Automatic Speech Recognition in Severe Environments (1984)

Pages
91

Size
8.5 x 10

ISBN
0309323975

Committee on Computerized Speech Recognition Technologies; Commission on Engineering and Technical Systems; National Research Council

84-0122

1100231 PB85-121697/XAB

Automatic Speech Recognition in Severe Environments
(Final rept)
National Research Council, Washington, DC.
Corp. Source Codes: 019026000
Sponsor: Office of Naval Research, Arlington, VA.
1984 92p
Languages: English
NTIS Prices: PC A05/MF A01 Journal Announcement: GRAI8505
Country of Publication: United States
Contract No.: N00014-830-G-0072

For more than a decade the United States government, private corporations, and universities have been engaged in research on human-machine interaction by voice. The Department of Defense, in particular, has recognized the potential for improving the safety and effectiveness of its forces by making electronic and electromechanical devices directly responsive to the human voice and able to respond by voice. The benefits of this capability would be especially noteworthy in situations where the individual is engaged in such hands/eyes-busy tasks as flying an airplane or operating a tank. Voice control of navigational displays, information files, and weapons systems could relieve the information loads on visual and manual channels.

Descriptors: *Speech recognition; Technology; Man machine systems; Automation

Identifiers: Computer applications; Speech synthesis; Interactive systems; NTISNASNRC; NTISDODN

Section Headings: 17B (Navigation, Communications Detection, and Countermeasures--Communications); 9B (Electronics and Electrical Engineering--Computers); 5H (Behavioral and Social Sciences--Man-machine Relations); 45F (Communication--Verbal); 95D (Biomedical Technology and Human Factors Engineering--Human Factors Engineering)

Automatic Speech Recognition In Severe Environments

A Report Prepared by the

Committee on Computerized Speech Recognition Technologies
Commission on Engineering and Technical Systems
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1984

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Research Council was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the federal government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which establishes the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine. The National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively, under the charter of the National Academy of Sciences.

This study was supported by grant N00014-830G-0072 between the National Academy of Sciences and the Office of Naval Research.

Copies available from:

[redacted]

National Research Council
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Order from
National Technical
Information Service,
Springfield, Va.
22161
Order No. PB85-121697

Printed in the United States of America

COMMITTEE ON COMPUTERIZED SPEECH RECOGNITION TECHNOLOGIES

James L. Flanagan (Chairman), Head, Acoustics Research Department, AT&T Bell Laboratories

N. Rex Dixon, Associate Editor, IBM Journal of Research and Development, IBM Corporation

George R. Doddington, Manager, Speech Systems Research Branch, Computer Science Laboratory, Texas Instruments

John I. Makhoul, Principal Scientist and Manager, Speech Signal Processing Department, Bolt Beranek & Newman Inc.

Michael E. McCauley, Vice President, Monterey Technologies, Inc.

Ellen F. Roland, Co-owner, Rolands and Associates Corporation

John C. Ruth, Vice President - Marketing and New Business Development, McDonnell Douglas Electronics Company

Carol A. Simpson, Co-owner, Psycho-Linguistic Research Associates

Beverly H. Williges, Senior Research Associate, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University

William A. Woods, Chief Scientist, Applied Expert Systems, Inc.

Victor W. Zue, Assistant Professor, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

---

COMMISSION ON ENGINEERING AND TECHNICAL SYSTEMS LIAISON MEMBERS

Erich Bloch, IBM Vice President - Technical Personnel Development

C. Kumar N. Patel, Executive Director - Research Physics Division, AT&T Bell Laboratories

---

STAFF

Dennis F. Miller, Study Director
Howard E. Clark, Staff Officer
Helen D. Johnson, Staff Associate
June F. Richardson, Administrative Secretary

PREFACE

For more than a decade the United States government, private corporations, and universities have been engaged in research on human-machine interaction by voice. The Department of Defense (DOD), in particular, has recognized the potential for improving the safety and effectiveness of its forces by making electronic and electromechanical devices directly responsive to the human voice and able to respond by voice. The benefits of this capability would be especially noteworthy in situations where the individual is engaged in such hands/eyes-busy tasks as flying an airplane or operating a tank.

In order to provide a forum for the exchange of ideas and information, and, additionally, to coordinate the activities of several federal agencies and research groups dealing with speech input/output systems, the Department of Defense established a Voice SubTechnical Advisory Group. This group functions within the Defense Department's Human Factors Technical Advisory Group (TAG). Thus, members of the Voice SubTAG are some of the government researchers and program managers who must respond to the increasing pressures to demonstrate and deliver computer-based, voice-interactive systems capable of reliable operation in severe and moderate military environments.

