be able to automatically acquire new knowledge about these features and trace these changes (Atal, in this volume; Bates, in this volume; Marcus, in this volume).

**Table 1** indicates broad projections for speech recognition technology that is/will become available in commercial systems in the next decade. The ultimate systems should be capable of robust speaker-independent or speaker-adaptive, continuous speech recognition. They should have no restrictions on vocabulary, syntax, semantics, or task.

These systems will probably be made possible by implementing automatic learning systems. For the projections in the table to come about, we need continued research in many aspects of speech recognition technology.

The following are also important from the viewpoint of applications:

- Incentive for customers to use the systems
- Low cost
- Creation of new revenues for suppliers
- Cooperation on standards and regulation
- Quick prototyping and development

One of the most useful applications of speech recognition technology in telecommunication networks is the directory assistance service. For this application, systems based on recognizing spoken spelled names are being investigated at many laboratories. However, it is not easy for users to correctly spell the names of persons whose telephone numbers are unknown. In addition, there are several sets of letters having similar pronunciations, such as the English E-rhyme set, and pronunciation of the spelling of other persons' names is often unstable, since this is not a familiar task for us. Therefore, it is not easy for recognizers to correctly recognize alphabetically spelled names. A more successful approach might be to recognize naturally spoken names using the most advanced speech recognition technology, even if the machines have to recognize hundreds of thousands of names (Minami et al., 1992).

The requirements that future speaker recognizers should satisfy are similar to those for future speech recognizers. They include the following:

- Few restrictions on text, speaking style, environmental noise, microphones, and telephones
- Robustness against speech variations
- Adaptation and normalization to variations due to environmental conditions and speakers
- Automatic acquisition of speaker-specific characteristics
- Naturalness and ease of human-machine interaction
- Incentive for customers to use the systems
- Low-cost creation of new revenues for suppliers
- Cooperation on standards and regulation
- Quick prototyping and development

One of the most serious problems arises from variability in a

person's voice. In speaker recognition there are always time intervals between training and recognition, and it is unrealistic to ask every user to utter a large amount of training data. Therefore, the variability problem is more serious for speaker recognition than for speech recognition (Rosenberg and Soong, 1992). Speaker normalization (adaptation) and recognition methods should be investigated using a common approach. This is because they are two sides of the same problem: how best to separate the speaker information and the phoneme information in speech waves or how best to extract and model the speaker-specific phoneme information (Matsui and Furui, 1993).

### TOWARD ROBUST SPEECH/SPEAKER RECOGNITION UNDER ADVERSE CONDITIONS

As described in the previous sections, robustness against speech variations is one of the most important issues in speech/speaker recognition (Furui, 1992b; Juang, 1991; Makhoul and Schwartz, in this volume; Weinstein, in this volume). Methods that are not robust in actual use cannot be considered authentic methods. There are many reasons for speech variations. Even the psychological awareness of communicating with a speech recognizer could induce a noticeable difference in the talker's speech. The main causes of speech variation can be classified according to whether they originate in the speaking and recording environment, the speakers themselves, or the input equipment ([Table 2](#); Furui, 1992b).

Additive noises can be classified according to whether they are correlated or uncorrelated to speech. They can also be classified as stationary or nonstationary. The most typical nonstationary noises are other voices. Although various kinds of signal-processing methods have been proposed to suppress additive noises, we still need to develop more flexible and effective methods, especially for nonstationary noises.

TABLE 2 Main Causes of Speech Variation

Environment	Speaker	Input Equipment
Speech-correlated noise—reverberation, reflection	Attributes of speakers—dialect, gender, age	Microphone (transmitter) Distance to the microphone
Uncorrelated noise—additive noise (stationary, nonstationary)	Manner of speaking—breath and lip noise, stress, Lombard effect, rate, level, pitch, cooperativeness	Filter Transmision system—distortion, noise, echo Recording equipment

Although the physical phenomena of variation can be classified as either noise addition or distortion, the distinction between these categories is not clear. When people speak in a noisy environment, not only does the loudness (energy) of their speech increase, but the pitch and frequency components also change. These speech variations are called the Lombard effect (Junqua, 1992). Several experimental studies have indicated that these indirect influences of noise have a greater effect on speech recognition than does the direct influence of noise entering microphones (Furui, 1992b).

Recognition performance under noisy conditions is often impaired by variations in the amount of speech-quality degradation rather than by the degradation itself. Problems are created, for example, by the variation of noise level associated with variations in the distance between the speaker and the microphone. To cope with these variations, it is essential to develop methods for automatically adapting to and normalizing these effects.

When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts. It is crucially important to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. Instability in the detection of end points is frequently observed. Additionally, the system is required to respond to the utterance as quickly as possible. To solve these problems, it is necessary to establish a method for detecting the time at which sufficient information has been acquired instead of detecting the end of input speech. How to extract contextual information, predict users' responses, and focus on key words are very difficult and important issues.

Style shifting also is an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers read lists of words rather than try to accomplish a real task. Users actually trying to accomplish a task, however, use a different linguistic style.

## SPEECH AND NATURAL LANGUAGE PROCESSING

Speech (acoustic) processing and language processing have usually been investigated in isolation, and the technologies of these two areas have merely been combined to obtain a final decision in speech recognition and understanding. However, the methods produced from the results obtained from natural-language-processing research are not always useful in speech recognition. Therefore, it has recently

become important to investigate new models that tightly integrate speech- and language-processing technology, especially for spontaneous speech recognition (Moore, in this volume; Proceedings of the Speech and Natural Language Workshop, 1992).

These new models should be based on new linguistic knowledge and technology specific to speaking styles, which are very different from read speech. It will be necessary to properly adjust the methods of combining syntactic and semantic knowledge with acoustic knowledge according to the situation. How to extract and represent concepts in speech, that is, how to map speech to concepts, and how to use conceptual associations in recognition processes are important issues in linguistic processing for spontaneous speech (Marcus, in this volume).

Statistical language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. It will also be useful to integrate unification grammars and context-free grammars for efficient word prediction. Adaptation of linguistic models according to tasks and topics is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly (Kuhn and DeMori, 1990).

## USE OF ARTICULATORY AND PERCEPTUAL CONSTRAINTS

Speech research is fundamentally and intrinsically supported by a wide range of sciences. The intensification of speech research continues to underscore an even greater interrelationship between scientific and technological interests (Flanagan, in this volume). Although it is not always necessary or efficient for speech synthesis/recognition systems to directly imitate the human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms to improve performance (Atal, in this volume; Carlson, in this volume; Furui, 1989).

For example, when sequences of phonemes and syllables are produced by human articulatory organs, such as tongue, jaw, and lips, these organs move in parallel, asynchronously, and yet systematically. Present speech analysis methods, however, convert speech signals into a single sequence of instantaneous spectra. It will become important to decompose speech signals into multiple sources based on the concealed production mechanisms (Atal, 1983). This approach seems to be essential for solving the coarticulation problem, one of

the most important problems in both speech synthesis and recognition.

Development of a new sound source model that precisely represents the actual source characteristics, as well as research on the mutual interaction between the sound source and the articulatory filter, would seem to be needed for faster progress in speech synthesis.

Psychological and physiological research into human speech perception mechanisms shows that the human hearing organs are highly sensitive to changes in sounds, that is, to transitional (dynamic) sounds, and that the transitional features of the speech spectrum and the speech wave play crucially important roles in phoneme perception (Furui, 1986). The length of the time windows in which transitions of sounds are perceived has a hierarchical structure and ranges from the order of several milliseconds to several seconds. The hierarchical layers correspond to various speech features, such as phonemes, syllables, and prosodic features. It has also been reported that the human hearing mechanism perceives a target value estimated from the transitional information extracted using dynamic spectral features.

The representation of the dynamic characteristics of speech waves and spectra has been studied, and several useful methods have been proposed. However, the performance of these methods is not yet satisfactory, and most of the successful speech analysis methods developed thus far assume a stationary signal. It is still very difficult to relate time functions of pitch and energy to perceptual prosodic information. Discovery of good methods for representing the dynamics of speech associated with various time lengths is expected to have a substantial impact on the course of speech research. This research is closely related to the analysis method based on the speech production mechanism described above.

The human hearing system is far more robust than machine systems—more robust not only against the direct influence of additive noise but also against speech variations (i.e., the indirect influence of noise), even if the noise is very inconsistent. Speech recognizers are therefore expected to become more robust when the front end uses models of human hearing. This can be done by imitating the physiological organs (Ghitza, 1992) or by reproducing psychoacoustic characteristics (Hermansky, 1990).

Basic speech units for speech synthesis and speech/speaker recognition should be studied from the following perspectives:

- Linguistic units (e.g., phonemes and syllables)
- Articulatory units (e.g., positions and motion targets for the jaw and tongue)

- Perceptual units (e.g., targets and loci of spectral movement and distinctive features)
- Visual units (features used in spectrogram reading)
- Physical units (e.g., centroids in vector/segment quantization)

These units do not necessarily correspond to each other. It will be important to establish new units based on combinations of these viewpoints.

Humans skillfully combine a wide variety of linguistic knowledge concerned with syntax and semantics according to the difficulty and characteristics of given sentences. It is necessary to investigate how to achieve these capabilities in speech recognition. The use of constraints imposed by articulatory and perceptual systems will also be useful for making speech synthesis/recognition systems more natural for the users.

## EVALUATION METHODS

It is important to establish methods for measuring the quality of speech synthesis/recognition systems. Objective evaluation methods that ensure quantitative comparison of a broad range of techniques are essential to technological development in the speech-processing field. Evaluation methods can be classified into the following two categories (Furui, 1991):

- *Task evaluation:* creating a measure capable of evaluating the complexity and difficulty of tasks.
- *Technique evaluation:* formulating both subjective and objective methods for evaluating techniques and algorithms for speech processing.

Task evaluation is very important for speech recognition, since the performances of recognition techniques can be compared only when they are properly compensated for the difficulty of the task. Although several measures for task evaluation have already been proposed, such as word and phoneme perplexity, none of them is good enough at evaluating the difficulty in understanding the meanings of sentences. It may be very difficult to achieve a reliable measure for such purposes, since it involves quantifying all sources of linguistic variability. Nevertheless, we should try to accomplish this target step by step, by creating several measures, such as "meaning perplexity" and "concept perplexity," since these steps are highly related to the basic principles pertaining to modeling the meanings and concepts conveyed by speech.

Technique evaluation must take the viewpoint of improving the human-machine interface (Kamm, in this volume). Ease in human-machine interaction must be properly measured. Recognition systems having minimum recognition errors are not always the best. There are various trade-offs among the categories of errors, such as substitution, insertion, and deletion. Even if the error rate is relatively high, systems are acceptable if the error tendency is natural and matches the principles of human hearing and perception. It is crucially important that recognition errors are easy to correct and the system does not repeat the same errors.

To correctly evaluate technologies and to achieve steady progress, it is important to comprehensively evaluate a technique under actual field conditions instead of under a single controlled laboratory condition. Even if recognition systems perform remarkably well in laboratory evaluations and during demonstrations to prospective clients, they often do not perform nearly as well in the "real-world." This is mainly because the speech that actually has to be recognized varies for many reasons, as mentioned above, and therefore usually differs from training speech. Recognition performance also generally varies with the motive and experience of users.

It was reported that people change their speaking styles when they notice that they are conversing with computers (Wilpon and Roe, 1992). This is another reason why it is important to develop experimental systems and test them under actual field conditions.

## CONCLUSION

This paper predicts speech recognition and synthesis technology for the year 2001 and describes the most important research problems to be solved for accomplishing those ultimate recognition and synthesis systems. The problems include automatic knowledge acquisition, speaking style control in synthesis, synthesis from concepts, robust speech/speaker recognition, adaptation/normalization, language processing, use of articulatory and perceptual constraints, and evaluation methods.

One important issue that is not included in this paper is language identification. It is usually assumed that the language of input speech for recognition, whether English, Japanese, French, etc., is known beforehand. However, in several cases, such as multilanguage interpreting systems, it is necessary to automatically identify the language of input speech. Methods that can be used for this purpose will probably be related to speaker recognition technology.

Although speech synthesis and recognition research have thus

far been done independently for the most part, they will encounter increased interaction until commonly shared problems are investigated and solved simultaneously. Only then can we expect to witness tremendous speech research progress and hence the fruition of widely applicable beneficial techniques.

## REFERENCES

- Atal, B. S., "Efficient coding of LPC parameters by temporal decomposition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston, 2.6 (1983).
- Basson, S., "Prompting the user in ASR applications," Proceedings of the COST 232 Workshop on Speech Recognition over the Telephone Line, Rome, Italy (1992).
- Flanagan, J. L., "Speech technology and computing: A unique partnership," Proceedings of the EUROSPEECH '91, pp. 7-22 (1991).
- Furui, S., "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80, pp. 1016-1025 (1986).
- Furui, S., "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, Inc., New York (1989).
- Furui, S., "Evaluation methods for speech recognition systems and technology," Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods, Chiavari, Italy (1991).
- Furui, S., "Speaker-independent and speaker-adaptive recognition techniques," in *Advances in Speech Signal Processing*, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, Inc., New York, pp. 597-622 (1992a).
- Furui, S., "Toward robust speech recognition under adverse conditions," Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu, pp. 31-42 (1992b).
- Ghitza, O., "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, Inc., New York, pp. 453-485 (1992).
- Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, 87 (4):1738-1752 (1990).
- Juang, B. H., "Speech recognition in adverse environments," *Computer Speech and Language*, 5:275-294 (1991).
- Junqua, J. C., "The variability of speech produced in noise," Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu, pp. 43-52 (1992).
- Kato, K., and S. Furui, "Listener adaptability for individual voice in speech perception," IEICE Technical Report, H85-5 (1985).
- Koizumi, N., "A review of control technology for sound field synthesis," *J. Inst. Television Eng. Jpn.*, 45:474-479 (1991).
- Kuhn, R., and R. DeMori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-12, 6:570-583 (1990).
- Matsui, T., and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, pp. 11-391-394 (1993).
- Minami, Y., K. Shikano, T. Yamada, and T. Matsuoka, "Very large-vocabulary continuous speech recognition for telephone directory assistance," Proceedings of the

- IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, VII.1 (1992).
- Proceedings of the Speech and Natural Language Workshop, Morgan Kaufmann Publishers, San Mateo, Calif. (1992).
- Rabiner, L. R., and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc., New Jersey (1993).
- Rosenberg, A. E., and F. K. Soong, "Recent research in automatic speaker recognition," Advances in Speech Signal Processing, ed. by Furui, S., and M. M. Sondhi, Marcel Dekker, Inc., New York, pp. 701-737 (1992).
- Wilpon, J. G., and D. B. Roe, "AT&T telephone network applications of speech recognition," Proceedings of the COST 232 Workshop on Speech Recognition over the Telephone Line, Rome, Italy (1992).

# Speech Technology in 2001: New Research Directions

*Bishnu S. Atal*

## SUMMARY

Research in speech recognition and synthesis over the past several decades has brought speech technology to a point where it is being used in "real-world" applications. However, despite the progress, the perception remains that the current technology is not flexible enough to allow easy voice communication with machines. The focus of speech research is now on producing systems that are accurate and robust but that do not impose unnecessary constraints on the user. This chapter takes a critical look at the shortcomings of the current speech recognition and synthesis algorithms, discusses the technical challenges facing research, and examines the new directions that research in speech recognition and synthesis must take in order to form the basis of new solutions suitable for supporting a wide range of applications.

## INTRODUCTION

After many years of research, speech recognition and synthesis systems have started moving from the controlled environments of research laboratories to applications in the real-world. Voice-processing technology has matured to such a point that many of us wonder why the performance of automatic systems does not approach the quality of human performance and how soon this goal can be reached.

Rapid advances in very-large-scale integrated (VLSI) circuit capabilities are creating a revolution in the world of computers and communications. These advances are creating an increasing demand for sophisticated products and services that are easy to use. Automatic speech recognition and synthesis are considered to be the key technologies that will provide the easy-to-use interface to machines.

The past two decades of research have produced a stream of increasingly sophisticated solutions in speech recognition and synthesis (Rabiner and Juang, 1993). Despite this progress, the perception remains that the current technology is not flexible enough to allow easy voice communication with machines. This chapter reviews the present status of this important technology, including its limitations, and discusses the range of applications that can be supported by our present knowledge. But as we look into the future and ask which speech recognition and synthesis capabilities will be available about 10 years from now, it is important also to discuss the technical challenges we face in realizing our vision of the future and the directions in which new research should proceed to meet these challenges. We will examine these issues in this paper and take a critical look at the shortcomings of the current speech recognition and synthesis algorithms.

Much of the technical knowledge that supports the current speech-processing technology was created in a period when our ability to implement technical solutions on real-time hardware was limited. These limitations are quickly disappearing, and we look to a future at the end of this decade when a single VLSI chip will have a billion transistors to support much higher processing speeds and more ample storage than is now available.

The speech recognition and synthesis algorithms available at present work in limited scenarios. With the availability of fast processors and a large memory, tremendous opportunity exists to push speech recognition technology to a level where it can support a much wider range of applications. Speech databases with utterances recorded from many speakers in a variety of environments have been important in achieving the progress that has been realized so far. But on the negative side, these databases have encouraged speech researchers to rely on trial-and-error methods, leading to solutions that are narrow and that apply to specific applications but do not generalize to other situations. These methods, although fruitful in the early development of the technology, are now a hindrance as we become much more ambitious in seeking solutions to bigger problems. The time has come to set the next stage for the development of speech technology, and it is important to realize that a solid base of scientific understanding is

absolutely necessary if we want to move significantly beyond where we are today.

The 1990s will be a decade of rising expectations for speech technology, and speech research will expand to cover many areas, from traditional speech recognition and synthesis to speech understanding and language translation. In some areas we will be just scratching the surface and defining the important issues. But in many others the research community will have to come up with solutions to important and difficult problems in a timely fashion. This paper cannot discuss all the possible new research directions but will be limited to examining the most important problems that must be solved during this decade.

## CURRENT CAPABILITIES

Voice communication from one person to another appears to be so easy and simple. Although speech technology has reached a point where it can be useful in certain applications, the prospect of a machine understanding speech with the same flexibility as humans do is still far away. The interest in using speech interface to machines stems from our desire to make machines easy to use. Using human performance as a benchmark for the machine tells us how far we are from that goal. For clean speech, automatic speech recognition algorithms work reasonably well (Makhoul and Schwartz, in this volume; Miller et al., 1961) with isolated words or words spoken in grammatical sentences, and the performance is continuing to improve. [Figure 1](#) shows the word error rate for various test materials and the steady decrease in the error rate achieved from 1980 to 1992. This performance level is not very different from that obtained in intelligibility tests with human listeners. The performance of automatic methods, however, degrades significantly in the presence of noise (or distortion) (Juang, 1991) and for conversational speech.

There are many factors besides noise that influence the performance of speech recognition systems. The most important of these are the size of the vocabulary and the speaking style. [Figure 2](#) shows examples of automatic speech recognition tasks that can be handled by automatic methods for different vocabulary sizes and speaking styles. Generally, the number of confused words increases with the vocabulary size. Current systems can properly recognize a vocabulary of as many as a few thousand words, while the speaking style can vary over a wide range, from isolated words to spontaneous speech. The recognition of continuously spoken (fluent) speech is significantly more difficult than that of isolated words. In isolated words, or speech

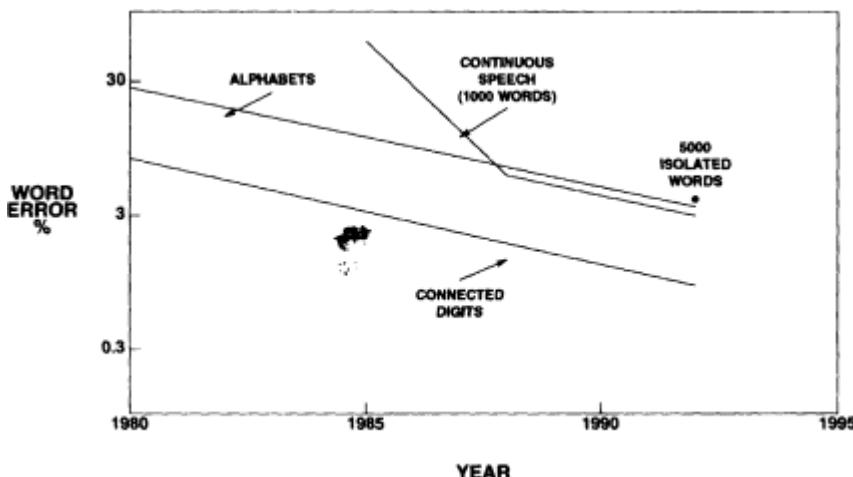


FIGURE 1 Reduction in the word error rate for different automatic speech recognition tasks between 1980 and 1992.

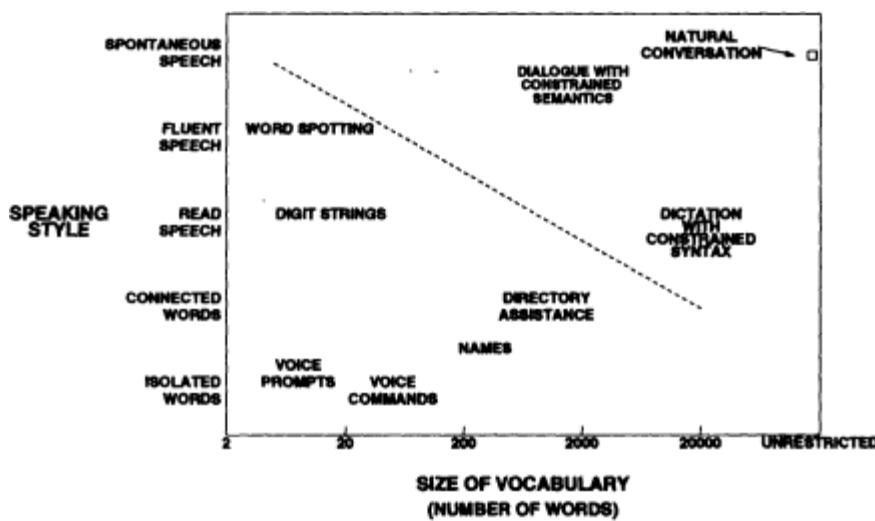


FIGURE 2 Different speech recognition tasks shown in a space of two dimensions: speaking style and size of vocabulary.

where words are separated by distinct pauses, the beginning and the end of each word are clearly marked. But such boundaries are blurred in fluent speech. The recognition of spontaneous speech, such as is produced by a person talking to a friend on a well-known subject, is even harder.

Examples of speech recognition applications that can be handled by the current technology are shown on the left side of the diagonal line in [Figure 2](#). These include recognition of voice commands (prompts), names, digit strings, and key-word spotting. New applications in speech recognition are rapidly emerging (Wilpon, in this volume). Commercial products are available for the recognition of isolated words, connected digit strings, and speech with vocabularies of up to several thousand words spoken with pauses between words.

The items on the right of the diagonal line in [Figure 2](#) are examples of speech recognition tasks that work in laboratory environments but that need more research to become useful for real applications (Roe, in this volume). Automatic recognition of fluent speech with a large-vocabulary is not feasible unless constraints on the syntax or semantics are introduced. The present knowledge in handling natural languages and in following a dialogue is very much limited because we do not understand how to model the variety of expressions that natural languages use to convey concepts and meanings.

Text-to-speech synthesis systems suffer from much of the same kinds of problems as speech recognition. Present text-to-speech systems can produce speech that is intelligible (although significantly lower intelligibility than natural speech) but not natural sounding. These systems can synthesize only a few voices reading grammatical sentences but cannot capture the nuances of natural speech.

## CHALLENGING ISSUES IN SPEECH RESEARCH

For speech technology to be used widely, it is necessary that the major roadblocks faced by the current technology be removed. Some of the key issues that pose major challenges in speech research are listed below:

- *Ease of use.* Unless it is easy to use, speech technology will have limited applications. What restrictions are there on the vocabulary? Can it handle spontaneous speech and natural spoken language?
- *Robust performance.* Can the recognizer work well for different speakers and in the presence of the noise, reverberation, and spectral distortion that are often present in real communication channels?
- *Automatic learning of new words and sounds.* In real applica

tions the users will often speak words or sounds that are not in the vocabulary of the recognizer. Can it learn to recognize such new words or sounds automatically?

- *Grammar for spoken language.* The grammar for spoken language is quite different from that used in carefully constructed written text. How does the system learn this grammar?
- *Control of synthesized voice quality.* Can text-to-speech synthesis systems use more flexible intonation rules? Can prosody be made dependent on the semantics?
- *Integrated learning for speech recognition and synthesis.* Current speech synthesis systems are based on rules created manually by an experienced linguist. Such systems are constrained in what they can do. Can new automatic methods be developed for the training of the recognizer and synthesizer in an integrated manner?

Some of the issues mentioned above, such as ease of use and robustness, need to be addressed in the near future and resolved. Others, such as automatic learning of new words and sounds or grammar for spoken language, will need major advances in our knowledge. Understanding of spontaneous speech will require tight integration of language and speech processing.

A number of methods have been proposed to deal with the problem of robustness. The proposed methods include signal enhancement, noise compensation, spectral equalization, robust distortion measures, and novel speech representations. These methods provide partial answers valid for specific situations but do not provide a satisfactory answer to the problem. Clean, carefully articulated, fluent speech is highly redundant, with the signal carrying significantly more information than is necessary to recognize words with high-accuracy. However, the challenge is to realize the highest possible accuracy when the signal is corrupted with noise or other distortions and part of the information is lost. The performance of human listeners is considered to be very good, but even they do not approach high intelligibility for words in sentences unless the signal-to-noise (S/N) ratio exceeds 18 dB (Miller et al., 1961).

## THE ROBUSTNESS ISSUE

Let us consider the robustness issue in more detail. Current speech recognition algorithms use statistical models of speech that are trained from a prerecorded speech database. In real applications the acoustic characteristics of speech often differ significantly from that of speech in the training database, and this mismatch causes a drop in the

recognition accuracy. This is illustrated for noise-contaminated speech in [Figure 3](#), which shows the recognition accuracy as a function of the S/N ratio for both matched and mismatched training and test conditions (Dautrich et al., 1983; Juang, 1991). These results point to a serious problem in current speech recognition systems: the performance degrades whenever there is a mismatch between levels of noise present in training and test conditions. Similar problems arise with spectral distortion, room reverberation, and telephone transmission channels (Acero and Stern, 1990). Achieving robust performance in the presence of noise and spectral distortion has become a major issue for the current speech recognition systems.

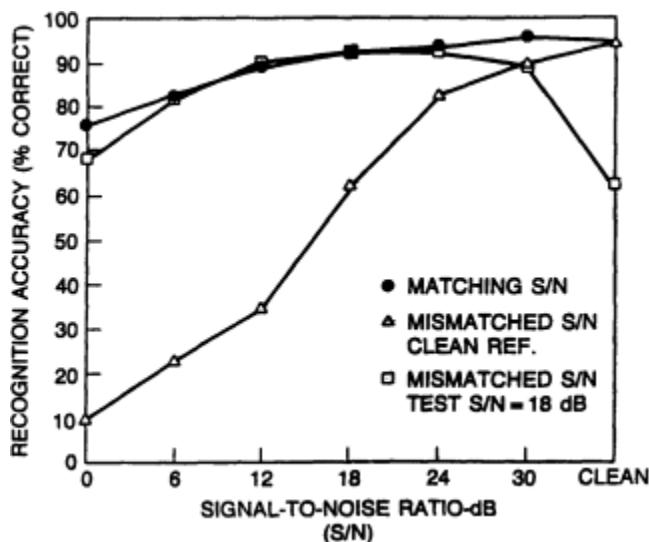


FIGURE 3 Speech recognition performances in noisy conditions: ·, training and testing have matched S/N ratios; A, only clean training data are used; □, training and testing S/N ratios are mismatched with test S/N ratio fixed at 18 dB (after Juang, 1991).

Robust performance does not come by chance but has to be designed into the system. Current speech recognition algorithms are designed to maximize performance for the speech data in the training set, and this does not automatically translate to robust performance on speech coming from different user environments. [Figure 4](#) shows the principal functions of an automatic speech recognition system. The input speech utterance is analyzed in short quasi-stationary segments, typically 10 to 30 ms in duration, to provide a parametric representation at the acoustic level. The parameters from the unknown

input utterance are then compared to patterns derived from a large training set of speech utterances collected from many speakers in many different speaking environments. This comparison provides a set of scores representing the similarity between the unknown pattern and each of the prestored patterns. The last step combines these scores together with other knowledge about the speech utterance, such as the grammar and semantics, to provide the best transcription of the speech signal. To achieve robustness, each function shown in the block diagram must be designed to minimize the loss in performance in situations when there is a mismatch between the training and test conditions.

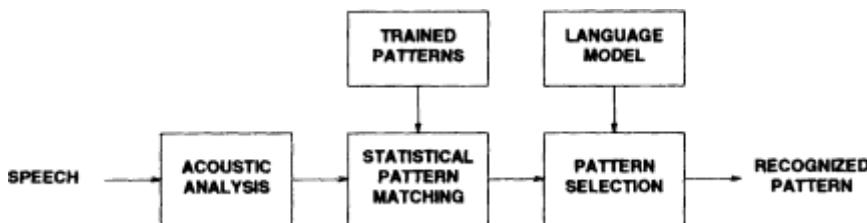


FIGURE 4 Principal functions of an automatic speech recognition system.

A speech recognizer can be regarded as a method for compressing speech from a high rate needed to represent individual samples of the waveform to a low phonemic rate to represent speech sounds. Let us look at the information rate (bit rate) at different steps in the block diagram of Figure 4. The bit rate of the speech signal represented by its waveform at the input of the recognizer is in the range of 16 to 64 kb/s. The rate is reduced to approximately 2 kb/s after acoustic analysis and to a phonemic rate in the range 30 to 50 b/s after pattern matching and selection.

The bit rate at the acoustic parameter level is large, and therefore the pattern-matching procedure must process speech in "frames" whose duration is only a small fraction of the duration of a sound. The scores resulting from such a pattern-matching procedure are unreliable indicators of how close an unknown pattern from the speech signal is to a particular sound. The reliability can be improved by reducing the maximum number of acoustic patterns in the signal (or its bit rate) that are evaluated for pattern matching. The bit rate for representing the speech signal depends on the duration of the time window that is used in the analysis shown in Figure 5 and is about 200 b/s for a window of 200 ms. Suppose we wish to compute the score for a speech segment 100 ms in duration, which is roughly the average length of a speech sound. The number of acoustic patterns

that the pattern-matching step has to sort out is 2200 at 2000 b/s, but that number is reduced to only 220 at 200 b/s. This is a reduction of 218° in the number of patterns that the pattern-matching procedure has to handle. The present speech analysis methods generate a static (quasi-stationary) representation of the speech signal. To achieve robust performance, it is important to develop methods that can efficiently represent speech segments extending over a time interval of several hundred milliseconds. An example of a method for representing large speech segments is described in the next section.

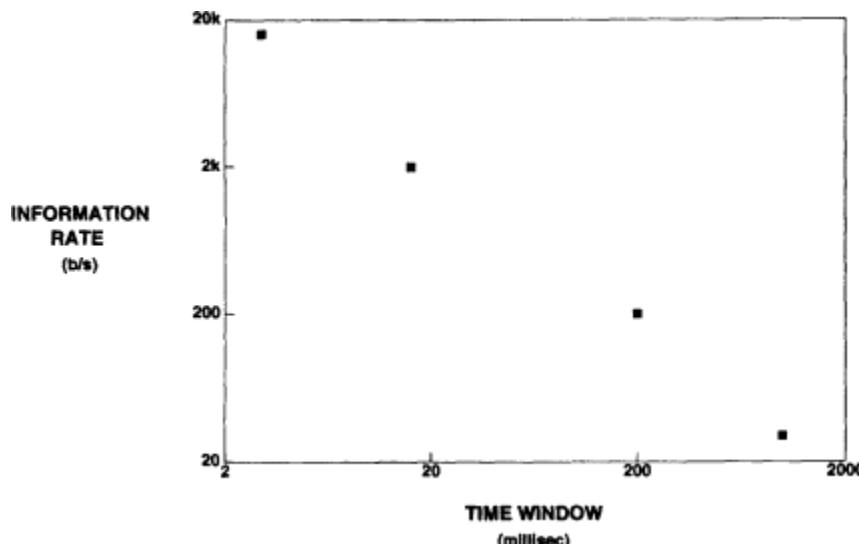


FIGURE 5 Information rate (b/s) of speech signal as a function of the length of the time window used in the analysis.

### SPEECH ANALYSIS

The goal of speech analysis is to provide a compact representation of the information content in the speech signal. In general, those representations that eliminate information not pertinent to phonetic differences are effective. The short-time power spectrum of speech, obtained either from a filter bank, Fourier transform, or linear prediction analysis, is still considered the most effective representation for speech recognition (the power spectrum is often converted into the cepstrum to provide a set of 10 to 15 coefficients). However, the power spectrum is affected by additive noise and linear-filtering distortions. We need new representations that go beyond the power spectrum and represent the frequency content of the signal.

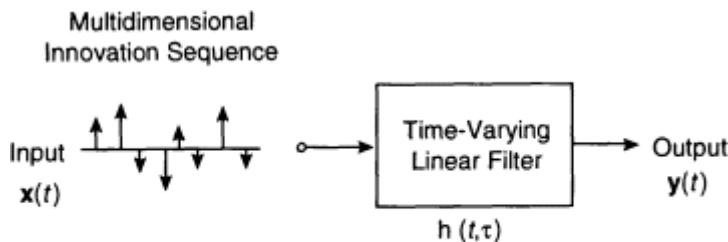
The cepstral coefficients are instantaneous (static) features. One of the important advances in the acoustic representation of speech has been the introduction of dynamic features (Furui, 1986), such as first- and second-order derivatives of the cepstrum. Recently, new representations based on human hearing have been proposed (Ghitza, 1992), but these representations have not yet been found to have significant advantage over the spectral representation. The following is a list of interesting new research directions in speech analysis:

- *Time-frequency and wavelet representations.* Time-frequency representations map a one-dimensional signal into a two-dimensional function of time and frequency (Daubechies, 1990; Hlawatsch and Boudreaux-Bartels, 1992; Rioul and Vetterli, 1991). The traditional Fourier analysis methods divide the time-frequency plane in an inflexible manner not adapted to the needs of the signal. New methods of time-frequency analysis are emerging that allow more general partitioning of the time-frequency plane or tiling that adapts to time as well as frequency as needed (Daubechies, 1990; Herley et al., 1993).
- *Better understanding of auditory processing of signals.* Although auditory models have not yet made a significant impact on automatic speech recognition technology, they exhibit considerable promise. What we need is a better understanding of the principles of signal processing in the auditory periphery that could lead to more robust performance in automatic systems.
- *Articulatory representation.* Models that take advantage of the physiological and physical constraints inherent in the vocal tract shapes used during speech production can be useful for speech analysis. Significant progress (Schroeter and Sondhi, 1992) has been made during the past decade in developing articulatory models whose parameters can be estimated from the speech signal.
- *Coarticulation models at the acoustic level.* During speech production, the articulators move continuously in time, thereby creating a considerable overlap in the acoustic realizations of phonemes. Proper modeling of coarticulation effects at the acoustic level can provide better accuracy and higher robustness in speech recognition.

### Temporal Decomposition

We discussed earlier the importance of extending the quasi-stationary static model of speech to a dynamic model that is valid over much longer nonstationary segments. We describe here one such model, known as temporal decomposition (Atal, 1983). The acoustics of the speech signal at any time are influenced not only by the sound

being produced at that time but also by neighboring sounds. Temporal decomposition seeks to separate the contributions of the neighboring sounds on the acoustic parameters by using a coarticulation model in which the contributions of sounds are added together with proper weights (Atal, 1983, 1989; Cheng and O'Shaughnessy, 1991, 1993).



$$y(t) = \int_{t_{\infty}}^t x(\tau) h(t, \tau) d\tau$$

FIGURE 6 Temporal decomposition model to represent coarticulation at the acoustic level.

In the temporal decomposition model the continuous variations of acoustic parameters are represented as the output of a linear time-varying filter excited by a sequence of vector-valued delta functions located at nonuniformly spaced time intervals (Atal, 1989). This is illustrated in Figure 6, where the linear filter with its impulse response specified by  $h(t, T)$  (response at time  $t$  due to a delta function input at time  $T$ ) has the role of smoothing the innovation  $x(t)$  that is assumed to be nonzero only at discrete times corresponding to the discrete nature of speech events. The number of nonzero components in the innovation in any given time interval is roughly equal to the number of speech events (and silence) contained in that interval of the spoken utterance. Speech analysis techniques have been developed to determine both the innovation and the time-varying impulse response of the filter for any utterance (Atal, 1989; Cheng and O'Shaughnessy, 1991, 1993). Figure 7 shows an example of this decomposition for the word "four." The three parts of the figure show: (a) even components of the linear predictive coding (LPC) line spectral frequencies as a function of time, (b) the filter impulse responses for each speech event, and (c) the waveform of the word "four." In this example the entire variations in the acoustic parameters over 0.5 s of the utterance for the word "four" can be represented as the sum of five overlapping speech events. We find that the information rate

of the innovation signal  $x(t)$  is about 100 b/s, which is much lower than the corresponding rate for the acoustic parameters  $y(t)$ .

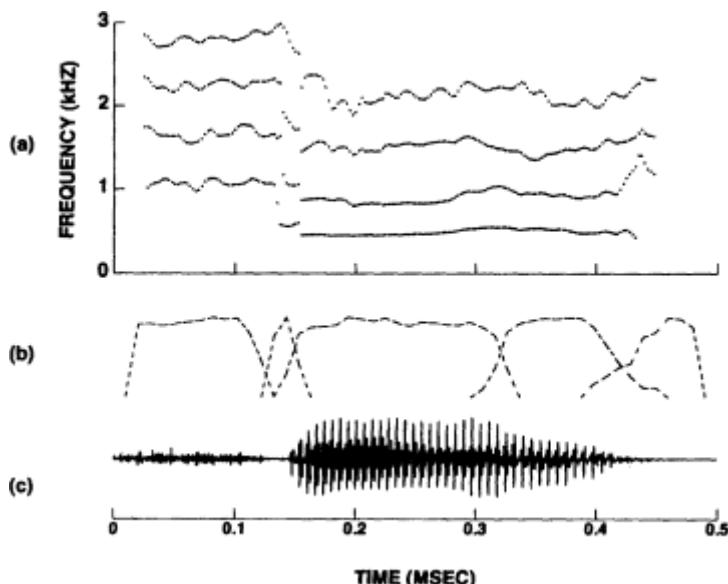


FIGURE 7 Temporal decomposition of the spoken word "four": (a) LPC line spectral parameters, (b) filter impulse responses for the different speech events, and (c) speech waveform for the speech utterance.

### TRAINING AND PATTERN-MATCHING ISSUES

The application of hidden Markov models (HMMs) has been a major factor behind the progress that has been achieved in automatic speech recognition (Rabiner and Juang, 1993). The HMM framework provides a mathematically tractable approach to the training and classification problems in speech recognition. While the speech recognition algorithms based on the HMM are important at the current state of the technology, these algorithms suffer from fundamental shortcomings (Juang and Rabiner, 1971) that must be overcome.

The HMM method is based on the Bayesian approach to pattern classification, which assumes that the statistical distributions of the HMM states are known or can be estimated. In the HMM, therefore, the problems of training and recognition are transformed to the problem of estimating distributions from the training data. In reality this is a difficult task requiring untested assumptions about the form and the

underlying parameters of the distributions. Moreover, the misclassification errors depend on the amount of overlap between the tails of the competing distributions and not on the exact shape of the distributions for the classes. Thus, the emphasis in the HMM approach on distribution estimation is unnecessary; a cost function defined in a suitable fashion is all that is required.

Other approaches to speech recognition based on discriminant functions are being investigated and appear to be promising. Significant progress has been made in formulating the discriminant approach for speech recognition and in developing methods that seek to minimize the misclassification errors (Juang and Katagiri, 1992). The major issues in training and recognition are listed below:

- *Training and generalization.* An important question is whether the trained patterns characterize the speech of only the training set or whether they also generalize to speech that will be present in actual use.
- *Discriminative training.* Although the discriminative training does not require estimation of distributions, they still need knowledge of the discriminant functions. What are the most appropriate discriminant functions for speech patterns?
- *Adaptive learning.* Can the learning of discriminant functions be adaptive?
- *Artificial neural networks.* Despite considerable research, neural networks have not yet shown significantly better performance than HMM algorithms. New research must address the important issue—what is the potential of neural networks in providing improved training and recognition for speech patterns?

### ADDITIONAL ISSUES IN SPEECH SYNTHESIS

Much of what has been discussed so far applies to speech synthesis as well. However, there are additional research issues that must be considered. We will discuss some of these issues in this section.

The core knowledge that forms the basis of current speech recognition and synthesis algorithms is essentially the same. However, there are important differences in the way the two technologies have evolved. Speech synthesis algorithms generate continuous speech by concatenating segments of stored speech patterns, which are selected to minimize discontinuities in the synthesized speech. Segmentation of speech into appropriate units, such as diphones or syllables, was therefore incorporated into the speech synthesis technology at an early stage and required the assistance of trained people (or phoneticians)

to learn the segmentation task. Lack of accurate models for representing the coarticulation in speech and the dynamics of parameters at the acoustic or the articulatory level has been the major obstacle in developing automatic methods to carry out the segmentation task. Without automatic methods, it is difficult to process large speech databases and to develop models that represent the enormous variability present in speech due to differences in dialects, prosody, pronunciation, and speaking style. Future progress in synthesizing speech that offers more than minimal intelligibility depends on the development of automatic methods for extracting parameters from speech to represent the important sources of variability in speech in a consistent fashion. Automatic methods for segmentation are also needed in order to develop multilingual capability in speech synthesis.

The primary goal of speech synthesis systems so far has been to synthesize speech from text—a scenario coming out of an earlier interest in "reading machines for the blind." New applications of speech synthesis that do not depend on synthesizing speech from text are rapidly emerging. As we proceed to develop new applications that involve some kind of dialogue between humans and machines, it is essential that the issue of synthesizing speech from concepts be addressed.

## CONCLUSIONS

Voice communication holds the promise of making machines easy to use, even as they become more complex and powerful. Speech technology is reaching an important phase in its evolution and is getting ready to support a wide range of applications. This paper discussed some of the important technical challenges in developing speech recognition and synthesis technology for the year 2001 and the new research directions needed to meet these challenges.

Robust performance in speech recognition and more flexibility in synthesizing speech will continue to be major problems that must be solved expeditiously. The solutions will not come by making incremental changes in the current algorithms but rather by seeking new solutions that are radically different from the present.

New speech analysis methods must move beyond quasi-stationary representations of the power spectrum to dynamic representations of speech segments. Solution of the coarticulation problem at the acoustic level remains one of the most important problems in speech recognition and synthesis. Temporal decomposition is a promising method along this direction.

In speech recognition, new training procedures based on discriminant functions show considerable promise and could avoid the limitations

of the HMM approach. The discriminant function approach achieves higher performance by using a criterion that minimizes directly the errors due to misclassification. In speech synthesis, articulatory models and automatic methods for determining their parameters offer the best hope of providing the needed flexibility and naturalness in synthesizing a wide range of speech materials.

## REFERENCES

- Acero, A., and R. M. Stern, "Environmental robustness in automatic speech recognition," Proc. ICASSP-90, pp. 849-852, Albuquerque, NM, 1990.
- Atal, B. S., "Efficient coding of LPC parameters by temporal decomposition," Proceedings of the International Conference IEEE ASSP, Boston, pp. 81-84, 1983.
- Atal, B. S., "From speech to sounds: Coping with acoustic variabilities," Towards Robustness in Speech Recognition, Wayne A. Lea (ed.), pp. 209-220, Speech Science Publications, Apple Valley, Minn., 1989.
- Cheng, Y. M., and D. O'Shaughnessy, "Short-term temporal decomposition and its properties for speech compression," IEEE Trans. Signal Process., vol. 39, pp. 1282-1290, 1991.
- Cheng, Y. M., and D. O'Shaughnessy, "On 450-600 b/s natural sounding speech coding," IEEE Trans. Speech Audio Process., vol. 1, pp. 207-220, 1993.
- Daubechies, I., "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Inf. Theory, vol. 36, pp. 961-1005, Sept. 1990.
- Dautrich, B. A., L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-31, pp. 793-806, Aug. 1983.
- Furui, S., "On the role of spectral transitions for speech perception," J. Acoust. Soc. Am., vol. 80, pp. 1016-1025, Oct. 1986.
- Ghitza, O., "Auditory nerve representation as a basis for speech processing," Advances in Speech Signal Processing, S. Furui and M. M. Sondhi (eds.), pp. 453-485, Marcel Dekker, New York, 1992.
- Herley, C., et al., "Time-varying orthonormal tilings of the time-frequency plane," IEEE Trans. Signal Process., Dec. 1993.
- Hlawatsch, F., and G. F. Boudreault-Bartels, "Linear and quadratic time-frequency signal representations," IEEE Signal Process. Mag., pp. 21-67, Apr. 1992.
- Juang, B. H., "Speech recognition in adverse environments," Comput. Speech Lang., vol. 5, pp. 275-294, 1991.
- Juang, B. H., and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Trans. Signal Process., vol. 40, pp. 3043-3054, Dec. 1992.
- Juang, B. H., and L. R. Rabiner, "Hidden Markov models for speech recognition," Technometrics, vol. 33, pp. 251-272, Aug. 1991.
- Miller, G. A., G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol., vol. 41, pp. 329-335, 1961.
- Rabiner, L. R., and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.
- Rioul, O., and M. Vetterli, "Wavelets and signal processing," IEEE Signal Process. Mag., pp. 14-38, Oct. 1991.
- Schroeter, J., and M. M. Sondhi, "Speech coding based on physiological models of speech production," Advances in Speech Signal Processing, S. Furui and M. M. Sondhi (eds.), pp. 231-267, Marcel Dekker, New York, 1992.