In view of these responsibilities, the Voice SubTAG asked the National Research Council to conduct an independent evaluation of the status and outlook for voice-interaction in moderate and severe environments. Recognizing also that the technology of speech synthesis is further advanced, the request to the National Research Council centered specifically upon the technology of speech recognition and upon the human factors issues associated with the application of speech recognition to government and military tasks. Speech synthesis is addressed, in this report, only in the context of the human-factors issues associated with the design of voice-interactive systems.

The study plan contained six objectives:

1. Identify major applications of automatic speech recognition that have existing or potential utility in both stressful and moderate environments characterized by both moderate and severe levels of psychological and physiological stress.

2. Review and summarize the properties of human speech under stressful conditions and in the presence of different types of background noise characteristic of the envisaged settings.
3. Review and summarize the state of the art and expected progress in speech recognition algorithms and systems, including procedures for enhancing information flow and rejecting noise.
4. Recommend the nature of performance guidelines and standards needed for further development and application.
5. Outline performance tradeoffs and assessment procedures, including consideration of the "systems cost" of the additional information channel offered by computerized speech recognition and synthesis.
6. Identify required research in fundamental areas of speech technology and its applications.

Responding to this request, the National Research Council's Commission on Engineering and Technical Systems assembled the Committee on Computerized Speech Recognition Technologies. The committee was comprised of eleven specialists from fields pertinent to speech recognition. These fields included speech communication, computer science, electrical engineering, human factors, acoustics, psycholinguistics, avionic systems, natural language processing, and artificial intelligence. The committee was drawn from industry and academia. The committee's task was to make projections and recommendations for future development of speech recognition technology, particularly for the military sector.

The committee met five times, for approximately two days each time, from July 1983 to March 1984. The initial meetings were devoted to information gathering. The later meetings were devoted to analysis of the information, specifically as it related to government applications, and to formulating this written report. Additionally, the results of the committee's study were presented to the committee's sponsors in a briefing on August 2, 1984.

In the course of its study, the committee:

- was briefed by members of the Voice SubTAG on the content of their programs and their expectations regarding the committee's role;
- heard presentations by invited consultants and experts on the state of the art of laboratory research and commercial systems;
- discussed the goals, procedures, and accomplishments of DOD's Advanced Research Projects Agency's Speech Understanding Research Project in the 1970s; and
- examined possible applications of speech recognition systems in the military sector; for example, in fighter aircraft and shipboard command/control centers.

These investigations involved briefings by equipment vendors and systems designers, interviews with test pilots, and field trips to development laboratories, test facilities, and operations environments.

This report should be of use to DOD and civilian policy makers who determine the capabilities that will be required in advanced combat systems. For these readers, the report gives the committee's best judgment regarding the probability of meeting various requirements in the future. Laboratory researchers and research managers may also look to the report for guidance in planning future activities.

The committee's ability to meet its obligations was heavily dependent upon assistance provided by a number of interested and accommodating parties. We are deeply appreciative of their kindness and efforts.

In particular, we thank Colonel Harry Heimple, director of the Advanced Fighter Technology Integrator (AFTI) F/16 Program, and all other members of his team, including test pilots, equipment vendors, General Dynamics staff, and National Aeronautics and Space Administration (NASA) staff for hosting the site visit to Edwards Air Force Base, California. We thank Captain James R. Williams, commanding officer of the Naval Surface Weapons Center and the members of his staff for briefing the committee and providing a tour of Systems Control Laboratory at Dahlgren, Virginia. We thank Leon Lerman of Lockheed Missiles and Space Company, Sunnyvale, California, for providing his first-person account of lessons learned in applying speech recognition systems to everyday industrial operations.

Additionally, we thank the commanding officer of the cruiser U.S.S. Long Beach, Captain Frederick Triggs, USN, his executive officer, Captain Clyde J. Vanarsdall, USN, and Master Chief Jess Mahon and other crew members for a memorable and productive tour of the ship's combat information center.

The committee values and appreciates the interest and counsel that it has received from its two liaison members from the Commission on Engineering and Technical Systems--C. Kumar N. Patel and Erich Bloch.

Finally, the committee wishes to recognize the efforts of Dennis F. Miller, study director, in developing and organizing this study and, for his guidance throughout this performance period. Further, the committee wishes to express its appreciation to Howard Clark, staff officer, for his assistance. Also, the committee wants to thank Helen Johnson, Patricia Wood, June Richardson, and Julia Torrence for their cheerful help and generous support.

CONTENTS

EXECUTIVE SUMMARY

HUMAN-MACHINE INTERACTION BY VOICE

Speech is a natural and convenient means for human communication. Information exchange between humans and complex machines could be facilitated if machines could respond appropriately to spoken commands through action, information processing, or machine-generated voice. In particular, interaction by voice could alleviate the information load on the human in many "hands-busy/eyes-busy" situations.