# New Trends in Natural Language Processing: Statistical Natural Language Processing

*Mitchell Marcus*

## SUMMARY

The field of natural language processing (NLP) has seen a dramatic shift in both research direction and methodology in the past several years. In the past, most work in computational linguistics tended to focus on purely symbolic methods. Recently, more and more work is shifting toward hybrid methods that combine new empirical corpus-based methods, including the use of probabilistic and information-theoretic techniques, with traditional symbolic methods. This work is made possible by the recent availability of linguistic databases that add rich linguistic annotation to corpora of natural language text. Already, these methods have led to a dramatic improvement in the performance of a variety of NLP systems with similar improvement likely in the coming years. This paper focuses on these trends, surveying in particular three areas of recent progress: part-of-speech tagging, stochastic parsing, and lexical semantics.

## SOME LIMITATIONS OF RULE-BASED NLP

Until about 3 or 4 years ago, all natural language processing (NLP) systems were entirely hand constructed, with grammars and semantic components made up of many carefully handcrafted rules. Often the target coverage of such systems was based on a small set of ex

emplar sentences; many such systems were originally developed on fewer than several hundred examples. While these systems were able to provide adequate performance in interactive tasks with typed input, their success was heavily dependent on the almost magical ability of users to quickly adapt to the limitations of the interface.

The situation is quite different, however, when these rule sets are applied open loop to naturally occurring language sources such as newspaper texts, maintenance manuals, or even transcribed naturally occurring speech. It now appears unlikely that hand-coded linguistic grammars capable of accurately parsing unconstrained texts can be developed in the near term. In an informal study conducted during 1990 (Black et al. 1992b), short sentences of 13 words or less taken from the Associated Press (AP) newswire were submitted to a range of the very best parsers in the United States, parsers expressly developed to handle text from natural sources. None of these parsers did very well; the majority failed on more than 60 percent of the test sentences, where the task was to find the one correct parse for each sentence in the test set. Another well-known system was tested by its developer using the same materials in 1992, with a failure rate of 70 percent.

This failure rate actually conflates two different, and almost contradictory, problems of this generation of parsers. The first is that the very large handcrafted grammars used by parsers that aim at broad coverage often generate very, very large numbers of possible parses for a given input sentence. These parsers usually fail to incorporate some source of knowledge that will accurately rank the syntactic and semantic plausibility of parses that are syntactically possible, particularly if the parser is intended to be domain independent. The second problem, somewhat paradoxically, is that these parsers often fail to actually provide the correct analysis of a given sentence; the grammar of a natural language like English appears to be quite vast and quite complex.

Why can't traditional approaches to building large software systems, using techniques like divide and conquer, solve this last problem? The problem is not that the grammar developers are not competent or that there is a lack of effort; a number of superb computational linguists have spent years trying to write grammars with broad enough coverage to parse unconstrained text. One hypothesis is that the development of a large grammar for a natural language leads into a complexity barrier similar to that faced in the development of very large software systems. While the human grammatical system appears to be largely modular, the interaction of the subcomponents is still sufficient to cause the entire system to be unmanageably com

plex. The net result is that the grammatical system does not appear to decompose easily into units that a team can develop and then join together. In support of this view is the fact that almost all of the large grammars extant are the result of a single grammar developer working over a long period of time. If this conclusion is correct, an approach to developing NLP systems must be found other than careful handcrafting.

### STATISTICAL TECHNIQUES: FIRST APPEARANCE

One of the first demonstrations that stochastic modeling techniques, well known in the speech-processing community, might provide a way to cut through this impasse in NLP was the effective application of a simple letter trigram model to the problem of determining the national origin of proper names for use in text-to-speech systems (Church, 1985). Determining the etymology of names is crucial in this application because the pronunciation of identical letter strings differs greatly from language to language; the string *GH*, for example, is pronounced as a hard *G* in Italian, as in *Aldrichetti*, while most often pronounced as *F* or simply silent in English, as in *laugh* or *sigh*. This system estimates the probability that a name *W* comes from language *L* as the product of the probabilities, estimated across a set of known names from *L*, of all the contiguous three-letter sequences in *W*. It then assigns *W* to the language *L*, which maximizes this probability. The success of this program came as a surprise to most of the NLP community, at the time completely wedded to the symbolic techniques of traditional artificial intelligence (AI). Many people in NLP thought this application was a fluke, that the task solved by the program was somehow special. In fact, this technique led the way toward application of statistical techniques to problems that one would have thought required an "AI-complete" solution, a full solution to the problem of modeling human understanding.

In another sense this work was an application of an approach to linguistic analysis called distributional analysis (Harris, 1951), which reached its zenith in the 1950s. This work suggested that the structure of language could be discovered by looking at distributional patterns of linguistic entities. While the work of Chomsky in the late 1950s showed that distributional analysis could not be the whole story, most linguists assumed that Chomsky's work implied that distributional techniques should be abandoned entirely. This application showed that simple distributional techniques were useful for solving hard engineering problems that looked resistant to the application of a priori knowledge.

## THE ARCHITECTURE OF AN NLU SYSTEM

Figure 1a gives an overview of a few of the crucial steps in the process of decoding a sentence in a conventional NLU system, given that the words that make up the sentence have been determined either by a speech recognition system or by tokenization of an ASCII source. When a new sentence comes in, it is analyzed by a parser that both determines what part of speech to assign to each of the words

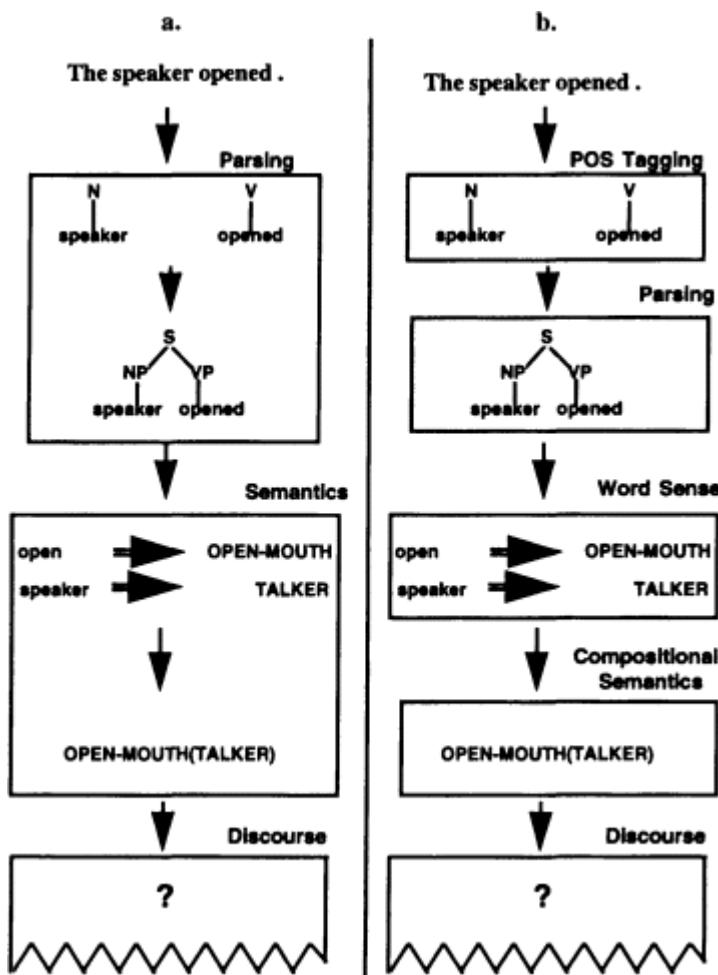


FIGURE 1 Two decompositions of the architecture of an NLU system. (a)—Standard decomposition. (b)—An alternate decomposition.

and combines these part-of-speech tagged words into larger and larger grammatical fragments, using some kind of grammar that tells what combinations are possible and/or likely. The output of this grammatical analysis, either a single-rooted tree or a string of tree fragments, then goes through semantic analysis, which determines the literal meaning of a sentence in isolation. This phase of analysis decides both what the individual words mean and how to combine the individual word meanings into larger semantical structures. Often, this last step is done using some form of *compositional semantics*, where the meaning of each larger unit is constrained to be a relatively simple function of the meaning of its parts. This meaning representation is then further analyzed by pragmatic and discourse components to determine what the sentence means given its particular context of use and to place this representation into a multisentence representation useful for such tasks as determining the referent of pronouns and other noun phrases.

For the purposes of the rest of this chapter, the problem will be subdivided into somewhat smaller functional units than given by the conventional model. This subdivision, given in [Figure 1b](#), reflects the development of statistical NLP techniques over the past several years, with rate of progress roughly proportional to height in the figure. The first success of statistical modeling techniques for NLU was in the area of part-of-speech determination—deciding, for example, whether the word *saw* functioned in a given linguistic context as a singular noun or a past tense verb. A variety of techniques now tag previously unseen material with 95 to 97 percent accuracy. Recently, purely context-free probabilistic parsing methods have been supplanted by parsing algorithms that utilize probabilities of context-free rules conditionalized on aspects of surrounding linguistic structure. Such parsers provide the correct parse as the first parse output between 60 to 80 percent of the time, by sentence, on naturally occurring texts from rich, but not entirely unconstrained, domains such as the *Wall Street Journal*. They have performed with up to 91 percent accuracy on spoken language tasks from limited domains like the Advanced Research Projects Agency's (ARPA) Air Travel Information Service (ATIS) domain. In the area of lexical semantics, a range of promising techniques for performing word-sense disambiguation have emerged in just the last year, as well as some preliminary work in automatically determining the selectional restrictions of verbs, that is, what kind of objects can serve as the subject or object of a given verb.

Finally, all of these methods depend crucially on the availability of training materials annotated with the appropriate linguistic structure. These advances were made possible by the development of cor

pura appropriately annotated with part-of-speech and syntactic structure. This paper will also touch on the development of such corpora.

## PART-OF-SPEECH TAGGING

The task of a part-of-speech tagger is to map an input stream of word tokens into the correct part of speech for each word token in context. To do this, it must disambiguate words that have the potential to be many different parts of speech. A part-of-speech tagger, for example, should map the sentence *Can we can the can?* into the string of parts of speech shown in [Figure 2](#). This problem of lexical disambiguation is a central problem in building any NLP system; given a realistically large lexicon of English, many common words are used in multiple parts of speech. Determining what function each word plays in context is a crucial part of either assigning correct grammatical structure for purposes of later semantic analysis or of providing a partial heuristic chunking of the input into phrases for purposes of assigning intonation in a text-to-speech synthesizer.

The problem of lexical disambiguation was thought to be completely intractable 10 years ago by most of the NLP community, and yet now a wide range of very different techniques can solve this problem with 95 to 97.5 percent word accuracy, for part-of-speech tag sets of between 40 and 80 tags, depending on the task and how accuracy is measured (see, e.g., Black et al., 1992a; Brill, 1992; Church, 1988; Cutting et al., 1992; Hindle, 1989; Merialdo, 1991; Weischedel et al., 1993; and others). It is worth noting that many of these errors are not very harmful; a significant fraction of the errors consist of cases where one kind of verb is confused with another kind of verb or one kind of noun with another. Many of the parsers for which these taggers serve as preprocessors are forgiving enough that the errors do not actually throw the parser off track.

Most part-of-speech taggers are implemented as hidden Markov models (HMMs). For an input sentence  $S = w_1, w_2, \dots, w_n$ , these taggers predict a tag  $t_i$  for each word  $w_i$  given two sets of probabilities: First,  $P(w | t)$  (the probability of  $w$  given  $t$ ), the probability for each possible word  $w$  and each part-of-speech tag  $t$  that if a given word is tagged with  $t$ , it is in fact the word  $w$ . Second,  $P(t_{i+1} | t_i)$ , the transition probability that the next tag is  $t_{i+1}$ , given that the current

Words in:	Can	we	can	the	can?
Part-of-speech stream out:	modal	pronoun	verb	det	noun

FIGURE 2 Part-of-speech taggers assign tags in context.

tag is  $t_i$ . These taggers use a linear time search utilizing the dynamic programming algorithm, often called the Viterbi algorithm, to find the string of tags  $T = t_1, t_2 \dots t_n$ , that maximize  $\pi P(w_i \mid t_i) P(t_i \mid t_{i+1})$ .

The question here is how to estimate the value of the parameters of the HMM. The standard approach for HMMs is to use the forward/backward algorithm to automatically estimate the parameters, as described by Jelinek elsewhere in this volume. However, systems that use the forward/backward algorithm do not perform quite as well as those that estimate parameters, at least initially, by simple counting, using a corpus of text that has been pretagged with part-of-speech information (Merialdo, 1991). In practice, such systems must use some technique to *smooth*<sup>1</sup> very small counts. One could then use the forward/backward algorithm to smooth these direct estimates, but there is little evidence that this helps. Currently, then, the best way to estimate the parameters of an HMM for part-of-speech tagging is to hand tag a corpus and simply count.

The theme that emerges here is true of most statistical NLP applications and will be a leitmotif in what follows below. What works best both for part-of-speech tagging using HMMs and for the entire range of statistical NLP applications considered in this paper, is some appropriate combination of stochastic techniques and linguistic knowledge. While earlier work provides evidence that handcrafted symbolic representations of linguistic knowledge are insufficient to provide industrial-strength NLP, it also appears that the use of statistical methods without some incorporation of linguistic knowledge is insufficient as well. This linguistic knowledge may either be represented in implicit form, as in the use of a pretagged corpus here, or encoded *explicitly* in the form of a grammar.<sup>2</sup> In the next few years, I believe we are going to see stochastic techniques and linguistic knowledge more and more deeply interleaved.

### The Problem of Unknown Words

In conjunction with this observation it is important to realize that if one simply implemented an HMM for part-of-speech tagging as

---

<sup>1</sup> Since sentence probabilities are estimated by multiplying together many estimates of local probabilities, any probability estimate of zero leads to a zero probability for the entire string. Since any direct estimate is based on only finite data, it is important to assume that any event not observed at all has some very small, but nonzero probability. How to best perform this smoothing of probability estimates is a central technical issue in applying any of the methods discussed in this chapter.

<sup>2</sup> For readers familiar with logic, this is the distinction between knowledge represented *extensionally* and knowledge represented *intensionally*.

discussed above, the performance of the resulting system on new material could well be no better than 70 or 80 percent correct. Without exception, input is preprocessed before parts of speech are assigned by an HMM; this preprocessing is often only partially discussed in technical descriptions of part-of-speech taggers. The preprocessing copes with "unseen words," words that were never seen in the training data and for which the system therefore has no prior knowledge of possible parts of speech. It turns out that about half of the word types in the Brown corpus (Francis, 1964; Francis and Kucera, 1982), a carefully balanced representative corpus of American English, appear exactly once (about 32,000 out of 67,000 word types). This is consistent with Zipf's law, the empirical law that the frequency of a word type is inversely proportional to its rank. Nor can this problem be circumvented by some appropriately huge lexicon; a very large number of proper names appear on any newswire for the first time each day.

How can this unseen word problem be handled? One simple but quite effective technique is to tag each unknown word with the most likely tag given its last three letters—an empirical approximation to simple morphological analysis (Brill, 1992). A useful heuristic for proper nouns in most English text is to use capitalization, often combined with some other heuristics to correct for unknown words used at the beginning of sentences (Weischedel et al., 1993). The key point here is that these techniques for unseen words go beyond using purely stochastic techniques to using implicit and explicit linguistic knowledge, although in a trivial way, to get the job done.

## STOCHASTIC PARSING

All work on stochastic parsing begins with the development of the inside/outside algorithm (Baker, 1979), which generalizes the Baum-Welch algorithm for estimating HMMs to the estimation of parameters of stochastic context-free grammars.<sup>3</sup> Just as each iteration of the Baum-Welch algorithm over some training corpus improves estimates of the parameters of the underlying HMM, as judged by the criterion of maximal likelihood, so the inside/outside algorithm improves parameter estimates of an underlying probabilistic context-free grammar, judged by this same criterion.

However, straightforward application of the inside/outside algorithm does not appear to produce effective parsers; the best results to

---

<sup>3</sup> For a tutorial introduction to probabilistic context-free grammars and the inside/outside algorithm, see Jelinek et al. (1991) Lari and Young (1990).

date have resulted in parsers with about 35 percent correct parses on fully reserved test material in simple parsing tasks (Fujisaki et al. 1989; Sharman et al., 1990). Two problems appear to lie behind this failure. First, for realistic probabilistic context-free grammars (PCFGs) the number of parameters that must be estimated is very large; unless some *a priori* constraint is provided,  $n^3$  parameters must be estimated for a grammar with  $n$  nonterminal categories, categories that label not words but structures, like *noun phrase*, *verb phrase*, or *sentence*.

But a worse problem is that the objective function that the inside/outside procedure maximizes, namely the probability of the training corpus given the grammar, is in fact not the objective function that one wants to maximize to train effective parsers. For parsing the goal is to maximize assignment of the *correct grammatical structure*, to recursively subdivide the sentence correctly into its constituent grammatical parts, determined, say, by examining similar sentences in a treebank of hand-parsed sentences. Unfortunately, there is no reason to expect that a PCFG whose parameters are estimated by the inside/outside algorithm will assign structures that have the desired constituent structure.

In recent years a range of new grammatical formalisms have been proposed that some suggest have the potential to solve a major part of this problem. These formalisms, called *lexicalized* grammar formalisms, express grammars in which the entire grammar consists of complex structures associated with individual words, plus some very simple general rules for combining these structures. Such grammar formalisms include combinatory categorial grammars (CCGs), lexicalized tree-adjoining grammars (LTAGs), and link grammars (Joshi and Schabes, 1992; Steedman, 1993). In these lexicalized formalisms each word can be thought of as a tree fragment; the full grammatical analysis of a sentence is formed by specifying how and in what order the fragments associated with each word in a sentence combine. Words may themselves be ambiguous between different "parts of speech," here differing tree fragments. In these grammar formalisms the bulk of parsing a sentence is just deciding on which part of speech to assign to each word. Given this property of these grammar formalisms, perhaps some way can be found to extend the inside/outside algorithm appropriately so that its objective function maximizes the probabilities of strings of *part-of-speech tagged* words. If so, it is just a matter of extending the search space to handle the large number of complex part-of-speech structures of lexicalized grammars.<sup>4</sup>

---

<sup>4</sup> I thank Aravind Joshi for the above observation.

### Constraining the Inside/Outside Algorithm

Recently, a number of experiments have been performed that combine the inside/outside algorithm with some form of linguistic knowledge. In a recent experiment by Pereira and Schabes (1992), a modified version of the inside/outside algorithm was applied to a corpus that was manually annotated with a skeletal syntactic bracketing by the Penn Treebank Project (Brill et al., 1990; Marcus et al., 1993). In this experiment the I/O algorithm was modified to consider only PCFG rules that did not violate the skeletal bracketing of the corpus, zeroing out many of the  $n^3$  parameters a priori. The algorithm was then trained on a corpus of only 770 sentences collected in the Air Travel Information System (ATIS) domain (Hemphill et al., 1990). The evaluation was based on the "crossing brackets" parser evaluation metric of Black et al. (1991). This crossing-brackets measure counts the number of brackets inserted during parsing that are consistent with the correct bracketing of the sentence.<sup>5</sup> Without constraint, the algorithm achieved 35 percent bracketing accuracy on reserved test materials but achieved 90 percent bracketing accuracy when constrained by the annotated corpus.

### Conditioning PCFG Rules on Linguistic Context

One new class of models uses linguistic knowledge to condition the probabilities of standard probabilistic context-free grammars. These new models, which in essence augment PCFG grammar rules with probabilistic applicability constraints, are based on the hypothesis that the inability of PCFGs to parse with high-accuracy is due to the failure of PCFGs to model crucial aspects of linguistic structure relevant to the appropriate selection of the next grammar rule at each point within a context-free derivation. Probabilities in the standard stochastic context-free model are conditioned only on the type of nonterminal that the grammar is about to expand; the key idea of these new models is that this provides insufficient linguistic context to adequately model the probabilities of rule expansion. One such parser, that of Magerman and Marcus (1991a, 1991b), assumes that expansion of any nonterminal is conditioned on the type of nonterminal, the most likely part-of-speech assignments for the next several words

---

<sup>5</sup> Notice that this is a much easier measure than the percentage of sentences parsed correctly; if one of, say, 33 brackets is inconsistent in a given sentence, the sentence is 97 percent correct by the bracket measure and simply wrong by the sentences-correct measure.

in the parser's input stream, and the rule that has generated the particular nonterminal that the parser is trying to expand. For example, the rule "NP—pronoun" might have a different probability when it expands the NP in the rule "S — NP VP" than when it expands the NP in the rule "VP—NP NP"). Tested on a corpus of sentences from the Massachusetts Institute of Technology's Voyager domain (Zue et al., 1990), this parser correctly parsed 89 percent of a reserved test set. A sample list of sentences from this corpus, with length distribution typical of the corpus as a whole, is given in [Figure 3](#). Although the performance of this algorithm is quite impressive in isolation, the sentences in this corpus are somewhat simpler in structure than those in other spoken language domains and are certainly much simpler than sentences from newswire services that were the target of the parser evaluation discussed in the introduction to this chapter.

I'm currently at MIT  
Forty-five Pearl Street  
What kind of food does LaGroceria serve  
Is there a Baybank in Central Square  
Where is the closest library to MIT  
What's the address of the Baybank near Hong Kong  
What's the closest ice cream parlor to Harvard University  
How far is Bel Canto's from Cambridge Street in Cambridge  
Is there a subway stop by the Mount Auburn Hospital  
Can I have the phone number of the Cambridge City Hall  
Can you show me the intersection of Cambridge Street and Hampshire Street  
How do I get to the closest post office from Harvard University  
Which subway stop is closest to the library at forty-five Pearl Street

FIGURE 3 Sample sentences from the Massachusetts Institute of Technology's Voyager corpus.

On the other hand, a simple PCFG for this corpus parses a reserved test set with only about 35 percent accuracy, comparable to PCFG performance in other domains. If the probability of each rule is conditioned on both the current nonterminal and on the rule immediately above that gave rise to the current nonterminal, then performance improves to about 50 percent accuracy. Conditioning each rule on the expected part of speech of the next several words in addition increases performance to 87.5 percent accuracy. The key point here again is that combining a very simple stochastic framework with a little bit of linguistic knowledge greatly increases performance over each alone.

Many parsing techniques are now emerging that combine stochastic techniques with linguistic knowledge in a number of different ways. Again, as discussed briefly above, linguistic knowledge can be

encoded *explicitly*, perhaps in the form of a grammar, or *implicitly* within the annotations of an annotated corpus.

For combination with stochastic methods, so-called covering grammars can be used, grammars that provide at least one correct parse for sentences of interest but that may also produce spurious impossible parses. While these spurious parses would be a problem if the grammar were used with a purely symbolic parser, the hope is that when used within a stochastic framework, spurious parses will be of much lower probability than the desired analyses. One simple method for combining explicit linguistic knowledge with stochastic techniques is to use a stochastic technique to estimate the probability distribution for all and only the rules within the grammar, drastically limiting the number of parameters that need to be estimated within the stochastic model. While advocated by many researchers, this method suffers from the potential defect that it cannot model grammar rules that the grammar writer overlooked or that occur rarely enough that they were unseen in training materials. A somewhat more powerful method is to (1) use the grammar to generate all potential parses of a set of example sentences, (2) create a training set of trees by either hand picking the correct parse for each sentence or simply using all potential parses (which works far better than might be expected), and then (3) use the usage count of each grammar rule within this training set to provide an initial estimate of the parameters of the associated stochastic grammar, which might then be smoothed using the inside/outside algorithm.

If the annotation of a corpus takes the form of syntactic bracketing, the implicit knowledge encoded in the annotations of a corpus can be used in exactly the same way as explicit knowledge, to a first approximation. The key idea is that a grammar can be simply extracted from the corpus, along with counts for each rule, and then the methods discussed immediately above are applicable. In fact, most grammatical annotation provided in corpora to date is *skeletal*, providing only partial structures, so the use of smoothing techniques is crucial.

To what extent might combinations of linguistic knowledge with stochastic techniques improve the performance of parsers in the near-term future? Two experiments, both on a corpus of short sentences from computer manuals, cast some light here. In the first experiment (Black et al., 1992b), both an explicit grammar and an annotated corpus were used to build a stochastic parser that parsed 75 percent of the sentences in a reserved test set completely consistently with a hand-assigned bracketing. The second experiment (Black et al., 1993) is attempting to leave explicit grammars behind, using instead a very

rich set of linguistically relevant questions in combination with decision tree techniques. These questions examine not only syntactic properties, but lexical and class-based information as well, thus combining a much richer set of linguistic knowledge sources than any other model to date. The decision tree uses this set of questions to search for the grammar implicit in a very large hand-annotated corpus. Published reports of early stages of this work indicate that this technique is 70 percent correct on computer manual sentences of length 7 to 17, where, to count as correct, each parse must exactly match the prior hand analysis of that sentence in the test corpus, a more stringent test criterion than any other result mentioned here.

While this last experiment uses one uniform statistical technique, decision trees, to make all parsing decisions, some recent work suggests that effective parsing might be done by a suite of interacting parsing experts, each handling a particular grammatical phenomenon. Perhaps the clearest example of this is a recent technique to resolve the ambiguous attachment of prepositional phrases. Consider the sentence *I saw the man with the telescope*; here the prepositional phrase *with the telescope* might modify *the man*, meaning *I saw the man who had a telescope*, or it might modify the main verb *saw*, meaning *I used the telescope to see the man*. If the sentence were instead *I saw the planet with the telescope*, the prepositional phrase would certainly modify the main verb, but if it were *I saw the man with the hat* the prepositional phrase would clearly modify *the man*. Here, as in many other cases, it becomes clear that a decision about grammatical structure depends crucially on the properties of the lexical items themselves. A technique that uses likelihood ratios to compare the strength of association between the preposition and the main verb with the strength of association between the preposition and the preceding noun correctly assigns about 80 percent of prepositional phrases in sentences from the AP newswire with structure identical to the examples here (Hindle and Rooth, 1993). It is interesting to note that human judges, given the same information, do this task at about 85 to 87 percent accuracy. This experiment also points out the key role of lexical properties in deciding grammatical structure. Its success suggests that the crucial role of grammar is just to mediate the properties of lexical items themselves. This would suggest, as does the recent work on lexicalized grammars discussed above, that the words themselves are primary.

### Annotated Corpora

Before leaving the question of syntax, I would like to say a word about the production of annotated corpora themselves. There are, at

the moment, two large grammatically annotated corpora for English—the IBM/Lancaster Treebank (Garside et al., 1987) and the Penn Treebank (Marcus et al., 1993). As of this time, only materials from the second are generally available; they are distributed through the Linguistic Data Consortium; because of this, and my familiarity with this corpus, that is the focus here.

The Penn Treebank now consists of 4.5 million words of text tagged for part of speech, with about two-thirds of this material also annotated with a skeletal syntactic bracketing. All of this material has been hand corrected after processing by automatic tools. The two largest components of the corpus consist of over 1.6 million words of material from the Dow-Jones News Service, hand parsed, with an additional 1 million words tagged for part of speech and a skeletally parsed version of the Brown corpus (Francis, 1964; Francis and Kucera, 1982), the classic 1-million-word balanced corpus of American English. This material has already been used for purposes ranging from serving as a gold standard for parser testing to serving as a basis for the induction of stochastic grammars to serving as a basis for quick lexicon induction.

There is now much interest in very large corpora that have quite detailed annotation, assuming that such corpora can be efficiently produced. The group at Penn is now working toward providing a 3-million-word bank of predicate-argument structures. This will be done by first producing a corpus annotated with an appropriately rich syntactic structure and then automatically extracting predicate-argument structure, at the level of distinguishing logical subjects and objects, and distinguishing a small range of particular adjunct classes. This corpus will be annotated by automatically transforming the current treebank into a level of structure close to the intended target and then completing the conversion by hand.

## LEXICAL SEMANTICS AND BEYOND

We now turn to an area of very recent progress—lexical semantics. At initial inspection, it would appear most unlikely that statistical techniques would be of much use for either the discovery or representation of the meanings of words. Surprisingly, some preliminary work over the past several years indicates that many aspects of lexical semantics can be derived from existing resources using statistical techniques.

Several years ago it was discovered that methods from statistics and information theory could be used to "tease out" distinctions between words, as an aid to lexicographers developing new dictionar

ies (Church et al., 1991). As an example, consider the following: How could one distinguish the meaning of *food* and *water*? Figure 4 shows the mutual information score,<sup>6</sup> an information theoretic measure, between various verbs and between *food* and *water* in an automatically parsed corpus, where either *food* or *water* is the object of that verb or, more precisely, where one or the other is the head of the noun phrase which is the object of the verb. The corpus used in this experiment consists of 25 million subject-verb-object triples automatically extracted from the AP newswire by the use of a parser for unrestricted text. The mutual information score is high if the verb and noun tend to occur together and will tend toward 0 if the verb and noun occur together no more often than expected by chance. Because this measure is the log of a ratio, scores such as those shown in the table are quite high. What kinds of things can you do with food? Well, according to the AP newswire, you can hoard it, go without it, eat it, consume it, etc. With water, you can conserve it, boil it, ration it, pollute it, etc.<sup>7</sup> This indeed begins to reveal something about the meaning of

Associated with food (y=food; fy=2240)				Associated with water (y=water; fy=3574)			
I(x:y)	fy	fx	x	I(x:y)	fy	fx	x
9.62	6	84	hoard	9.05	16	208	conserve
8.83	9	218	go_without	8.98	18	246	boil
7.68	58	3114	eat	8.64	6	104	ration
6.93	8	722	consume	8.45	10	198	pollute
6.42	6	772	run_of	8.40	20	408	contaminate
6.29	14	1972	donate	8.37	38	794	pump
6.08	17	2776	distribute	7.86	6	178	walk_on
5.14	51	15900	buy	7.81	43	1320	drink
4.80	53	21024	provide	7.39	15	618	spray
4.65	13	5690	deliver	7.39	9	370	poison

Computed over Parsed AP Corpus (N = 24.7 million SVO triples)

FIGURE 4 What do you typically do with *food* and *water*? (Adapted from Church et al., 1991.)

<sup>6</sup> The mutual information statistic is a measure of the interdependence of two signals. It is defined as  $M(x,y) = I \log [P(x,y)/P(x)P(y)]$ .

<sup>7</sup> Because English contains a large number of so-called phrasal verbs, whenever the parser encounters a verb followed immediately by a prepositional phrase, as in *go without food*, the parser creates a potential phrasal verb, for example, *go\_without* and a triple where the object of the preposition (here *food*) is taken as the object of the putative phrasal verb (here *go\_without*).

these verbs, based on *distributional* properties of these words, that is, what other words they cooccur with.

More with food				More with water			
t	food	water	w	t	food	water	w
7.47	58	1	eat	-6.93	0	50	be_under
6.26	51	7	buy	-5.62	1	38	pump
4.61	31	6	include	-5.37	3	43	drink
4.47	53	25	provide	-5.20	0	29	enter
4.18	31	9	bring	-4.87	1	30	divert
3.98	21	3	receive	-4.80	0	25	pour
3.69	14	0	donate	-4.25	0	20	draw
3.55	13	0	prepare	-4.01	0	18	boil
3.31	13	1	offer	-3.89	0	17	fall_into
3.08	13	2	deliver	-3.75	1	20	contaminate

Computed over Parsed AP Corpus (N = 24.7 million SVO triples)

FIGURE 5 What do you do more *with food* than with *water*? (Adapted from Church et al., 1991.)

To differentiate these two words somewhat more sharply, one might ask what might be done more or less to one than the other. Here, an appropriate metric is the t-score, which contrasts the conditional probability of seeing food as object given a particular verb with the conditional probability of seeing water as object given that verb.<sup>8</sup> Figure 5 shows that one eats or buys food far more than water and that one pumps or drinks water far more than food. Perhaps surprisingly, these descriptions get quite close to the heart of the difference between *food* and *water*.

These experiments show that statistical techniques can be used to tease out aspects of lexical semantics in such a way that a human lexicographer could easily take advantage of this information. However, for computers to utilize such information, some kind of representation must be found to encode semantical information in the machine. What kinds of representations might be appropriate for automatic discovery procedures? Much research on automatic machine translation is now being done using the parallel French-English transcripts of the proceedings of the Canadian parliament. This corpus has been used to test a statistical technique that finds the most reliable local clue in context to tease apart different senses of the same word in the

<sup>8</sup> The formula computed is

source language, representing that meaning as the translation in the target language (Brown et al., 1991). An example of this technique is shown in [Figure 6](#). Here, the most useful single clue for the translation of an instance of the French word *prendre* is the particular word that occurs as the first noun to its right. As shown in [Figure 6\(a\)](#), the identity of this word provides, on average, 0.381 bits of information as to how to distinguish between two different senses of *prendre*. [Figure 6\(b\)](#) shows some of the French words that distinguish between one sense of *prendre* (*part, mesure, note, exemple, etc.*) and another (*decision, parole, connaissance, etc.*). As shown in [Figure 6\(c\)](#), the most likely translation into English for sense 1 is the verb *take*; for sense 2, the most likely translation is the English verb *make*. This technique has shown itself to be quite useful in current work in machine translation. As a technique for word sense disambiguation, it

Word:	prendre
Informant:	Right noun
Information:	.381 bits

(a) For *prendre* the noun to the right is maximally informative.

Sense I	Sense 2
part	decision
mesure	parole
note	connaissance
exemple	engagement
temps	fin
initiative	retraite

(b) Some French words that are informant values for each sense.

Pr(English —	Sense I)	Pr(English—	Sense 2)
to_take	.433	to_make	.186
to_make	.061	to_speak	.105
to_do	.051	to_rise	.066
to_be	.045	to_take	.066
		to_be	.058
		decision	.036
		to_get	.025
		to_have	.021

(c) Sense one translates as *take*, sense two as *make*.

FIGURE 6 The two senses of *Prendre* translate as *take* or *make*. (Adapted from Brown et al., 1991.)

has the clear drawback that often a word with multiple senses in language 1 has many of the same senses in language 2, so it can only be used when the target language *does* split the senses of interest.

Other researchers have found a richer and more robust representation for words in the internal representation used within WordNet, a large hand-built computer-based lexicon (Beckwith et al., 1991). Because WordNet is fundamentally computer-based, it can be organized as a large graph of different kinds of relations between words. These relations include not only relatively standard ones such as *X is a synonym of Y*, or *X is an antonym of Y* but also many other relations such as *X is a kind of Y* and *X is a part of Y*. Concepts within WordNet are represented by "synonym sets," sets of words that all share a core meaning. One simple representation of a meaning of a word, then, is just the synonym set, or *synset*, of the words that share that meaning.

This hierarchy has been used to investigate the automatic classification of verbs by the kinds of objects that they take, a first step toward determining the *selectional restrictions* of verbs automatically (Resnik, 1992). In this work, synonym sets are used to represent classes of objects in both the input and output of a program that computes a variant of the mutual information statistic discussed above. Using synonym sets for the output provides the general classification one seeks, of course. Using synonym sets for input as well has an important added advantage: it provides a solution to the sparse data problem that plagues work in lexical statistics. Many of the counts of verb-object pairs that make up the input to this program are very small and therefore unreliable, in particular given a corpus as small as the million-word Brown corpus (Francis, 1964; Francis and Kucera, 1982) used in this experiment. By pooling data for particular nouns into the synonym sets they fall into, much of this sparse data problem can be solved.

Figure 7 gives one example of the performance of Resnik's statistic. These are the highest-ranking synonym sets for objects of the verb *open*. The names of the synonym sets were hand assigned within WordNet. Figure 8 gives the single highest ranking synonym set for a list of common verbs. These two experiments show that a statistical approach can do surprisingly well in extracting major aspects of the meaning of verbs, given the hand encoding of noun meanings within WordNet. These experiments suggest that it might be possible to combine the explicit linguistic knowledge in large hand-built computational lexicons, the implicit knowledge in a skeletally parsed corpus, and some novel statistical and information theoretic methods to automatically determine a wide variety of aspects of lexical semantics.

The work described above is typical of much recent work in the

area of lexical discovery. Other recent work has focused on, for example, the use of distributional techniques for discovery of collocations in text (Smadja, 1993) and of subcategorization frames for verbs (Brent, 1993) and to uncover lexical semantics properties (Pustejovsky et al., 1993). Much other work has been done in this area; the references given here are typical rather than exhaustive.

SynSet Name	Typical Members
entrance	door
mouth	mouth
repository	store, closet, locker, trunk
container	bag, trunk, locker, can, box, hamper
time_period	tour, round, season, spring, session, week, evening,
morning, saturday	
oral_communication	discourse, engagement, relation, reply, mouth, program, conference, session
writing	scene, book, program, statement, bible, paragraph, chapter

FIGURE 7 Classes of things that are opened. (Adapted from Resnik, 1992.)

Verb	Most Highly Associated Object SynSet
ask	question
call	someone
climb	stair
cook	repast
draw	cord
drink	beverage
eat	nutrient
lose	sensory_faculty
play	part
pour	liquid
pull	cover
push	button
read	written_material
sing	music

FIGURE 8 "Prototypical" classes of objects for common verbs. (Adapted from Resnik, 1992.)

## A QUESTION FOR TOMORROW

While the focus above has been on the effectiveness of recent stochastic and statistical techniques, there is some evidence that this

effectiveness is due in large measure to the *empirical corpus-based* nature of these techniques rather than to the power of stochastic modeling. Surprisingly, symbolic learning techniques have performed as well as stochastic methods on two tasks considered above, despite the fact that they learn only simple symbolic rules, with only simple counting used during training, and then only to choose one potential rule over another. This raises the question of whether the effectiveness of the stochastic techniques above is essentially due to the fact that they extract linguistic structure from a large collection of natural data or is the result of their statistical nature. This issue, I believe, will be resolved in the next several years.

In one experiment a very simple symbolic learner integrated with a parser for free text produced a set of symbolic lexical disambiguation rules for that parser. The parser, running with the new augmented grammar, if viewed only as a part-of-speech tagger, operates at about 95 percent word accuracy (Hindle, 1989). What makes this result all the more surprising is that this parser works strictly left to right in a fully deterministic fashion. Recently, a very simple symbolic learning technique called error-based transformation learning was applied to the tagging problem (Brill, 1992). The resultant tagger operates with a set of only 160 simple rules, plus a table of the most likely tag in isolation for each word. The tagger begins by tagging each word with the most likely tag in isolation and then applies each of the 160 rules in order to the entire corpus. These rules are of the form *In a context X, change tag A to tag B*. This tagger operates at about 96 percent word accuracy.

This learning algorithm has also been applied to the problem of bracketing English text, with very surprising results (Brill, 1993). The learner begins by assuming that English is strictly right branching and then learns a set of rules using exactly the same learning technique as used in the tagger discussed above, except that here the potential environments for rule application are very simple configurations in the bracketed string, for example, *if some category X is preceded by a right paren then . . . or if a left paren falls between category X and category Y then . . .* There are only two possible rule operations that simply transform the binary branched trees with bracketings  $((A\ B)\ C)$  and  $(A\ (B\ C))$  into each other. The parser was trained on a small 500-sentence bracketed subset of material from the *Wall Street Journal*, of sentences less than 15 words in length, and acquired about 160 rules. Tested on a reserved test set of sentences of the same length, the parser bracketed 54 percent of the sentences completely consistently with the original bracketing; 72 percent of the sentences were bracketed with one bracketing error or less. Trained on only 250 sen

tences of length  $n$ ,  $n < 25$ , the parser again acquired about 160 rules and parsed a similar reserved test set at 30 percent of sentences bracketed correctly. In recent unpublished experiments this same technique was applied to the problem of labeling these bracketed but unlabeled structures, achieving about 95 percent correct labeling, by node.

Perhaps this learning technique will lead to an even more powerful stochastic method of some kind. What is unique about this learner is that each rule applies to the output of the previous rule. But perhaps it will turn out that the power of these methods comes from use of a corpus itself. Time will tell.

## ACKNOWLEDGMENTS

This work was partially supported by Defense Advanced Research Projects Agency (DARPA) Grant No. N0014-85-K0018, by DARPA and (AFOSR) jointly under Grant No. AFOSR-90-0066, and by Grant No. DAAL 03-89-C0031 PRI. Thanks to Eric Brill, Ken Church, Aravind Joshi, Mark Liberman, David Magerman, Yves Schabes, and David Yarowsky for helpful discussions. Thanks also to two anonymous reviewers for many excellent comments and suggestions.

## REFERENCES

- Baker, J. K. 1979. Trainable grammars for speech recognition. Proceedings of the Spring Conference of the Acoustical Society of America.
- Beckwith, R., C. Fellbaum, G. Gross, and G. Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, U. Zernik (ed.), Lawrence Erlbaum, Hillsdale, N.J., pp. 211-232.
- Black, E., S. Abney, F. Flickenger, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, February.
- Black, E., F. Jelinek, J. Lafferty, R. Mercer, and S. Roukos. 1992a. Decision tree models applied to the labeling of text with parts-of-speech. *Proceedings of the DARPA Speech and Natural Language Workshop*, February, pp. 117-121.
- Black, E., F. Jelinek, J. Lafferty, D. M. Magerman, R. Mercer, and S. Roukos. 1992b. Towards history-based grammars: Using Richer models for probabilistic parsing. *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*.
- Black, E., J. Lafferty, and S. Roukos. 1993. Development and evaluation of a broad coverage probabilistic grammar of English-language computer manuals. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*.
- Brent, M. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243-262.

- Brill, E. 1992. A simple rule-based part of speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy.
- Brill, E. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics.
- Brill, E., D. Magerman, M. Marcus, and B. Santorini. 1990. Deducing linguistic structure from the statistics of large corpora. Proceedings of the DARPA Speech and Natural Language Workshop, June, pp. 275-282.
- Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. A statistical approach to sense disambiguation in machine translation. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, Calif.
- Church, K. 1985. Stress assignment in letter to sound rules for speech synthesis. Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, pp. 246-253.
- Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of the Second Conference on Applied Natural Language Processing. 26th Annual Meeting of the Association for Computational Linguistics, pp. 136-143.
- Church, K., W. Gale, P. Hanks, and D. Hindle. 1991. Using Statistics in Lexical Analysis. AT&T Bell Laboratories Technical Memorandum.
- Cutting, D., J. Kupiec, J. Pederson, and P. Sibun. 1992. A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, ACL.
- Francis, W. N. 1964. A Standard Sample of Present-Day English for Use with Digital Computers. Report to the U.S Office of Education on Cooperative Research Project No. E-007. Brown University, Providence, R.I.
- Francis, W. N., and H. Kucera. 1982. Frequency Analysis of English Usage. Lexicon and Grammar. Houghton Mifflin, Boston.
- Fujisaki, T., F. Jelinek, J. Cocke, E. Black, and T. Nishino. 1989. A probabilistic method for sentence disambiguation. Proceedings of the 1st International Workshop on Parsing Technologies, Carnegie-Mellon University, Pittsburgh, Pa.
- Harris, Z. 1951. Methods in Structural Linguistics. University of Chicago Press, Chicago.
- Garside, R., G. Leech, and G. Sampson. 1987. The Computational Analysis of English. A Corpus-Based Approach. Longman, London.
- Hemphill, C., J. Godfrey, and G. Doddington. 1990. The ATIS spoken language systems pilot corpus. Proceedings of the Third DARPA Speech and Natural Language Workshop, February.
- Hindle, D. 1989. Acquiring disambiguation rules from text. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.
- Hindle, D., and M. Rooth. 1993. Structural ambiguity and lexical relations. Computational Linguistics, 19:103-120.
- Jelinek, F., J. D. Lafferty, and R. L. Mercer. 1991. Basic methods of probabilistic context-free grammars. Continuous Speech Recognition Group, IBM T. J. Watson Research Center.
- Joshi, A., and Y. Schabes. 1992. Tree adjoining grammars and lexicalized grammars. Tree Automata and Languages, M. Nivat and A. Podelski (eds.). Elsevier, New York.
- Lari, K., and S. J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language 4:35-56.
- Magerman, D., and M. Marcus. 1991a. Parsing the Voyager domain using pearl. Pro

- ceedings of the Fourth DARPA Speech and Natural Language Workshop, February.
- Magerman, D., and M. Marcus. 1991b. PEARL—A probabilistic chart parser. Proceedings, Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL), Berlin, April.
- Marcus, M., B. Santorini, M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313-330.
- Merialdo, B. 1991. Tagging text with a probabilistic model. Proceedings of ICASSP-91, pp. 809-812.
- Pereira, F., and Y. Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.
- Pustejovsky, J., S. Bergler, and P. Anick. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19:331-358.
- Resnik, P. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. Workshop Notes AAAI-92 Workshop in Statistically-Based NLP Techniques, July.
- Schabes, Y., M. Roth, and R. Osborne. 1993. Parsing the *Wall Street Journal* with the inside-outside algorithm. Proceedings, Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL), Utrecht, April.
- Sharman, R.A., F. Jelinek, and R. Mercer. 1990. Generating a grammar for statistical training. Proceedings of the Third DARPA Speech and Natural Language Workshop, February.
- Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143-178.
- Steedman, M. 1993. Categorial grammar, Lingua, 90:221-258.
- Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19:359-382.
- Zue, V., J. Glass, D. Goodine, H. Leung, M. McCandless, M. Philips, J. Polifroni, and S. Seneff. 1990. Recent progress on the VOYAGER system. Proceedings of the Third DARPA Speech and Natural Language Workshop, June.

# The Future of Voice-Processing Technology in the World of Computers and Communications

*Yasuo Kato*

## SUMMARY

This talk, which was the keynote address of the NAS Colloquium on Human-Machine Communication by Voice, discusses the past, present, and future of human-machine communications, especially speech recognition and speech synthesis. Progress in these technologies is reviewed in the context of the general progress in computer and communications technologies.

## EXPECTATIONS FOR VOICE INTERFACE

Many of us have now experienced conversation with computers by voice. When I first experienced it more than a quarter-century ago I was a little shocked, even though it was just a simple conversation. It gave me the feeling that I was actually communicating with a person; that is, I had the feeling that the machine had a personality. And it was fun! Since then I have come to believe that the voice interface is very special. It should not be regarded simply as one alternative for ordinary human-machine interface.

Speech communication is natural, common, and easy for us. Practically speaking, there are at least three advantages to a voice interface. First, it is an easy interface to use. Second, it can be used while the user is engaged in other tasks. Third, it accommodates multimodal interfaces. But it seems to me that there are other important or essen

tial reasons for voice communication for humans, and this thought has driven me to pour energy into speech recognition and synthesis for a long time—for the past 35 years.

I think we are still only halfway to our goal of an advanced or smart interface. From here on the scientific path to our goal only gets steeper.

Today we live in the age of information. Five billion people can benefit from an economically efficient society supported by computers and communications. This will become truer as we become more information oriented.

The NEC Corporation recognized the importance of integrating computers and communication long ago and adopted "C&C" (computer and communications) as its corporate identity in 1977. In the future, C&C will become an indispensable tool for everyone—for many it already has. In this kind of environment, C&C must be easy to use. It must provide a friendly, natural, smart interface for people. Thus, the voice interface is an important component for the C&C of the future.

Recently, we have seen significant progress in speech-processing technologies, especially in the United States. At NEC we have also made a little progress with large-vocabulary recognition (10,000 words), speaker-independent continuous speech recognition (1000 words), and naturally spoken language understanding. Some of this technology is close to being commercialized.

I will not spend a lot of time discussing speech synthesis, but I must make one important comment. Prosodic features such as pauses, stresses, and intonation are related to semantic structure and emotion and are difficult to convey in written messages. But the role of these features will become very important in sophisticated smart interfaces.

## VOICE INTERFACE IN THE C&C INFORMATION SOCIETY

There are many applications for new C&C technologies. In the area of public information services there are applications in employment, securities, risk management, taxation, health, education, marriage, entertainment, shopping, and even funeral arrangements. For business and personal use, there are applications in document design, presentations, publications, information access, and inter/intraoffice communication. And personal applications are also important: text preparation (dictation), e-mail, telephone, scheduling, and personal database management.

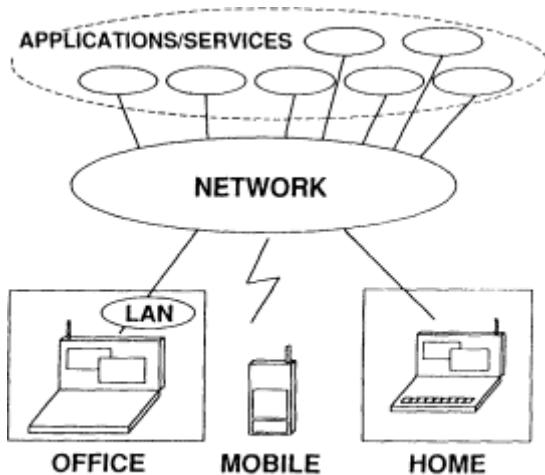


FIGURE 1 C&C information society.

In the future information society, people will access and exchange a variety of services, applications, and information through networks. Access will be obtained through terminals in offices and homes and even through mobile terminals, as shown in Figure 1.

The speech-processing function for a voice interface can be located in those terminals or in central machines. I anticipate that this function will generally be handled by terminals. I will touch on this again later, but until now it has been easiest to locate this function centrally because of space or cost constraints. Next I will discuss some examples of centrally located speech-processing functions.

At NEC, we commercialized a speaker-independent recognition system, the SR-1000, and first used it for banking information services in 1979. Fourteen years ago is ancient history in voice-processing technology, but hindsight is 20-20, and to understand the future it is often helpful to know the past. This system could recognize 20 words (numerals plus control words, spoken separately) and could process 64 telephone lines simultaneously.

Our current model, the SR-3200, is also used for banking services as well as order entry and reservations. In all cases, centrally located speech recognizers must be speaker-independent, which makes producing them more difficult.

For a moment I would like to go back even further. In 1959 we developed a spoken digit recognizer. I believed that word recognition for a small vocabulary was within reach, and we challenged ourselves to produce a voice dialing mobile telephone system. This

experimental system was another example of a centralized voice interface. We made it in 1960, and long-run field evaluations were done inside our laboratory building. Although the system was primitive, and of course not commercialized, we did confirm a future for the voice interface.

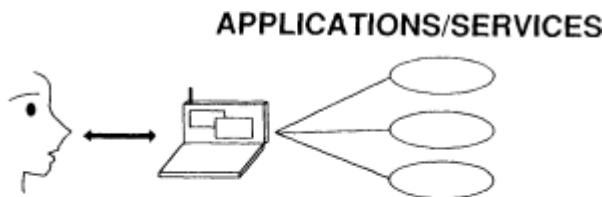
When we demonstrated the system at an EXPO in Japan, many customers were interested in using the recognizer in their work. However, most of them needed more than a simple word recognizer, and this forced me to begin research into continuous speech recognition.

Today, no one thinks about centralized machines because it is easy to install a voice dialing function inside the telephone. Soon you will be able to carry a voice dial mobile telephone in your pocket.

### A FRIENDLY, SMART INTERFACE

A human-machine interface can be friendly and smart only if it is customized to individuals. By "smart" I mean that it is multimodal, accepts spoken language, and allows spontaneous and incomplete utterances. The personal terminal will become a tool for accessing and exchanging information from an infinite variety of applications and services, as shown in [Figure 2](#). To customize terminals, however, it will be necessary for the terminal to know about the user, that is, to have knowledge of a person's job, application preferences, and acoustic and linguistic speech characteristics.

Similarly, the interface will also require knowledge of the applications or services being used or accessed. Because of difficulties with centralizing knowledge about users, this information will prob



- **CUSTOMIZED TO A PERSON, SMART**
- **INTERFACES BETWEEN A PERSON AND APPLICATIONS/SERVICES**

FIGURE 2 Terminal as a personal tool.

ably need to be contained in the terminal, and all speech recognition functions may also need to be carried out at the terminal.

In our speech recognition research and development at NEC, we have pursued terminal-type speech recognizers. The DP-100 was our first commercial speech recognizer. It was a speaker-dependent continuous speech recognizer with a 120-word vocabulary that we produced in 1978. Its operation is based on dynamic programming (DP) matching, which we developed in 1971. At that time we were very excited because DP matching significantly improved recognition performance. The DP-3000 is our current model; it has a 500-word vocabulary. To accomplish this we developed a high-speed DP matching LSI to reduce hardware size; it also tripled recognition performance.

The history of speech recognition research at NEC has been a series of studies and struggles for applications. The first application of the DP-100 was in parcel sorting, a job that requires hands-free data entry. Before we accepted contracts I conducted many joint studies with material-handling machine manufacturers and did frequent field experiments at end users' sorting yards. To promote sales of the system, I also visited many end users, including an airline company in the United States. The Federal Aviation Administration was one of our earliest customers for speech recognition products and tested our technology for applications in flight control training in 1979.

Other examples of the DP-series speech recognizer in use include inspection data entry for used-car auctions, meat auctions, and steel sheet inspection. The DP-3000 is also being used in a voice-controlled crane.

DP-series recognizers have been used primarily in eyes-busy or hands-busy situations. Some might say that we have spent too much time on speech recognition and that we started too soon, but I do not agree. Speech is difficult to handle, and speech-processing technological advances are closely tied to device technologies and market needs. So someone must make repeated efforts to show new practical possibilities and obtain funding for continued research.

From this perspective I do not think that we started too early or that our struggles have been useless. Through these experiences we have learned a lot. But clearly, more basic scientific research is needed because we do not yet have sufficient knowledge of speech.

As I mentioned earlier, the speech interface is for people, and I have always believed that it should be applied to personal systems or personal computers. We developed a speech recognition board for PCs in 1983 that was intended to popularize the speech interface. This board was a speaker-dependent word recognizer that had a 128word vocabulary and was priced at \$250. It saw limited sales, mainly

because the software environment was inadequate. Recently, because of increases in CPU (central processing unit) power, speech recognition is finding its way into personal computers, and a good software environment is developing.

We had a similar experience in the area of speech synthesis. In 1965 I made a computer-controlled terminal analog speech synthesizer (TASS), and it was really big. To control it you needed a computer that was 10 times bigger than the synthesizer. LSI technology enabled us to shrink the size of the huge TASS, and in 1982 we commercialized a cartridge-type text-to-speech-type voice synthesizer. It was designed as a plug-in option to NEC's PCs. Initially, the synthesizer was not very popular at all. We subsequently included it as a standard feature of our PC, and the result was that we successfully marketed hundreds of thousands of speech synthesizers. Now though, even in a PC, we can easily accomplish speech synthesis with software and the CPU. To get 10 times as much linguistic processing power, we only have to add one VLSI (very-large-scale integrated) chip. In the year 2000 all of this will be on a small fraction of a single chip.

## VOICE INTERFACE AND VLSI TECHNOLOGY

From now on, integrating a voice interface function in terminals will be related to progress in VLSI technology. Let us take a look at progress in VLSI technology. In the 2000s, with advances in device technology, gate numbers in chips will increase a hundred times, and clock frequencies will increase about 10 times. Most principal processing functions will be integrated on a single chip. Further advances in CAD (computer-aided design) technology will enable the development of large-scale application-specific ICs (ASICs). In the 2000s, because of these improvements, large-scale ASICs with 50 Megagates will be easily designed in less time than it takes to design today's 10k-gate ASICs.

Roughly speaking, our speech recognition processor, which was used at the TELECOM'91 demonstration, performed 1000-word continuous speech recognition at 1 GIPS processing with 100 Megabits of memory. If we apply the previous estimate for VLSI technological advances to speech recognition processing, we will see 10,000-word continuous speech recognition with 100 GIPS CPUs and 1 Gbit memories in the 2000s, even if we assume a 10 to 1 increase in hardware for vocabulary improvement. This would correspond to three CPU chips and one memory chip. If we are clever enough to reduce increased hardware requirements by one-third, which does not appear to be an

unreasonable expectation, we will be able to produce a two-chip high-performance speech recognizer.

This kind of VLSI progress was beyond the imagination of an engineer who wanted to make a phonetic typewriter 30 years ago. But even back then we youngsters believed in the future progress of electronics, and such belief enabled us to challenge some big mountains. In 1960 we developed a phonetic typewriter that exemplifies the technological improvements I have been talking about. It was built by NEC for Kyoto University as an experimental tool, and it could recognize a hundred kinds of Japanese monosyllables. Though solid-state technology was very new, I elected to use fully transistorized circuitry in this system, and it took 5000 transistors and 3000 diodes. Roughly speaking, this corresponds to 1/250 of CPU chip space in today's technology and only 50k primitive instructions per second.

## FUTURE RESEARCH AND DEVELOPMENT ISSUES

Progress in VLSI technology is going to enable us to do an enormous amount of speech processing in a chip, and as a result the contents of—the methods for—speech processing become more important. There will be two issues in future voice technology research and development (R&D). The first is the development of speech interfaces for PCs and personal terminals. The second is basic research toward the smart interface. In both cases, R&D needs to be oriented toward, and with an awareness of, applications.

Because of increases in processing power, multimedia functions, including images and sound, are becoming popular in PCs and terminals. Recently, pen input has been presented as a hopeful solution for making the interface more friendly. Personally, I prefer to use fingers. But, in addition to pens and fingers, the importance of speech input and output needs to be recognized. Today, speech recognition performance is not perfect, but continued development of practical speech interfaces and platforms is necessary to extend areas of application and to popularize PCs and terminals.

With a speech interface, as shown in [Figure 3](#), three processing layers need to be considered. They are the application interface, the operating system, and the device for speech processing. The implementation of speech processing will depend heavily on the processing power of the CPU.

Today, NEC manufactures the Hobbit chip developed by AT&T. This chip is suitable for utilizing a pen input function in PCs. NEC also produces the V810, NEC's original multimedia processing chip.

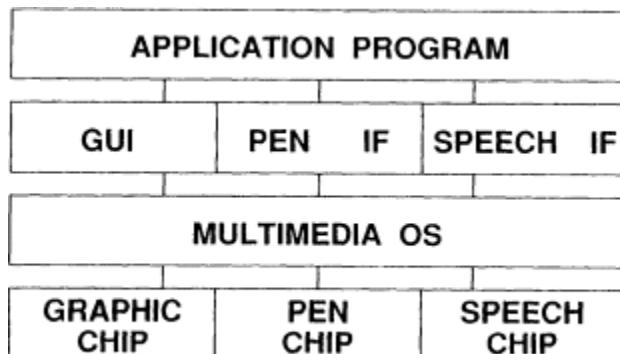


FIGURE 3 Speech interface for PCs.

However, we believe that a speech-processing chip is essential to fully implement the functions of a speech interface.

The second issue for future R&D is the smart interface. Areas of importance for the smart interface include speech-to-text conversion, or dictation, and a conversational spoken language interface. Important functions include the allowance of spontaneous and incomplete utterances. Also, because it is often impossible to understand or correctly interpret what was intended by an utterance without knowledge of the situation and speaker, a dialogue function becomes an important component of the smart interface. It will also be important to utilize knowledge of applications, service, and users' characteristics in the implementation of a smart interface.

To produce a smart interface, a fusion of speech and language (i.e., spoken language processing) is necessary. So far, language-processing people have not been so interested in speech, while speech people have been interested in language processing. That has been an unrequited love. The importance of this fusion between language and speech is now being recognized. As I stated earlier, some speech phenomena, including pause, stress, and intonation, are valuable in this fusion because they are gifts from speech to language.

In 1983 I organized a new laboratory that integrated two groups: a speech group and a language group, which until then had been working independently. Our natural-language-processing research has also been working toward machine translation. In 1985 NEC commercialized a machine translation system, PIVOT, that translates between Japanese and English text. To do this, we developed an intermediate language called PIVOT Interlingua, which is suitable for

multilingual translation. In fact, we recently added translation in Korean, Spanish, and French to the system.

With this fusion between speech and language in mind, I would like to mention three future research needs. The first is to develop more precise models of human spoken language production and perception. The second is to develop sophisticated models for computer learning and recognition. Third, we need to develop a knowledge base of language and domains. Research on these topics involves many areas of science and technology; psychology, cognitive science, bioscience, linguistics, mathematics, computer science, and engineering, to name a few. However, interdisciplinary collaboration between them will be extremely important, and I hope that people in the United States will help initiate this collaboration. For this collaboration there is a good common vehicle, and that is automatic interpretation.

### TOWARD AUTOMATIC INTERPRETATION

When NEC first advocated the concept of C&C, we recognized that the future C&C would be large and complex. As a result, it must be intelligent enough to be reliable, self-controlling, and self-organized or autonomous, and it must offer people smart interfaces, which we called MandC&C (huMan and C&C). In 1983 we suggested automatic interpreting telephony as the ultimate goal of C&C. We think that the purpose of research into automatic interpreting telephony is not merely to realize an interpreting machine. This research involves the most important scientific and technological issues for future C&C.

At TELECOM'83 in Geneva, we demonstrated such a system. This was actually a small experimental system to suggest that future R&D be directed toward automatic interpretation. You can imagine my pleasure when, right after TELECOM'83, the ministry of post and telecommunications in Japan announced the start of a national project on automatic interpretation. They say that this was the founding of ATR Interpreting Telephony Research Laboratories, and I offer them my congratulations on having recently developed a very sophisticated system. I am glad, as a proposer and supporter, that automatic interpretation research is now receiving worldwide attention including work in the United States, Germany, and Japan. In 1991 NEC demonstrated its improved automatic interpretation system, INTERTALKER, at TELECOM'91 in Geneva. INTERTALKER is an integrated speaker-independent, continuous speech recognition, text-to-speech conversion system that utilizes the PIVOT machine translation system. The system recognizes Japanese and English speech and

translates into English, Japanese, French, and Spanish, as shown in [Figure 4](#).

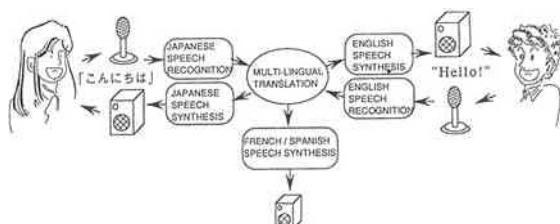


FIGURE 4 Automatic interpretation system INTERTALKER.

Automatic interpretation will help us realize the dream of human global communication that bridges the language gap. It is also an important goal for voice technology. Through the development of automatic interpreting telephony, we will be able to obtain and develop the technologies necessary for maintaining a C&C society.

In concluding, I want to reemphasize the importance of collaboration. It is the collaboration of interdisciplinary research areas: (1) collaboration between speech, language, and other disciplines and (2) collaboration between those involved in science and technology. It must also be an international collaboration.

## Author Biographies

JONATHAN ALLEN received his Ph.D. from MIT in 1968, after six years at AT&T Bell Laboratories in Human Factors Engineering. He then joined the faculty of MIT, where he is Professor of Electrical Engineering and Computer Science. His interests include speech synthesis and recognition, as well as computer-aided design for integrated circuit synthesis. Since 1981 he has been Director of the Research Laboratory of Electronics at MIT. He is a Fellow of the IEEE, and past President of the Association for Computational Linguistics.

BISHNU S. ATAL is Head of the Speech Research Department at AT&T Bell Laboratories, Murray Hill, New Jersey. He has been with Bell Laboratories since 1961, and his research there has covered a wide range of topics in acoustics and speech. He has made major contributions in the field of speech analysis, synthesis, and coding. His research in linear predictive coding of speech has established linear predictive analysis as one of the most powerful speech analysis techniques for applications in speech coding, recognition, and synthesis. His current research interests include low bit rate speech coding, and new robust and accurate methods for automatic speech recognition. Dr. Atal is a member of the National Academy of Engineering and the National Academy of Sciences. He is a Fellow of the Acoustical Society of America and of the IEEE. He received the IEEE ASSP Soci

ety Award in 1993 for contributions to linear prediction of speech, multipulse, and code-excited source coding.

MADELEINE BATES is the Assistant Department Manager for Speech and Natural Language Processing at BBN Systems and Technologies. She is responsible for the technical and administrative direction of research and development efforts, including spoken language understanding, human-machine interfaces incorporating natural language processing, and the development of evaluation methodologies for those areas. She has more than 20 years of experience in research, development, and application in many aspects of artificial intelligence and computational linguistics, including syntactic processing of English by computer, speech understanding, knowledge acquisition for NL systems, computer assisted language instruction, interfaces to data bases, and human factors studies. She is a past president of the Association for Computational Linguistics.

ROLF CARLSON joined the Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden in 1969. He has been active in the department since that time, with the exception of two years of employment at MIT in 1978-1979 and 1990-1991. He also has academic merits in general linguistics and phonetics. His first studies concerned perception of speech, e.g., models of vowel perception. In 1977 he received his Doctor of Science degree on the subject "Perception and Synthesis of Speech." His main activity has been to create possibilities for converting text to speech. This research formed a base of the company Infovox, which was created in 1983. Since 1970, he has published extensively on speech synthesis, speech perception, and general phonetics. His recent activity includes the development of a dialog system based on speech technology. He is currently Advisory Editor of the Journal of Phonetics.

PHILIP R. COHEN is a Senior Computer Scientist in the Artificial Intelligence Center at SRI International. He is also a Principal Researcher at the Center for the Study of Language and Information at Stanford University, and a Consulting Associate Professor with the Symbolic Systems Program at Stanford. After receiving his Ph.D. in Computer Science in 1978 from the University of Toronto, he worked at Bolt Beranek and Newman, Oregon State University, and the Fairchild Camera and Instrument Corporation before joining SRI in 1984. His research interests include multimodal human-computer interaction, natural language processing, spoken dialogue, and intelligent com

puter agents. Dr. Cohen is a Fellow of the American Association for Artificial Intelligence.

JAMES FLANAGAN is Vice President for Research and Board of Governors Professor in Electrical and Computer Engineering at Rutgers University. Flanagan joined Rutgers after extended service in research and research management positions at AT&T Bell Laboratories. Flanagan holds the S.M. and Sc.D. degrees in Electrical Engineering from MIT. He specializes in voice communications, computer techniques, and electroacoustic systems, and has authored papers, books, and patents in these fields. Flanagan is a Fellow of the IEEE, the Acoustical Society of America, and the American Academy of Arts and Sciences. He is a member of the National Academy of Engineering and of the National Academy of Sciences.

SADAOKI FURUI received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University in 1968, 1970, and 1978, respectively. Since joining the Electrical Communications Laboratories, Nippon Telegraph and Telephone Corporation in 1970, he has been working on speech analysis, speech recognition, speaker recognition, and speech perception. From 1978 to 1979 he was with AT&T Bell Laboratories, Murray Hill, New Jersey. He is currently the Research Fellow and the Director of Furui Research Laboratory at NTT Human Interface Laboratories. He is a Fellow of the IEEE and a Distinguished Lecturer of the IEEE Signal Processing Society. He has received awards from several institutions, including the IEEE.

LYNETTE HIRSCHMAN heads the Speech and Natural Language group at MITRE Corporation in Bedford, Massachusetts, where she is also responsible for the corporate Human-Computer Interaction initiative. She received her Ph.D. in formal linguistics from the University of Pennsylvania in 1972 and has been involved in the field of natural language processing and, more recently, in spoken language understanding. She has been active in both the speech and natural language evaluation efforts, and chaired the original Multisite ATISData Collection Working group (MADCOW) that coordinated the spoken language data collection and evaluation methods for the Air Travel Information System (ATIS) task.

FREDERICK JELINEK received the S.B., S.M., and Ph.D. degrees in electrical engineering from MIT in 1956, 1958, and 1962 respectively. From 1972 to 1993, he was with the Computer Sciences Department, IBM

Thomas J. Watson Research Center, Yorktown Heights, New York, where he managed research on automatic recognition (transcription) of speech. Since November 1993, he has been the Director of the Center for Speech Processing and Professor of Electrical and Computer Engineering, with joint appointments in Cognitive Science and Computer Science at Johns Hopkins University. His principal interests are in speech recognition, language processing, and information theory. He is the author of "Probabilistic Information Theory" (New York: McGraw-Hill, 1968). Dr. Jelinek was the recipient of the 1971 Information Theory Group Prize Paper Award and was recognized in 1981 as one of the top 100 innovators by TECHNOLOGY magazine.

CANDACE KAMM received her Ph.D. in Cognitive Psychology from UCLA, where she was a research associate at UCLA School of Medicine, studying speech recognition and loudness perception of normal and hearing-impaired listeners. She joined AT&T Bell Laboratories in 1982, evaluating automatic speech recognition technology, and she has been at Bellcore since 1984. At Bellcore, her work has focused on user-interface design for voice applications, telephone speech database collection, and the application of artificial neural networks to speech recognition problems. She currently directs the Speech Recognition Applications Research Group at Bellcore.

YASUO KATO received the B.E. degree in Electrical Engineering from the Tokyo Institute of Technology in 1958. After graduation, he joined NEC Corporation and has been working for NEC's Central Research Laboratories. He was involved in the research and development of digital signal processing, speechband-compression, speech synthesis, and speech recognition. He was a visiting researcher in speech communications at MIT during 1964 and 1965. From 1980 to 1986, he was General Manager of NEC's C&C Systems Research Laboratories and became Vice President and Director in 1986. He is now Executive Vice President for Research and Development at NEC Corporation.

STEPHEN E. LEVINSON received the B.A. degree in Engineering Sciences from Harvard in 1966, and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Rhode Island in 1972 and 1974, respectively. From 1966 to 1969 he was a design engineer at Electric Boat Division of General Dynamics in Groton, Connecticut. From 1974 to 1976 he held a J. Willard Gibbs Instructorship in Computer Science at Yale University. In 1976, he joined the technical staff of AT&T Bell Laboratories in Murray Hill, New Jersey, where he conducted research in the areas of speech recognition and cybernetics. Dr. Levinson

is Head of the Linguistics Research Department at AT&T Bell Laboratories where he directs research in speech synthesis, speech recognition, and spoken language translation. Dr. Levinson is a member of the Association for Computing Machinery, a fellow of the Institute of Electrical and Electronic Engineers and a fellow of the Acoustical Society of America. He is a member of the editorial board of *Speech Technology*, a founding editor of the journal *Computer Speech and Language* and a member of the IEEE Signal Processing Society Technical Directions Committee.

HARRY LEVITT obtained his Ph.D. in Electrical Engineering from the Imperial College of Science and Technology, London, England, in 1964. He then crossed the Atlantic to become a member of the AT&T Bell Laboratories technical staff where he did research on binaural hearing, computer-assisted adaptive testing, speech synthesis, and telephone aids for people with hearing loss. In 1969, he joined the faculty of The City University of New York, where he is now Distinguished Professor of Speech and Hearing Sciences. His current research interests include digital hearing aids, video signal processing for deaf people, and other applications of assistive technology.

MARK LIBERMAN received his Ph.D. in 1975 from MIT. He is currently Trustee Professor of Phonetics at the University of Pennsylvania, where he also holds an appointment in the Department of Computer and Information Science. He came to Penn in 1990 after fifteen years at AT&T Bell Laboratories, where he served as Head of the Linguistics Research Department. Liberman is a member of the editorial advisory boards of the journals *Cognition*, *Computer Speech and Language*, and *Speech Communication*. He is the chair of the Data Collection Initiative of the Association for Computational Linguistics and the director of the Linguistic Data Consortium.

JOHN MAKHOUL is a Chief Scientist at Bolt Beranek Newman Inc. (BBN), Cambridge, Massachusetts. He is also an Adjunct Professor at Northeastern University and a Research Affiliate at the MIT Research Laboratory of Electronics. An alumnus of the American University of Beirut and the Ohio State University, he received a Ph.D. from MIT in 1970 in electrical engineering, with specialization in speech recognition. Since that time, he has been with BBN directing various projects in speech recognition, spoken language systems, speech coding, speech synthesis, speech enhancement, signal processing, and neural networks. Dr. Makhoul is a Fellow of the IEEE and of the Acoustical Society of America.

MICHAEL MARCUS has been RCA Professor in Artificial Intelligence at the University of Pennsylvania since 1987. Prior to that, he was in the Linguistics and Artificial Intelligence Research Department at AT&T Bell Laboratories. He received his Ph.D. from the Artificial Intelligence Laboratory, MIT, in 1978. His research in natural language deterministic parsing (*A Theory of Syntax for Natural Language Understanding*, MIT Press, 1980) is discussed in a number of textbooks in artificial intelligence and natural language understanding. His current research interests include statistical and corpus-based natural language processing, the processing of grammatical processing in humans, and cognitive science in general.

ROBERT C. MOORE received his Ph.D. in Artificial Intelligence from MIT in 1979. He is currently a Principal Scientist in the Artificial Intelligence Center of SRI International, Menlo Park, California, having held previous positions as Director of SRI's Natural Language Research Program and Director of SRI's Computer Science Research Centre in Cambridge, England. Dr. Moore's current research interests lie in the area of integration of speech recognition and natural language processing, particularly how to use natural language processing to improve the accuracy of speech recognition.

RYOHEI NAKATSU received the B.S., M.S., and Ph.D. degrees in electronic engineering from Kyoto University in 1969, 1971, and 1982, respectively. After joining NTT in 1971, he worked on speech recognition technology. Particular interests at that time included isolated and connected word recognition. He moved to NTT Yokosuka Laboratories in 1980. There he took part in a project to develop a voice response and recognition system to be used for banking services. Since 1990, he has been with NTT Basic Research Laboratories in Musashino, where he is currently Director of the Information Science Research Laboratory. His current research interests include auditory perception, speech production, visual perception and discourse understanding. He is also a member of the IEEE and the Acoustical Society of Japan.

JOHN A. OBERTEUFFER is Vice President for New Business Development at Voice Processing Corporation in Cambridge, Massachusetts. Voice Processing Corporation is a developer and vendor of continuous speech recognition technology for over-the-telephone and personal computer applications. Prior to joining VPC, Oberteuffer was editor of ASR News, a monthly newsletter that reports on market and technical developments in the advanced speech technology industry. Before

founding Voice Information Associates (publisher of ASR News) in 1989, Oberteuffer held senior marketing, technical, and management positions in several emerging technology-based companies, including Iris Graphics, which he cofounded, and which is now a subsidiary of Scitex Corporation. Oberteuffer received his doctorate in physics from Northwestern University and his bachelor's and master's degrees from Williams College.

SHARON L. OVIATT is a Senior Cognitive Scientist in the Artificial Intelligence Center at SRI International. After receiving her Ph.D. in Experimental Psychology from the University of Toronto in 1979, she taught at the University of Illinois, Oregon State University, and the University of California before joining SRI International in 1987. Her research interests include next generation human language technology, design of spoken language and multimodal systems, modality effects in communication (speech, pen, keyboard, etc.), and research design and evaluation for emerging technologies.

LAWRENCE R. RABINER has worked for AT&T Bell Laboratories since 1962 and has been involved in research in the area of digital signal processing with applications to speech processing. Currently, as Director of the Information Principles Research Laboratory, he directs research in speech coding, speech synthesis, speech recognition, speaker recognition, language translation, acoustics, communications, image processing, interactive systems, and systems using digital signal processor chips. He is a member of both the National Academy of Engineering and National Academy of Sciences.

DAVID B. ROE is Head of the Applied Speech Research Department at AT&T Bell Laboratories, Murray Hill, New Jersey. He has been with AT&T since 1986. He is active both in long-range speech research, such as spoken language identification and a Voice English/Spanish Translator, and in business applications of speech recognition. In the latter capacity, he has consulted with several AT&T organizations that have developed speech recognition products and services. David received a Ph.D. in experimental low-temperature physics in 1976 from Duke University and joined Bell Laboratories in 1977 after a year of teaching at Duke.

RONALD W. SCHAFER received the B.S.E.E. and M.S.E.E. degrees from the University of Nebraska, Lincoln, in 1961 and 1962, respectively, and the Ph.D. degree from MIT in 1968. From 1968 to 1974 he was a member of the Acoustics Research Department, AT&T Bell Laborato

ries, Murray Hill, New Jersey. In 1974 he joined the faculty of the Georgia Institute of Technology, where he is Institute Professor of Electrical and Computer Engineering and holder of the John O. McCarty Chair. His current research interests include speech processing, video processing, and nonlinear signal processing systems. He is coauthor of three widely used textbooks on digital signal processing. Dr. Schafer is a Fellow of the IEEE and the Acoustical Society of America, and he is a member of the National Academy of Engineering. In 1992 he received the IEEE Education Medal.

RICHARD SCHWARTZ is a Principal Scientist at Bolt Beranek and Newman Inc. (BBN), Cambridge, Massachusetts. He joined BBN in 1972 after receiving an S.B. in electrical engineering from MIT. Since then, he has worked on phonetic recognition and synthesis, narrowband speech coding, speech enhancement in noise, speaker identification and verification, speech recognition and understanding, neural networks, and statistical message processing. He is currently the lead scientist on the ARPA project on spoken language systems.

CHRIS SEELBACH has provided management direction and consulting for the development of voice processing systems, services, and markets since 1982. He was President of Verbex and founded Voice Systems to commercialize voice processing systems. He has provided consulting and research services on international voice processing market and product development. Mr. Seelbach is currently Executive Vice President and COO of Viatel, the leading international value-added telephone light carrier. At Viatel, Mr. Seelbach is leading the introduction of the latest voice processing services to Viatel's global customers. He has a B.S. from the U.S. Naval Academy and an M.B.A. from Columbia University.

YOSHITAKE SUZUKI received the B.E. and M.E. degrees in electrical engineering from Waseda University, Japan, in 1979 and 1981, respectively. After joining NTT Electrical Communication Laboratories in 1981, he worked on designing speech recognition systems. He was a visiting scholar at the University of Washington in 1988, where his study included designing digital architectures for artificial neural networks. Mr. Suzuki is currently a senior research engineer at NTT Human Interface Laboratories, Tokyo, Japan. He is also a member of the IEEE and the Acoustical Society of Japan.

CLIFFORD WEINSTEIN is Group Leader of the Speech Systems Technology Group at MIT Lincoln Laboratory. He received the S.B., S.M.,

and Ph.D. degrees from MIT and has been with Lincoln Laboratory since 1967. He has made technical contributions and has been responsible for initiation and leadership of research programs in speech recognition, speech coding, speech enhancement, packet speech communications, integrated voice/data communication networks, finite-word-length effects in digital signal processing, and radar signal processing. Since 1986, Dr. Weinstein has been the U.S. Technical Specialist on the NATO RSG10 Speech Research Group. Since 1989, he has been Chairman of the Coordinating Committee for the ARPA Spoken Language Systems Program, which is the major U.S. research program in speech recognition and understanding. Dr. Weinstein was elected a Fellow of the IEEE in 1993. Also in 1993, he was elected to the Board of Governors of the IEEE Signal Processing Society. From 1991 to 1993, he was Chairman of the Signal Processing Society's Technical Committee on Speech Processing, and from 1976 to 1978 he was Chairman of that society's Technical Committee on Digital Signal Processing.

JAY G. WILPON is a Distinguished Member of the Technical Staff in the Speech Research Department at AT&T Bell Laboratories. Since joining AT&T in June 1977, Mr. Wilpon has concentrated on problems in automatic speech recognition. He has published extensively in this field and has been awarded several patents. His current interests lie in keyword spotting techniques, speech recognition training procedures, and determining the viability of implementing speech recognition systems for general usage over the telephone network. Mr. Wilpon holds the B.S. in Mathematics, the A.B. degree in Economics, both from Lafayette College, and the M.S. degree in Electrical Engineering/Computer Science from Stevens Institute of Technology. In 1987 Mr. Wilpon received the IEEE Acoustics, Speech and Signal Processing Society's Paper Award for his work on clustering algorithms for use in training automatic speech recognition systems. Mr. Wilpon is Chairman of the IEEE Digital Signal Processing Society's Speech Committee.



# Index

## A

Abbreviations, pronunciation of, 142-143  
Acoustic interactions, 122  
inventory elements, 126  
models/modeling, 26, 36, 85, 95, 117, 122, 182-183, 476  
Kratzenstein's resonators, 78, 80  
phonetics, 85, 95  
speech recognition, 64, 182-183  
speech synthesis, 137  
terminal analog synthesizer, 117  
Advanced Research Projects Agency.  
*See also* Airline Travel Information System  
Benchmark Evaluation summaries, 224-225  
common speech corpora, 181-182  
continuous speech recognition program, 175-176, 181-182  
Human Language Technology Program, 108

research funding, 349  
Speech and Natural Language Workshop, 359  
Speech Language Understanding Program, 262-263  
Spoken Language Systems program, 218-219, 220, 230, 232-233, 250, 254, 255-256, 262-263, 265, 405  
Resource Management corpus, 181-182, 184, 185, 188, 376, 377  
*Wall Street Journal* corpus, 184-185, 186, 187  
Air traffic control, 365-366  
Airline Travel Information System (ATIS), 376  
context-dependent utterances, 61  
corpus, 61, 184-185, 234, 219, 250, 256, 257-258, 491  
degree of difficulty, 383-385  
error rates, 252, 486  
human performance on, 162  
interactive dialogue, 227, 228, 233

- language understanding methods, 258, 268  
N-best filtering in, 227  
and Online Airline Guide, 219  
order in problem solving, 229  
overview of, 46  
recognition errors, 261, 262  
spontaneous input, 234  
template-based approach, 259  
understanding errors, 262
- Algorithms  
ambiguity-handling, 56  
assessment of, 391-392, 405, 409-412  
Baum-Welch training (forward-backward), 178-179, 199, 202-207, 489  
beam search, 202, 210, 212, 214  
compression, 83, 381  
databases, 405-409  
inside/outside, 263-264, 489-490, 491  
intonation contours, 45  
large vocabularies, 307  
learning, 249, 250, 263-264  
nonlinear interpolation, 97  
part-of-speech assignment, 143  
probabilistic parsing, 56  
prosodic phrase generation, 146, 147, 151  
reference resolution, 57  
robustness, 391, 392, 405, 412-416  
search, 180-181, 189, 199, 202, 208, 209, 264-265  
speech processing, 21, 393  
speech recognition, 28, 409-411, 412, 417-418, 431, 468, 469  
speech synthesis, 468  
Stack, 202, 208  
standardization, 7  
text-to-speech, 25  
Viterbi, 173, 180, 199, 202, 208-209, 210, 213  
voice coding, 7  
wordspotting, 404, 431
- Allophone models, 182  
American Automobile Association, 354  
Ameritech, 291, 292, 293, 302  
Analog-to-digital converter, 22-23, 189, 350
- Analysis-by-synthesis systems  
articulatory data, 125  
and automatic learning, 127  
bit-rate reduction and, 23  
and "break index" data, 148-149  
defined, 118  
linear predictive coding, 24, 26-27, 119  
PSOLA methods 119-120  
source-filter technique, 119  
in speech analysis, 26-27  
in speech recognition, 30  
text-to-speech conversion as, 136
- Apple Macintosh, 52
- Applications of voice communications.  
*See* Assistive technology for disabled persons  
Deployment of applications  
Military and government applications  
Telecommunications  
Telephony  
air travel information systems, 46, 85-86, 162  
aircraft pilots, 40, 41, 44, 45, 359, 365, 509  
assessment criteria, 409-410  
automatic teller machines, 86  
computer-aided instruction, 151  
databases for, 406-408  
baggage handlers, 40  
consumer electronics  
programming, 43, 353  
development environment, 400-401  
driving instructions, 354  
economic impact of, 280  
expectations for, 505-506  
force feedback glove, 98, 101  
foreign language learning, 44

- hands/eyes-busy tasks and, 39-41  
 in information society, 506-508  
 limited keyboard/screen option and, 41-43  
 medical report generation, 351  
 motor vehicle navigation, 44  
 multimodal systems, 63-64  
 parcel sorters, 40, 509  
 portable computers, 43  
 reading lessons, 44  
 real-time support, 403-405  
 smart interfaces, 508-509  
 speech interface technology, 347-356  
 success factors, 289-290  
 security, 7  
 speech recognition, 28-29, 30-32, 81, 275-282, 283-284, 318, 377-379, 451, 457, 458, 471, 508-510  
 stock quotation service, 283, 292, 293, 299, 354, 437, 438, 439  
 technology trends, 399-405  
 text-to-speech synthesis, 43, 109, 280, 282, 302, 354, 451  
 user-friendly, 508-510  
 video/audio conferencing system, 99-100  
 and VLSI technology, 40, 54, 510-511  
 voice input, 39-44  
 voice output, 44-45  
 wire installers, 40
- Army.  
*See also* Military and government applications  
 Avionic Research and Development Activity (AVRADA), 362  
 Communications and Electronics Command (CECOM) program, 361-362  
 Articulatory models, 88, 95, 117, 118, 120, 122, 124-125, 152-153, 461-463, 476  
 Artificial intelligence, 484  
 Artificial neural networks, 2, 21, 124, 190, 191-193, 381, 479
- Assembler language, 399-400, 401  
 Assistive technology for disabled persons  
 assistive listening devices, 315-316  
 augmentative and alternative communication, 130, 335-337  
 captioning, 314-315, 322-323  
 carpal tunnel syndrome, 43  
 categories of sensory aids, 316  
 cochlear implants, 314, 328-331, 332-333  
 computer-assisted instruction, 336  
 deaf-blind, 327  
 direct stimulation of auditory system, 328-331  
 dysarthric speech, 337  
 extracochlear implant, 329-330  
 eyeglass speechreader, 320-322  
 hearing aids and assistive listening devices, 278, 311, 312, 315-318, 328-331, 332  
 hearing impaired, 43, 292, 302-304, 312, 314-333  
 limitations of, 318  
 mobility control, 312  
 noise reduction, 331-333  
 reading machines for blind, 349  
 research and development efforts, 313-314  
 sound/speech spectrograph, 319, 325, 349  
 with speech/language disabilities, 311, 313, 325  
 speech recognition, 275-279  
 speech processing for sightless people, 279, 313, 329, 333-335, 349  
 speechreading cues, 320-321, 325, 327, 328  
 tactile sensory aids, 314, 324-328  
 talking books, 333  
 Telephone Relay Services (TRS), 292, 302-303, 322  
 teletypewriters, 323

- Terminal Device for the Deaf (TDD), [302](#), [314](#), [322](#)  
 text telephone, [322](#), [323](#), [324](#)  
 use trends, [312](#)-[313](#)  
 visible speech translator, [319](#)-[320](#)  
 visual sensory aids, [319](#)-[324](#)  
 voice control, [278](#)-[279](#), [313](#), [337](#)  
 voice output devices, [334](#), [336](#)  
 ATR International, [9](#), [10](#), [42](#), [83](#), [108](#)-[109](#),  
[115](#), [119](#), [128](#), [130](#), [176](#), [513](#)
- AT&T  
 800 Speech Recognition service, [298](#)  
 articulatory models, [124](#)-[125](#)  
 Directory Assistance Call Completion, [292](#), [301](#)-[302](#)  
 control of network fraud, [291](#)  
 government funding, [349](#)  
 hidden Markov models, [175](#)  
 Hobbit chip, [511](#)-[512](#)  
 HuMaNet, [454](#)  
 Intelligent Network, [292](#), [298](#)  
 operator services automation, [291](#), [292](#),  
[293](#)  
 packet data network [XUNET], [99](#)  
 speech synthesis technology, [107](#)-[108](#),  
[112](#), [124](#)-[125](#)  
 spoken language translation, [9](#), [10](#), [130](#)  
 Telephone Relay Services (TRS), [292](#),  
[302](#)-[303](#)  
 telephone speech database, [407](#)  
 Terminal Device for the Deaf (TDD), [302](#)  
 text-to-speech system, [348](#)-[349](#)  
 voice dialing system, [300](#), [383](#)-[385](#)  
 Voice English-Spanish Translation  
 (VEST), [10](#)  
 Voice Interactive Phone, [292](#), [300](#)-[301](#)  
 voice processing vision, [285](#)-[286](#)  
 Voice Prompter, [292](#)  
 Voice Recognition Call  
 Processing (VRCP), [292](#), [293](#)-[295](#),  
[383](#)-[385](#)
- voice response systems market, [281](#), [282](#)  
 Who's Calling service, [282](#)  
 wordspotting techniques, [305](#)  
 Auditory modeling, [24](#), [26](#), [91](#), [92](#), [94](#), [97](#)
- B**
- Bandwidth compression, [81](#)  
 Basilar membrane filtering, [97](#)  
 Bell, Alexander Graham, [77](#)-[78](#)  
 Bell Atlantic, [291](#)  
 Bell Mobility (BM), [302](#)  
 Bell Northern Research (BNR), [176](#), [282](#),  
[283](#), [292](#), [293](#), [294](#), [295](#), [299](#),  
[383](#)-[386](#), [437](#), [438](#), [439](#)  
 Bell System, [6](#)  
 Bellcore, [291](#)-[293](#)  
 Bigram models, [201](#), [209](#), [211](#), [213](#), [214](#),  
[222](#)  
 Bit rates  
   and image processing, [101](#)  
   speech coding and, [23](#), [24](#), [81](#), [83](#)-[84](#)  
   text-to-speech synthesis, [29](#), [77](#)  
 Bolt, Beranek, and Newman (BBN) Systems and Technologies  
 ATIS, [46](#), [261](#)  
 Delphi system, [259](#)  
 directory service, [438](#)  
 hidden Markov models, [175](#)  
 N-best filtering and rescoring, [267](#)  
 word lattice parsing, [265](#)  
 "Break index" data, [147](#), [148](#)
- C**
- C cross compiler, [399](#)-[400](#)  
 Cambridge University, [176](#)  
 Carnegie-Mellon University (CMU).  
*See also* Airline Travel Information System  
 ATIS, [46](#), [261](#)  
 dialogue state information, [229](#)

- HMM applications in speech recognition, 175
- multilingual systems, 42, 83
- Phoenix system, 258, 259, 260
- recursive transition networks, 222
- spoken language translation, 9
- Cepstrum techniques, 28, 86, 178, 182-183, 476
- Chaos, 21, 26
- Classification and decision tree techniques, 152
- Classification and regression tree techniques, 147
- CNET, 130
- COCOSDA group, 130
- Coding. *See* Linear Predictive coding; Music coding; Speech coding
- Compact disc technology, 334
- Compound words, 142, 147
- Compression
  - algorithms, 83, 381
  - bandwidth, 81
  - image, 99
  - speech, 23, 83, 474
  - two-channel amplification, 332
- Computation
  - models of language, 78, 81, 86, 90-91
  - of pronunciation, 139
  - research needs, 30
  - speech recognition systems, 30
  - speech synthesis, 137
  - speed, 19-20, 97
  - teraflop capability, 97
  - Viterbi algorithm, 173
- Computer-aided tools, 21, 510
- Computer Search and Language*, 2-3
- Consonants
  - alveolar flapped, 142
  - clusters, 138, 140
  - modeling, 123
- Consortium for Lexical Research, 241
- Context-oriented clustering, 126
- Corpora
  - Airline Travel Information Service, 61, 184-185, 219, 250, 256, 257-258, 491
  - American English, 489, 495
  - annotated, 493, 494-495
  - Brown, 489, 495, 499
  - common speech, 181-182
  - connected digit, 184-185
  - IBM/Lancaster Treebank, 495
  - large linguistic, 447
  - Penn Treebank, 241, 491, 495
  - Resource Management, 181-182, 184, 185, 188, 376, 377
  - optimization, 113
  - telephone speech, 408-409
  - Wall Street Journal*, 184-185, 186, 187
  - Creak (vocal), 122
  - CRIM, 176
  - Cross-word effects, 182
  - CSELT, 176
  - CSTR, 130
  - Currency, pronunciation of, 143
  - Cybernetics, 445-446, 448-449
- D**
- Databases.
  - See also* Corpora
  - algorithms, 405-409
  - for applications, 406-408
  - dialect considerations, 409
  - interfaces, 240, 252
  - large tagged, 152
  - natural language interfaces, 240
  - NTIMIT, 409
  - Official Airline Guide, 46, 219
  - relational, 53-54
  - for research, 405-406
  - remote access to, 42, 44, 278, 296-299, 348, 349, 351
  - retrieval system, product quality, 57
  - simulated telephone lines, 408-409
  - speech, 387, 405, 407, 468, 472
  - StockTalk, 383-386, 437, 438, 439
  - WordNet, 499
- DEC, 130
- Decision criteria, 305

- Defense Advanced Research Projects Agency. *See* Advanced Research Projects Agency
- Deployment of applications
  - degree of difficulty and, 375-386
  - hardware considerations, 381, 382-383
  - language understanding task dimensions and, 379-381
  - military technology transfer, 367-369
  - obstacles to, 374-375
  - procedure for, 386-388
  - speech recognition task dimension and, 377-379
  - speech synthesis task dimensions and, 381-382
  - system integration requirements, 383
  - technical challenges in, 280-281
- Desert Storm, 360
- Dialogue
  - capabilities, 85, 403-405
  - clarification/confirmation, 56, 62-63
  - continuous speech, 431-432
  - convergence of styles, 60
  - conversational dynamics, 431-432
  - engineering constraints, 387-388
  - feedback and confirmation, 437-438
  - finite-state transition network, 63, 85
  - flow, 435-436
  - grammars, 62, 63
  - interaction and, 61-63
  - models, 62-63
  - natural language, 17, 56, 61-63
  - quantity of text and, 381
  - real-time processing function, 403-404
  - research, 63, 66
  - robustness of, 66
  - speech recognition, 63
  - spoken language systems, 47, 60, 61-63, 66, 229
  - talk-over, 431
  - task-specific voice control, 452
- transcript, 433-434
- Dictation devices, automatic, 50, 77, 81, 426, 428, 437-438.
  - See also* Text, typewriters
- Digital
  - encryption, 83
  - speech coding, 25, 82-83, 85
  - filtering, 19
  - telephone answering machines, 7-8
- Digital computers.
  - See also* Digital signal processors and speech signal processing, 19, 78, 81, 189, 393-396
  - and microelectronics, 19-21, 81
- Digital-to-analog converter, 23, 398
- Digital signal processors/processing applications, 350, 400-401
  - capabilities, 391, 393-394
  - development environment, 399-405
  - distributed control of, 404-405
  - floating-point, 383, 394-396
  - growth of, 19, 78, 81
  - integer, 383
  - for LSP synthesis, 398
  - mechanisms, 393
  - microphone arrays, 97
  - technology status, 393-396
  - transputer architecture, 396, 397
  - workstation requirements, 189
- Digitizing pens, 52
- Diplophonia, 122
- Discourse
  - natural language processing, 246
  - and prosodic marking, 149-151
  - speech analysis, 145, 149-151
  - in spoken language systems, 227-230
  - in text-to-speech systems, 145
- Dragon Systems, Inc., 176, 380, 401, 402
- Dynamic grammar networks, 265-266
- Dynamic time warping (DTW), 28

**E**

Electronic mail (e-mail), 8, 306, 381  
 ESPRIT/Polyglot project, 123, 129, 130,  
 406

Etymology  
 proper name estimates, 92  
 trigram statistics, 141

Experiments  
 capabilities, 32  
 real-time, 32  
 research cycle, 183-184  
 Extralinguistic sounds, 122

**F**

Fallside, Frank, 1-3, 445-446

Fast Fourier Transform (FFT), 28, 84, 475

FAX machines, 5

Feature  
 extraction, 177-178  
 delta, 182-183  
 vectors, 182-183

Federal Aviation Administration,  
 365-366, 509

Federal Bureau of Investigation, 367

Fiber optics, 6

Filter bank outputs, 28, 475

Filters/filtering  
 adaptive, 332, 414, 456-457  
 basilar membrane, 97  
 digital, 19  
 high-pass, 332  
 language understanding component for,  
 22

linear time-varying, 477-478

N-best, 227, 267

transverse, 415

Flex-Word, 292

Fluid dynamics, principles in speech pro-  
 duction, 87-90

Force feedback glove, 98, 101

Foreign language.  
 See also Multilingual systems;

Spoken language translation  
 learning, 44  
 word incorporation in text-to-speech  
 systems, 138  
 Formants, 122-123, 125  
 Fractals, 21, 26  
 Frequency-domain representation, 24, 476

**G**

Gestural inputs, 65

Government. *See* Military and govern-  
 ment applications

Grammars  
 ambiguity, 380  
 bigram, 179  
 combinatory categorical, 490  
 context-free, 264, 461, 490, 491-494  
 covering, 493  
 dialogue, 62, 63  
 dynamic grammar networks, 265-266  
 features-value structures in, 264  
 finite-state, 266, 379-380  
 formalisms, 490  
 hand-coded linguistic, 483  
 lexicalized, 490  
 lexicalized tree-adjoining, 490  
 Markov, 179-180  
 modeling, 28, 63  
 natural language understanding and,  
 37-38, 264, 380, 491-494  
 perplexity, 180, 185, 229, 378  
 probabilistic context-free, 491-494  
 size, 37-38  
 speech analysis and, 28, 36-38  
 speech recognition, 36-37, 41-42, 63,  
 81, 85-86, 179-180, 185-186, 265-66  
 statistical n-gram, 183, 224  
 training speech, 179-180, 185-186  
 trigram, 141, 179-180, 183  
 unification, 461

Graphical user-interface.  
*See also* User interfaces

- growth of, 108
  - guidelines, 66-67
  - hierarchical menu structure, 54
  - speech compared with, 54-55
  - strengths, 52-53
  - weaknesses, 53, 57-58
- H**
- Handwriting
    - recognition, 402-403
    - screen-based channel, 64
  - Hardware technology.
    - See also* Digital signal processors;
    - Microcomputers;
    - Personal computers;
    - Workstations
    - advances in, 391
    - CISC architecture, 392, 392
    - Hobbit chip, 511-512
    - Intel x86 series, 392
    - microprocessors, 383, 391, 392-393, 396
    - Motorola 68000 series, 392
    - RISC chips, 383, 392-393
    - speech-processing equipment and systems, 383, 396-405, 510-511
    - V810 multimedia processing chip, 511-512
  - Health Interview Survey on Assistive Devices, 312
  - Hidden Markov models (HMM)
    - bigram, 201, 211, 213, 214
    - defined, 171-173
    - estimation of statistical parameters of, 199, 202-208
    - feature extraction, 177-178
    - fenonic case, 207
    - grammar-state-transition table, 266
    - limitations of, 189-190
    - Markov chains, 170-171, 172
    - and mel-frequency cepstral coefficients, 178
    - neural nets combined with, 193-194
  - part-of-speech tagging, 487-488, 490
    - phonetic, 166, 173-175, 178-179, 182, 188
  - and semantics, 221
  - speaker recognition systems, 30, 85
  - speech recognition, 28, 30, 85, 170-175, 177-178, 199, 200-208, 377, 394, 396, 397, 478-479
    - speech variability and, 28, 415-416
    - and talker verification, 86
    - three-state, 172
    - theory development, 175
    - training and analysis, 30, 178-179, 181-182, 478-479
  - trellis representation, 203, 208, 212
  - trigram, 201-202, 212, 213-214
  - unigram, 210
  - Viterbi algorithm and, 210
  - word models, 179
  - wordspotting, 397
  - Human-human communication
    - conversational dynamics, 431-432
    - language imitation, 60
    - repair rates, 260
    - studies, 50-51
- I**
- IBM, 9, 175, 349, 380, 495
  - Image compression, 99
  - Image processing, 78, 101
  - Information processing
    - in auditory systems, 91, 94
    - speech technologies, 453
  - Information retrieval, 54-55, 57
  - INFOVOX, 130
  - Institute for Defense Analyses, 175, 234-235
  - Institute for Perception Research, 127
  - INTELLECT, 57
  - Integrated Services Digital Network (ISDN), 84

- Interaction.  
*See also* User interfaces  
 acoustic, 122  
 and dialogue, 61-63  
 failures, cost of, 426-427  
 large-vocabulary conversational, 101-102  
 natural language, 51-57, 58  
 speech recognition, 36  
 spoken language systems, 51-57, 60, 61-63  
 system requirements for voice communications applications, 383
- Intonation  
 contours, 45  
 cues, 432  
 parts-of-speech distinctions and, 151  
 structures, 129  
 models, 127
- J**
- Joysticks, 52
- K**
- Karlsruhe University, 83  
 Klatt, Dennis, 111, 123  
 Kratzenstein's acoustic resonators, 78, 80  
 Kurzweil Applied Intelligence, 380
- L**
- Language  
 acquisition, theory of, 2  
 generation, 38, 241  
 imitation, 60  
 processing, 239;  
*see also* Natural language processing  
 variability, 380
- Language modeling.  
*See also* Natural language  
 bigram, 201, 209, 211, 213, 214, 222, 461  
 computational, 78, 81, 86, 90-91
- etymology estimates for proper names, 92  
 future of, 307  
 research needs, 26, 29  
 speech recognition, 29, 81-82, 90-91, 168-169, 183, 263, 307  
 speech synthesis, 128  
 statistical, 263-264, 461, 472-473  
 trigram, 92, 183, 209-210, 212, 213-214, 461  
 by users, 60
- Laryngalization, 122  
 Law enforcement, 367  
 Lexicons, 138, 140, 141-142, 178-179, 188, 296, 499
- LIMSI, 176
- Linear predictive coding  
 analysis by synthesis, 24, 26-27, 119  
 mapping code book, 128  
 code-excited (CELP), 24, 26, 83, 101  
 mixed-excitation (MELP), 24  
 multipulse excited (MPLPC), 24, 26  
 pitch-excited, 24  
 robustness of, 97  
 self-excited (SEV), 24, 26  
 AND SPEECH ANALYSIS, 575
- Linguistic analysis, 59-60, 259, 263, 382, 461, 484
- Linguistic Data Consortium, 181, 241, 252
- Linguistics.  
*See also* Parsing;  
 Semantics;  
 Syntax  
 after-thoughts, 256  
 consonant cluster, 138  
 discourse-level effects, 149-151  
 English lexical stress system, 141-142  
 letter-to-sound relationships, 138, 140-141  
 metonymy, 256-257, 257-258  
 morphophonemics, 141-142  
 orthographic conventions, 142-143

- parts-of-speech assignment, 143, 151  
prosodic marking, 145-149  
spontaneous speech, 255-258  
vocalic suffixes, 139  
word-level analysis, 138-139
- M**
- Machine translation, 240  
Markov chains, 170-171, 172  
Masking, time and frequency, 84, 93-94, 177-178  
Massachusetts Institute of Technology  
  articulatory models, 124  
  ATIS, 46, 227, 261  
  HHMs for speech recognition, 176  
  MITalk, 123  
  multilingual synthesis, 130  
  speech synthesis, 111, 123, 124  
TINA language understanding system, 222, 223, 259  
Matsushita, 130  
MCI, 300  
Mel-frequency cepstral coefficients (MFCC), 178, 182-183  
Message processing, 241, 251  
Microcomputers. See also Personal computers  
  computation speed, 19-20, 97  
  device density, 20  
  digital signal processing, 19  
  projected advances in, 102-103  
  speech processing and, 19-20, 81, 396-399  
Microelectronics  
  chip densities, 102  
  digital computation and, 19-21  
  research, 21  
  revolution, 108  
  speech signal processing, 19-20  
Microphones  
  applications, 86-87, 102
- autodirective arrays, 86-89, 96, 97, 99-100, 102  
beam forming systems, 87, 88, 99  
characteristics, 414  
digital signal processors, 97  
directional, 333, 414-415  
electret, 87, 88, 97, 102  
environmental variation in speech input, 412-413, 460  
in hearing aids, 331-332, 333  
noise reduction, 331-332, 414-415  
reflection and reverberation, 414  
speaker distance from, 414  
and speech recognition, 379, 414  
technology projections, 102  
three-dimensional, 96, 97, 99-100  
track-while-scan mode, 87, 89
- Microsoft Windows, 52
- Military and government applications.  
*See also* Advanced Research Projects Agency;  
*other government agencies*  
Agent's Computer, 367  
Air Force, 359, 365  
air traffic control, 365-366  
aircraft carrier flight deck control and information management, 363  
Army, 359, 360-363  
combat team tactical training, 364-365, 366  
Command and Control on the Move (C2OTM), 360-361  
law enforcement, 367  
Multi-Role Fighter, 365  
Navy, 363-365  
Pilot's Associate system, 365  
Soldier's Computer, 360, 361-362, 367  
SONAR supervisor command and control, 363-364  
technology transfer issues, 367-369  
voice control of systems, 360, 362, 365

- Mixed-mode communication. *See* Multi-modal systems
- Models/modeling.  
*See also* Hidden Markov models; Language modeling  
 acoustic, 26, 36, 64, 85, 95, 117, 122, 182-183, 476  
 allophone, 182  
 articulation, 88, 95, 117, 118, 120, 122, 124-125, 152-153  
 auditory, 24, 26, 91, 92, 94, 97  
 bigram, 201, 209, 211, 213, 214  
 computational, 78, 81, 86, 90-91  
 consonants, 123  
 context-dependent, 182, 246  
 cross-word effects, 182  
 dialogue, 62-63  
 grammar, 28, 63, 380  
 intonation, 127  
 Klatt, 123  
 left-to-right, 175  
 natural language understanding, 238-253, 262-264  
 noise excitation, 122  
 phonetic, 173-174, 190-191, 193  
 prosody, 117  
 segmental, 125, 173-174, 190-191, 193  
 signal, 19, 101  
 sinusoidal, 24  
 sound source, 462  
 source/system, 22, 118, 120-122  
 speech perception, 26  
 speech production, 22  
 speech recognition requirements, 168-169  
 speech synthesis, 109, 116-130  
 speech variability, 176  
 spoken language systems, 48  
 stochastic segment, 190-191  
 trigram, 201-202  
 vocal tract, 95, 118, 122, 124, 125  
 wave propagation, 26  
 word, 179, 207  
 Modulation theory, 26
- Morphemes, 137, 139, 140  
 Morphology, speech synthesis, 110, 111, 112, 113, 137, 141-142, 489  
 Morphs, 138-139, 140  
 Motorola, 383, 392  
 Mouse, 52, 350-351, 402-403  
 Multilingual systems.  
*See also* Foreign language; Spoken language  
 translation;  
 Telephony  
 future of, 513-514  
 INTERTALKER, 513-514  
 Japanese kana-kanji preprocessor, 403  
 MITalk, 130  
 PIVOT, 512-513  
 speech synthesis, 42, 101, 117, 129-130, 151-152
- Multimodal systems.  
*See also* User interfaces  
 advantages of, 426  
 error avoidance, 64  
 error correction, 64  
 HuMaNet, 454  
 referent determination difficulties, 61  
 robustness, 64  
 situational and user variation, 64-65  
 synergistic integration of sensory modalities, 100-101, 102  
 user interfaces, 32, 56, 63-65
- Multiprocessing, 21
- Music coding, 84
- N**
- N-Best interface, 217, 221, 226, 233  
 N-Best Paradigm, 191, 193  
 National Institute of Standards and Technology, 377  
 Natural language.  
*See also* Speech recognition; Spoken language  
 anaphora, 55  
 dialogue, 17, 56, 61-63  
 interaction, 51-57, 58

- modeling, 128;
  - see also* Language modeling
- research directions, 56, 241
  - and speech recognition systems, 17, 262-267, 388
  - and spoken-language systems, 59-61
  - spontaneous speech, 59, 263
  - typed, 57
- Natural language processing
  - ambiguity-handling algorithms, 56
  - applications, 240, 241, 250-253
  - clarification/confirmation subdialog, 56, 62-63
  - components of, 243-250
  - constraints on, 17, 59-61, 262-268, 388, 482-484, 491
  - context modeling, 57, 246
  - cooperating process view of, 248-250
  - database interfaces, 240
  - domain model extraction, 250
  - evaluation, 250-252, 483
  - grammars, 264, 483
  - history of, 240-241
  - ideal systems, 55-56
  - inputs to, 241-243
  - integration architecture, 265-267
  - machine translation, 240
  - menu-based system, 56
  - outputs, 243
  - parsers, 59, 247, 483, 489-495
  - portability of systems, 252
  - pragmatics, 246, 250
  - problems, 241-243
  - product quality database retrieval system, 57
  - prosodic information in, 268-269
  - reasoning, 246-247
  - reference resolution algorithms, 57
  - research directions, 460-461, 500-502
  - response planning and generation, 246-247
  - rule-based, 482-484
  - semantics, 245-246, 247, 250, 486, 495-500
  - simplified systems, 247-248
  - speech processing and, 460-461
  - state of the art, 252-253
  - statistical techniques, 484
  - strengths, 55-56, 58
  - syntactic, 244-245, 247
  - training [learning], 56, 57, 58, 249, 250, 252
  - verbal repair detection, 269
  - weaknesses, 56-57
- Natural language understanding.
  - See also* Linguistics;
  - Speech recognition;
  - Spoken language understanding
  - accuracy/error rates, 47, 251, 252, 255, 261, 262, 388
  - applications, 379-381
  - architecture, 485-487
  - background, 238-239
  - current capabilities, 10, 506
  - defined, 239
  - grammar, 37-38, 263, 380, 491-494
  - language variability and, 380
  - models of, 238-253, 262-264
  - off-the-subject input and, 287, 380, 388
  - part-of-speech tagging, 487-489
  - preprocessing and, 489
  - search process, 248-249
  - speech constraints in, 268-269
  - stochastic parsing, 489-495
  - task difficulty and, 379-381
  - TINA system, 222
  - vocabulary size and, 37-38
  - unknown words, 488-489
- Naval.
  - See also* Military and government applications
  - Air Technical Training Center (Orlando), 363-364
  - Combat Team Tactical training, 366
  - Ocean Systems, 363
  - Personnel Research and Development Center, 364-365

- Resource Management task, 376
  - Underwater Systems Center, 363
  - Navier-Stokes equation, 89
  - Neural nets. *See* Artificial neural networks
  - Neural transduction, 97
  - New Mexico State University, 241
  - Nippon Electric Corporation (NEC), 9, 10, 42, 82, 176, 383, 506, 507-511
  - Nippon Telephone and Telegraph (NTT) analysis-synthesis systems, 119
  - ANSER (Automatic Answer Network System for Electrical Requests), 283, 291, 292, 296-297, 398-399, 407-409, 410, 417
  - concatenative synthesis, 126
  - HMM applications, 176
  - systematic optimization techniques, 115
  - telephone speech database, 407
  - Noise
    - additive, 459
    - and algorithm robustness, 413
    - excitation, 122
    - immunity, 305
    - Lombard effect, 415, 460
    - reduction technology, 331-333, 414-415
    - sources, 122
    - and speaker variation, 415-416
    - and speech recognition, 288, 305, 379, 388, 414-415, 469, 473-474
    - white, 122
  - Northern Telecom, 278, 291, 295, 299
  - Numbers, pronunciation of, 143, 288
  - NYNEX, 282, 283, 291, 292, 300, 301-302, 407, 409, 436
  - O**
    - Occam parallel programming language, 396
    - Octel, 281
  - Official Airline Guide database, 46, 219
  - Olive, Joseph, 107
  - Operating systems
    - pen, 402, 511-512
    - speech, 417
  - Optical character recognition technology, 43, 349
  - Oregon Graduate Institute, 407
- P**
- Packet data network (XUNET), 99
  - Paget, Richard, 15-16
  - Palantype keyboard, 335
  - Parallel processing, 89, 383, 400
  - Parsing/parsers
    - ambiguous, 147-148
    - clause-level, 144, 145
    - crossing brackets, 491
    - natural language, 59, 247, 483, 489-495
    - phrase-level, 144-145, 146
    - probabilistic, 56
    - and prosodic marking, 56, 144, 146-147
    - in speech synthesis, 137, 139, 144-145
    - stochastic, 489-495
    - of unrestricted text, 144
    - word lattice, 265
  - Pause insertion strategies, 129
  - Performance structures, 146
  - Personal Communication Devices, 306
  - Personal Communication Networks, 306
  - Personal Communication Services, 306
  - Personal computers
    - hand-held, 355
    - portable, 64-65
    - sound boards, 350, 353, 397
    - speech interfaces for, 511
    - speech processing technology, 108, 374, 401-403, 509-510

- Phoneme**  
 conversion to acoustic events, 429  
 intelligibility, 411  
 recognition systems, 182  
 sequences, 138, 175, 461-462
- Phonetics**  
 acoustic, 85, 95  
 hidden Markov models, 166, 173-175, 178-179, 182, 188  
 segmental models, 125, 173-174, 190-191, 193  
 and speech recognition, 167, 169-170, 188  
 in training speech, 30, 178-179, 182-183  
 text-to-speech synthesis, 85, 125, 174  
 typewriter, 511
- Pierce, John, 283
- Pitch-synchronous overlap-add approach (PSOLA)**, 114, 119-120, 128-129
- Pitch-synchronous analysis**, 127
- Pragmatic structure**, 144, 246, 150, 246, 250
- Pronunciation**  
 abbreviations and symbols, 142-143  
 computational, 139  
 numbers and currency, 143, 288  
 part of speech and, 144  
 speech recognition, 44  
 surnames/proper names, 140-141, 288  
 symbols, 142-143
- Proper names**, 92, 288, 458, 484
- Prosodic phenomena**  
 algorithm, 146, 147, 151  
 articulation as a basis for, 152-153  
 and conversational dynamics, 431  
 defined, 144, 145-146  
 discourse-level effects, 149-151  
 marking, 144, 145-149
- modeling**, 117  
 multiword compounds, 147  
 in natural language processing, 268-269  
 parsing and, 56, 144, 146-147  
 pauses, 431  
 phrasing, 146-147, 151  
 PSOLA technique for modifying, 128-129  
 in speech synthesis, 88, 117, 119, 124-125, 128-129, 145-149, 288-289  
 and speech quality, 88, 118, 288-289
- Psychoacoustic behavior**, 78, 91, 94
- Pulse Code Modulation, adaptive differential (ADPCM)**, 24, 82-83, 101
- Q**
- Quasi-frequency analysis**, 177
- Query language, artificial**, 57
- R**
- Rabiner, Lawrence, 111, 113
- Recursive transition networks (RTNs)**, 222
- Repeaters, electromechanical**, 81
- Resonators**, 78, 80
- Research methodology, spoken language vs. types language**, 47-48
- Robust processing techniques**, 259-260, 263
- Robustness**  
 algorithms, 391, 392, 405, 412-416  
 ATIS system, 262  
 case frames and, 258  
 classification of factors in, 412-413  
 dialogue systems, 66  
 environmental variation in speech input and, 412-414  
 lexical stress system, 142

- linear predictive coding, 97
- multimodal systems, 64
- natural language systems, 56, 59, 262
- noise considerations, 413
- research, 417-418
- speaker variation and, 415-416
- speech analysis, 97
- speech recognition systems, 29-30, 44, 184, 261-262, 459-460
- speech synthesis, 139
- speech variation and, 413
- spoken-language understanding systems, 66, 258-259
- templates and, 258-259
- user interfaces, 56
- word error rates, 182-183, 184, 185-186
- Royal Institute of Technology (KTH), 122, 123, 124, 125, 129
- Rutgers University, CAIP Center, 98, 99
  
- S**
- Sampled-data theory, 78, 81
- Security applications
  - speaker verification, 9, 30, 86, 300, 305
  - low bit-rate coding for transmission, 7
- Semantics
  - ambiguity, 380
  - compositional, 486
  - First-Order Predicate Logic, 245-246
  - lexical, 486, 495-500
  - natural language, 245-246, 247, 250
  - pragmatics and, 144
  - propositional logic, 245
  - and speech recognition, 305-306
  - and spoken-language understanding, 220-221
- Sensimetrics Corporation, 123
- Siemens A. G., 9, 42, 83
- Signal modeling techniques, 19, 101
- Signal processing
  - digital, 19, 97
  
- enhancement, 102
- research, 21
- Sinusoidal models, 24
- Software technologies, 391
- Sound
  - generation, 118, 119, 124
  - source model, 462
- Sound Pattern of English*, 126
- Sound/speech spectrograph, 319, 325, 349
- Source-filter decomposition, 128
- Speak 'N Spell, 110
- Speaker
  - adaptation, 459, 460
  - atypical, 187-188
  - dependence, 36
  - recognition/identification, 9, 30, 85, 348
  - style shifting, 460, 461
  - variation, 415-416
  - verification, 9, 30, 86, 300, 305
- Speaking characteristics and styles, 128-129, 378-379
- Spectrum analysis, 19
- Speech
  - behaviors, conversational, 430-432
  - casual informal conversational, 82
  - compression, 23, 83, 474
  - connected, 97
  - continuous, 36, 78, 95, 323, 427-428, 430-431
  - constraints on, 77, 268-269
  - databases, 405, 407-409, 468
  - dialect, 409
  - digitized, 38, 45, 189, 428
  - dysarthric speech, 337
  - gender differences, 129
  - information processing technologies, 453
  - interactive, 36
  - intonation, 45, 127, 129, 432
  - knowledge about, 117
  - machine-generated, 335

- noninteractive, 48
- pause insertion strategies, 129
- perception models, 26
- preprocessor, 403
- production, 21-22, 26, 77, 87-90, 137-138
- prolongation of sounds, 322
- psychological and physiological research, 462
- self-correction, 256, 432
- signal processing systems, 19
- slips of the tongue, 257
- spontaneous, 58-59, 185, 255-260, 303, 460, 461, 469-471
- standard model of, 267
- synthetic, 428-428;
  - see also* Speech synthesis;
  - Speech synthesizers
- toll quality, 23, 24
- training, 322, 325
- type, 36
- ungrammatical, 257
- units of, 168-170, 462-463
- variability, 28, 176, 378, 413, 459-460, 480
- waveforms, 24, 136, 137
- Speech analysis
  - acoustic modeling, 26
  - analysis-by-synthesis method, 26-27
  - auditory modeling, 26
  - defined, 22
  - dimensions, 36-38
  - importance, 21
  - interactivity, 36
  - language modeling, 26
  - linear predictive coding, 24
  - robustness, 97
  - speech continuity, 36
  - speech type, 36
  - vocabulary and grammar, 28, 36-38
  - vocal tract representation in, 90, 91
- Speech coding, 26
  - applications, 82-83
  - articulatory-model-based, 125
- audio perception factors in, 84, 85
- in cochlear implants, 331
- concatenation using speech waveforms, 117
- bit rates and, 23, 24, 81, 83-84
- digital, 25, 82-83, 85
- and masking, 84, 93
- predictive, 117
- psychoacoustic factors in, 101
- research challenges in, 76
- rule-based diphone system, 118
- stereo coding, 84-85
- technology status, 82-85, 281
- terminal analog, 118
- wideband audio signals, 84
- Speech processing
  - algorithms, 21, 393
  - articulatory and perceptual constraints in, 461-463
  - digital, 22-23, 76
  - equipment and systems, 19-20, 81, 396-399
  - evaluation methods, 463-464
  - in hearing aids, 317
  - and natural language processing, 460-461
  - obstacles to, 373
  - research challenges, 76-77
  - psychoacoustic behavior and, 94
  - for sightless people, 333-335
  - and speech technology development, 76, 78
- Speech recognition
  - accuracy, 28, 37, 41, 46-47, 86, 159, 181-189, 377, 378, 470, 473
  - acoustic modeling, 64, 182-183
  - adverse conditions, 459-460
  - algorithms, 28, 409-411, 412, 417-418, 469
  - alternative models, 189-193
  - analysis-by-synthesis, 30
  - applications, 28-29, 30-32, 81, 275-282, 283-284, 318, 377-379, 451, 457, 458, 471, 508-510

- articulation and, 152-153  
assessment techniques, 410-411, 463-464  
"barge in" (interruption of conversation)  
and, 277, 287, 292, 295, 298-299,  
388, 404  
common speech corpora, 181-182  
complexity, 17  
connected digit corpus, 184-185  
continuous speech, 78, 165-194, 323,  
471, 506  
defined, 7, 239, 348  
decision criteria, 305  
decoding, 209-214  
dialogue grammar approach models, 63  
dimensions of task difficulty, 376,  
377-379  
domain independent (DI), 187  
dynamic grammar networks, 265-266  
dynamic programming matching, 509  
environmental factors, 413-414  
error correction, 64, 261-262, 388  
feature extraction, 177-178, 180  
Flexible Vocabulary Recognition, 295  
future, 307-309, 456-459  
generalization, 479  
Hidden Markov models and, 28, 30, 85,  
170-175, 177-178, 199, 200-208,  
377, 397, 478  
historical overview, 175-176  
improvements in performance, 181-184,  
388  
interactivity, 36  
language modeling, 29, 81-82,  
90-91, 168-169, 183, 263  
large-vocabulary systems, 183, 193,  
277, 292, 506  
lip reading, 64  
linguistic rules, 82  
market for technology, 350-351, 416-417  
microphones and, 305, 414  
most likely path, 208-209  
most likely word sequence, 209-214  
N-best filtering or rescoring, 267  
natural language and, 17, 262-267, 388  
naturalness, 45, 153  
neural networks, 191-193  
new words, 188-189  
noise immunity and channel equaliza-  
tion, 288, 305, 379, 388, 414-415,  
469, 473  
normalization of speakers in, 30,  
456-457, 459, 460  
pattern matching, 474, 478-479  
perplexity of language model and, 37,  
180, 185, 229, 378, 463  
phonetics and, 167, 169-170, 188, 410  
processes, 167-168, 180-181, 199,  
451, 453-454, 473-474  
pronunciation and, 44  
prototype systems, 34  
real-time, 189  
rejection of irrelevant input, 287, 388  
and repetitive stress injuries, 43  
research challenges, 29-30, 44, 76, 108,  
183-184, 304-306, 417-418  
robustness, 29-30, 44, 184, 261-262,  
459-460, 473, 474  
sample performance figures, 184-185  
search algorithms, 180-181, 248, 264-265  
segmental models, 190-191, 473-474  
sheep and goats phenomenon, 456  
speaker-adaptive, 36, 187-188, 288,  
388, 479  
speaking characteristics and styles and,  
128, 377, 378-379, 415-416, 460

- speaker-dependent, 28, 36, 54, 186-187, 292, 509-510  
 speaker expertise and, 378  
 speaker-independent, 28, 36, 37, 46, 184, 186-187, 188, 362-363, 378, 397, 425, 433-434, 506, 507  
 spontaneous speech and, 58-59, 185, 460, 461, 469, 471  
 SR-1000 system, 507  
 SR-3200 system, 507  
 subword units, 287-288, 299, 388  
 successful systems, 239  
 system structure, 27-28, 398, 401, 402  
 talker verification, 86  
 task completion rate, 410  
 technology status, 8-9, 18, 81, 85-86, 112-113, 159-164, 165-166, 181-189, 286-288, 428, 468  
 templates, 258-259, 425  
 terminal-type, 508-510  
 training data, 178-180, 185-186, 457, 459, 473, 478-479  
 transputer-based, 397  
 trials, 417  
 units of speech and, 168-170  
 user tolerance of errors and, 379  
 vocabulary and grammar and, 36-37, 41-42, 81, 85-86, 185-186, 265-266, 277, 378, 457  
 Wizard of Oz assessment technique, 410-411, 439  
 word lattice parsing, 265  
 wordspotting, 286-287, 292, 295, 298-299, 305, 387, 388, 397, 404  
**Speech research**  
 computational models of language, 90-91  
 critical directions in, 87-101  
 historical background, 78-82  
 language modeling, 26  
 physics of speech generation, 87-90  
 unification of coding, synthesis, and recognition, 94-95, 97  
 Speech synthesis.  
 See also Text-to-speech synthesis  
 acoustic models, 85, 95, 117, 122, 476  
 analysis-synthesis systems, 117, 118, 119, 125  
 applications, 30-32, 108, 109, 110, 278, 381-382  
 articulatory models, 88, 117, 118, 120, 124-125, 152-153, 476, 480  
 assessment of, 411-412  
 automatic learning, 127  
 concatenative, 110, 114, 117, 118-119, 126, 168, 406  
 concept-to-speech systems, 38-39  
 content, 45  
 control, 124, 118, 125-127  
 corpus-based optimization, 113  
 defined, 22, 109, 110, 116, 348  
 digitized speech, 22-23, 25, 38  
 dimensions of task difficulty, 381-382  
 discourse-level effects, 149-151  
 error rates, 112  
 evaluation of, 130  
 expectations of listeners, 382  
 flexibility needs, 117-118  
 fluid dynamics in, 89-90  
 formant-based terminal analog, 117, 118, 122-123, 125  
 forms, 38-39  
 frequency domain approach, 119  
 future of, 152-153, 455-456  
 higher-level parameters, 123-124  
 history of development, 111-115  
 individual voices, speaking styles, and accents and, 117-118  
 input, 109  
 intelligibility, 44-45, 129, 130, 149, 382, 429  
 large-vocabulary systems, 101-102, 351

- letter-to-sound rules, 140-141  
linguistic aspects of, 135-153  
market for, 351  
microelectronics revolution and, 108  
models, 109, 116-130  
morphophonemics and lexical stress, 110, 111, 112, 113, 137, 141-142  
multilingual, 42, 101, 117, 129-130, 151-152  
natural speech coding and, 117, 128  
naturalness, 129, 149, 381, 429, 456  
noise sources, 122  
and objective distortion metrics, 114-115  
obstacles to, 117  
orthographic conventions, 142-143  
output, 118  
parsing, 137, 139, 144-145  
part-of-speech assignment, 143  
phonetic HMM functions and, 174, 429  
predictive coding, 117  
process, 167-168, 135, 428-429, 453, 454, 479  
prosody, 88, 117, 119, 124-125, 128-129, 145-149, 288-289  
PSOLA (pitch-synchronous overlap-add approach), 114, 119-120, 128-129  
quantity of text and, 381  
real-time, 108  
research, 25-26, 29-30, 44-45, 76, 108, 113-114, 128  
rule-based, 111, 118, 125, 126-127, 140-145, 429  
segmental, 113-114, 115, 125, 145, 479-480  
sentence length and grammatical complexity, 382  
sound generation, 118  
source/system models, 22, 118, 120-121  
speech quality, 130  
structures and processes, 109-110
- systematic optimization methods, 114  
techniques, 118  
text analysis, 110, 112, 113  
technology status, 18, 29, 81, 85-86, 107-115, 411-412, 468  
testing, 114-115  
time functions, 111, 113, 118, 119, 476-478  
variability of text and, 381-382  
vocabulary, 119  
vocal tract model, 95, 118, 122, 125  
waveform concatenation (simple), 118-119, 383, 476  
word-level analysis, 138-139
- Speech synthesizers  
acoustic terminal analog, 117  
cartridge-type, 510  
cascade, 122-123  
future, 455-456  
large-vocabulary, 349  
neural network controller, 124  
OVE, 123  
parallel, 123, 125  
terminal analog, 510  
voice quality, 456
- Speech technology, *See Deployment of applications*  
capabilities and limitations, 427-430  
challenges in, 284, 471-475  
commercial developments, 352-354  
foundations, 77-78  
growth of, 2  
information processing, 453  
market, 350-352, 416-418  
projections, 101-102, 355-356  
readiness evaluation, 440  
research on, 65-67, 417-418  
service trials, 417  
status, 82-87  
trends, 117

- voice input, 427-428
  - voice output, 428-429
  - Speech Technology Laboratory, 123
  - Speech transmission, low-bit-rate, 23, 24, 29, 77, 81, 83-84, 97, 474
  - Speech understanding, 17, 34, 37-38, 307, 379
  - Spoken language systems (SLS) ARPA, 218-220
    - comparison of modalities, 46-58
    - constraints on, 227-230
    - defined, 38, 241
    - dialogue, 47, 60, 61-63, 66, 229
    - discourse in, 227-230
    - efficiency of language-based modalities, 48-51
    - error metrics, 224-225, 259
    - error recovery, 439
    - evaluation of, 230-233, 251
    - human factors obstacles to, 58-63
    - interaction, 51-57, 60, 61-63
    - interfacing speech and language, 221-224
    - linguistic analysis, 59-60, 259
    - mixed initiative, 228-229
    - N-best interface, 217, 221, 233
    - natural language, 51-57, 59-61
    - order in problem solving, 229
    - prototypes, 46-47, 438
    - reference, 227-228
    - robustness, 66, 259
    - simulation methods, 66
    - speaker-independent, 65
    - spontaneous speech and, 58-59, 234, 255-260, 427-428
    - SUNDIAL, 228-229
    - research methodology, 47-48
    - technology development, 81
    - training, 60, 260
    - typed language contrasted with, 47-51, 60
    - user adaptation to, 60
  - Spoken language translation
    - current capabilities, 9-10, 42
    - defined, 9-10
    - directory assistance, 295-296
  - laboratory systems, 9-10
  - projections, 102
  - VEST (Voice English-Spanish Translator), 10, 42
  - voice output, 29
  - Spoken language understanding, 47
    - approaches to, 220-221
    - defined, 255
    - error repair, 260
    - limits on, 379
    - process, 452, 453
    - progress in, 224-226
    - spontaneous speech and, 258-260
  - Sprint, 300
  - SQL, 57
  - SRI International, 52, 176, 213
    - ATIS, 46, 261
    - Gemini system, 259, 260
    - Template Matcher, 258, 259
  - Stenograph, 322, 335
  - Stereo coding, 84-85
  - StockTalk, 383-386, 437, 438, 439
  - Stored voice, 110
  - Subband coders, 24, 83, 101
  - SUNDIAL spoken language systems, 229
  - Surnames, pronunciation of, 140-141, 288
  - Symbols, pronunciation of, 142-143
  - Symbolic learning techniques, 501
  - Syntax, 137.
    - See also* Parsing
    - natural language processing system, 244-245, 247, 269
    - speech recognition systems, 305-306
    - and spoken language understanding, 220-221
  - Syntactico-semantic theory, 447
  - System technologies. *See* Hardware technology;
  - Workstations
- T**
- Tactile technology, 101, 324-328
  - Talker. *See* Speaker
  - Talking statues, 78, 79

- Technology transfer issues, 367-369  
Telecommunications.  
*See also* Telephony  
banking services, 292, 507  
Baudot code, 323  
conferencing, 101  
cost-reduction applications, 290-291  
digital speech coding, 82-83  
information access from remote databases, 42, 44, 278, 296-299, 348 , 349  
interfaces, 397  
market for speech technology, 290-304  
personal communication networks and services, 306  
predictions, 307-308  
revenue opportunities in, 291-293  
shaping user language, 60-61  
speaker verification, 305  
speech technology and, 7, 41-42, 285-286  
technical challenges, 304-306  
Telefonica, 9, 10, 298  
Telegraph, 80-81  
Telephony.  
*See also* Telecommunications  
Automated Alternate Billing Services, 292, 293, 431  
Automated Customer Name and Address, 302  
automatic interpreting, 513-514  
bandwidth conservation, 19  
banking by phone, 283, 291, 398-399, 407-408, 425  
cellular, 6, 7, 81, 83, 374, 383-385, 507-508  
deaf user aids, 43, 302-304  
digital channels, 101  
directory assistance, 41, 278, 282, 283, 291, 292, 295-296, 301-302, 355-356, 438, 458  
history, 81  
language translation, 10, 42, 77, 81, 82, 83, 108-109, 513-514  
operator services, 8-9, 277, 282, 284, 291, 292, 293-296, 351, 353-354, 374, 380, 383-385, 387  
simulated telephone lines, 278, 408-409  
speech databases, 407  
speech recognition technology, 428  
teleconferencing, 454-455  
telephone relay service, 302-304, 322  
text telephone, 322, 323  
voice-controlled automated attendant, 356  
voiced-based dialers, 40, 292, 299-300, 355, 374, 376, 383-386, 436 , 507-508  
voice-interactive phone service, 292, 300-301, 351  
Voice Recognition Call Processing (VRCP), 292, 293-295, 376, 383-385  
TELECOM, 510, 513  
Telephone answering machines, digital, 7-8  
Texas Instruments (TI), 110, 176, 184-185, 291, 300, 349, 377, 407  
Text analysis, 110, 112, 113  
Text-to-speech synthesis.  
*See also* Speech analysis  
acoustic phonetics and, 85  
address, date, and number processing, 288  
advances in, 288-289  
algorithms, 25  
applications, 43, 109, 280, 282, 302, 354, 451  
articulatory synthesis in, 124-125  
cartridge-type device, 510  
components of, 38  
constraints on speech production, 137-137  
development tools, 126-127  
discourse analysis in, 145

- error rate, 262  
 formant-based terminal analog, 122-123  
 future of, 152-153, 308  
 hardware requirement, 383  
 language modeling and, 26, 78, 90-91  
 linguistic analysis in, 382  
 multilingual, 42, 129, 397-398  
 naturalness, 381  
 output, 29  
 parsing, 144-145  
 part-of-speech assignment, 143  
 phonemic-based, 348  
 phonetic factors, 125  
 problems, 120, 303-304, 471  
 proper name pronunciation, 288  
 prosody, 288-289, 306  
 research challenges, 26, 304, 306, 324  
 rule system, 125  
 sound generation, 124  
 source models and, 120  
 speaker identity and normalization, 30  
 speaking characteristics and styles and, 128-129  
 structural framework, 136-137, 398  
 waveform approach, 24-25  
 word-level analysis, 138-139
- Text preprocessors, 381-382  
 Time Assignment Speech Interpolation, 81  
 Tools. *See* Computer-aided tools  
 Touch screens, 50  
 Touch-Tone keypad, 335  
 Trackballs, 52  
 Training  
     natural language interactive systems, 56, 57, 58  
     neural nets, 193  
     shaping user language, 60-61  
     speech, 322  
     tactical, combat team, 364-365  
 Training speech [learning]  
     automatic, 263-264  
     databases for, 387, 405, 407, 468, 472
- discriminative, 479  
 effects of, 185-186, 473  
 grammar, 179-180, 185-186  
 natural language processing, 56, 57, 58, 249, 250, 252, 263-264  
 phonetic HHMs and lexicon, 30, 178-179, 182-183  
 speech recognition, 178-180, 185-186, 457, 459, 473, 478-479  
 syntactico-semantic theory and, 447  
 Transatlantic radio telephone, 81  
 Transatlantic telegraph cables, 81  
 Transform coders, 24  
 Treebank Project, 241, 491, 495  
 Trigrams, 92, 183, 201-202, 209-210, 212, 213-214, 229  
 Triphones, 182  
 Turing's test, 35  
 Tuttle, Jerry O., 363
- U**
- United Kingdom, Defense Research Agency, 365  
 University of Indiana, 130  
 University of Pennsylvania, 181, 241, 252, 491, 495  
 US West, 300-301  
 Usability/usefulness.  
*See also* Applications of voice communications  
 determinants of, 31-32  
 issues, 18, 30-32  
 pronunciation and, 44  
 voice input, 39-44  
 voice output, 44-45
- User interfaces.  
*See also* Graphical user-interface  
 artificial query language, 57  
 capabilities and limitations, 51-52, 387, 427-430, 434  
 cost of interaction failures, 426-427  
 databases, 240, 252

- design strategies, 387, 423-424, 426, 433-440
  - dialogue flow, 435-436
  - direct manipulation, 51, 52-55, 57-58
  - error recovery, 438-440
  - evaluation of, 440
  - feedback and confirmation, 434, 437-438, 445
  - heirarchical, 454
  - information requirements of, 425-426
  - instructions, 438
  - keyboard dialogs, 49-50
  - metaphor, 54
  - multimodal systems, 32, 56, 63-65, 505, 508-510
  - N-best, 217, 221, 226, 233
  - natural language interaction, 55-57
  - personal computer, 511-512
  - prompts, 435-436, 471
  - research directions, 56, 511-512
  - revisions suggested, 435
  - robustness, 56
  - smart, 512-513
  - system capabilities, 429-430
  - task modalities, 426
  - task requirement considerations, 424-427
  - telecommunications, 397
  - training issues, 58
  - user expectations and expertise and, 430-432
  - voice-actuated, 360
  - voice input, 427-428
  - Users
    - conversational speech behaviors, 430-432
    - expectations and expertise, 430-432
    - language modeling by, 60
    - novices vs. experts, 432
    - satisfaction, 429-430
    - tolerance of speech recognition errors, 379
  - USS Ranger, 363
- V**
- Vector quantization, 28
  - Verbal repair, 269
  - Videophones, 5-6
  - Virtual reality technology, 454-455
  - Visual sensory aids, 319-324
  - Vocabulary
    - algorithms, 307
    - confusability, 378
    - conversational, 101-102
    - Flexible Vocabulary Recognition, 295
    - large, 101-102, 183, 193, 277, 292, 307, 349, 351, 506
    - and natural language understanding, 37-38
    - operator services, 277
    - speech analysis and, 28, 36-38
    - speech recognition and, 36-37, 41-42, 81, 85-86, 183, 185-186, 193, 265-266, 277, 292, 378, 457, 506
    - speech synthesis, 101-102, 119, 349, 351
    - user-specific dictionaries, 335-336
    - wordspotting techniques, 292, 305
  - Vocal tract modeling, 95, 118, 122, 124, 125
  - Vocoder, 48, 81, 83, 119, 325
  - Voice
    - control, assistive, 278-279, 313, 337, 360, 452
    - conversion system, 128-129
    - dialog applications, 375-377
    - fundamental frequency, tactile display, 326-327
    - input, 39-44, 50, 427-428
    - mail, 7, 81, 83, 101, 110
    - messaging systems, 281
    - mimic, 94-95
    - output, 44-45, 428-429
    - response, 25
    - task-specific control, 452
    - typewriters, 97, 376, 380, 451

- Voice coding  
 algorithm standardization, 7  
 current capabilities, 7-8  
 defined, 7-8  
 research challenges, 306  
 security applications, 7  
 source models, 120-122  
 storage applications, 7-8
- Voice communication, human-machine  
 advantages, 16, 48-51  
 art of, 387-388  
 current capabilities, 469  
 degree-of-difficulty considerations, 375-386  
 expectations for, 505-506  
 implementation issues, 18  
 natural language interaction, man-farm  
 animal analogy, 16  
 process, 374  
 research and development issues, 511-513  
 research methodology, 47-48  
 role of, 34-67;  
*see also* Applications  
 scientific bases, 15-33  
 scientific research on, 65-67  
 simulations, 47-48, 50, 51  
 successful, 423  
 system elements, 17-18  
 and task efficiency, 48-49  
 transcript, 433-434  
 voice control, 337  
 VLSI technology and, 510-511
- Voice processing  
 network-based, 292  
 market share, 281  
 research, 6  
 technology elements, 6-7  
 technology status, 467-468  
 telecommunications industry vision, 285-286
- Voice synthesis  
 current capabilities, 8  
 defined, 8  
 output, 23, 17-18, 29  
 text-to-speech, 99
- von Kempelen's talking machine, 78, 80
- Vowel  
 clusters, 140  
 digraphs, 140  
 reduction, 129
- VLSI technology, 468, 510-511
- W**
- Wave propagation, 26  
 Waveform coding techniques.  
*See also* Speech coding  
 adaptive differential PCM (ADPCM), 24  
 speech synthesis, 118, 119, 136, 137, 381, 474  
 Wavelets, 21  
 Wideband audio signals, 84  
 Windows, 52, 350, 353  
 Wizard of Oz (WOZ) assessment technique, 410-411, 439  
 Word-level analysis, 138-139  
 Word models, 179, 207  
 Word processors, speech only, 50  
 Word recognition systems, 182, 188  
 Workstations  
 Hewlett-Packard 735  
 RISC chips in, 393  
 Silicon Graphics Indigo R3000, 189  
 speech input/output operating systems, 401-403  
 speech processing board, 397  
 Sun SparcStation 2, 189  
 Wheatstone, Charles, 80
- X**
- Xerox, 52
- Z**
- Zipf's law, 489