Heavy information loads place severe demands on personnel such as pilots of single-seat, high-performance aircraft, tank crews in the field, and combat information center staff. Voice control of navigational displays, information files, and weapons systems could relieve the information loads on visual and manual channels.

The techniques of automatic speech recognition allow the machine to respond to spoken commands. The techniques of speech synthesis permit the machine to generate spoken responses. Over the past decade notable advances have been made in the development of automatic speech recognition and speech synthesis devices. These advances have been fueled by explosive progress in integrated circuit technology. Rudimentary systems are now being commercialized and more sophisticated techniques are under laboratory study.

But present speech recognition systems are rigidly limited in capability and require significant complexity for reliable performance. And, most speech systems have been used by forgiving users in benign environments. In contrast, military applications often involve harsh environmental conditions and demanding tasks where humans may be exposed to high ambient acoustic noise, encumbered by equipment (such as an oxygen or gas mask), and subjected to significant physiological and psychological stress during combat conditions.

Applications of speech synthesis are presently more extensive than those of recognition because message-generation techniques are somewhat better understood. Speech synthesis has been, and will

continue to be, a vital adjunct to recognition for voice-interactive systems, and its development must go apace. But the challenges in automatic recognition are greater, and this report focuses primarily on the issues of speech recognition.

ORIGIN OF THE PRESENT STUDY

Several military and government organizations requested that the National Research Council organize a study of the status and outlook for automatic speech recognition. The study was organized to consider the level of performance that present speech technology can support under hostile conditions, and what research and development might be undertaken to create a speech technology that could serve usefully in severe environments. The study aims to complement and expand the scope of previous efforts, such as the assessment conducted by Beek, Neuburg, and Hodge in 1977.

COMMITTEE ON COMPUTERIZED SPEECH RECOGNITION TECHNOLOGIES

In response to this request, the National Research Council convened a committee of eleven specialists to make projections and recommendations for future development of speech recognition technology, particularly for the military sector. The committee met five times, for approximately two days each time, over the interval July 1983 to March 1984. The initial meetings were devoted to information gathering--from presentations by the sponsors, current contractors, invited consultants and experts, and from several site visits to military installations. The later meetings were devoted to analysis of the information, specifically as it related to government applications, and to developing this written report. Additionally, in August 1984, the committee conducted an open briefing for its sponsors and other interested parties.

CONCLUSIONS

Based on its exposure to the issues and its familarity with the field, the committee concluded* that:

- The use of speech for communication between humans and machines has distinct potential for military and other government purposes.
- Current technology for automatic speech recognition is not sufficiently advanced to provide robust, reliable performance in hostile and high-stress environments.
- Current speech recognition technology is not sufficiently advanced to achieve high performance on continuous spoken input with large vocabularies and/or arbitrary talkers.
- Current technology is mature enough to support restricted applications in benign environments, with disciplined use under low-stress conditions. Success strongly depends upon the integration of speech recognition with improved automation techniques.
- No standardized techniques exist for evaluating and comparing the performance of speech recognizers.
- No established human-factors methodologies exist for analyzing and evaluating human-machine performance in integrated voice-interactive systems or for systematically quantifying the benefits of speech input as compared to related automation techniques.
- There is insufficient fundamental understanding of how human speech degrades under severe environmental and stress conditions and of how to design recognition algorithms for these conditions.
- Government-sponsored efforts are currently insufficient to sustain major advances in speech recognition technology.
- Laboratory studies of speech recognition algorithms will probably require sophisticated computational resources that are not widely available.
- Successful deployment of advanced speech recognition systems will be directly related to, and in part dependent upon, continued advances in integrated circuit technology and computer architecture.
- Speech synthesis is an important adjunct to automatic speech recognition for voice-interactive systems.
- No central focus exists in the U.S. government to manage research and development in speech recognition.

---

*Abstracted from detailed conclusions in Chapter 6.

RECOMMENDATIONS

The committee's conclusions lead to corollary recommendations. These recommendations* aim to achieve a speech recognition technology that can provide utility, accuracy, and reliability in severe as well as benign environments.

- A basic research program is needed to characterize speech and its variabilities, including the study of the acoustic properties of speech in various contexts and for different speakers.
- Because an automatic speech recognizer is limited by the information delivered to it, new methods for sound transduction (including microphone systems designed for severe environments) and for electronic signal enhancement should be sought and studied.
- Significant research efforts are required in the design of algorithms and systems for the recognition of continuous speech in complex application domains, for speaker-independent operation, and for robust performance under conditions of degraded input.
- Research is necessary to establish human-factors procedures for analyzing human-machine communication tasks, to quantify the benefits that speech input/output can contribute, and to develop systematic techniques for integrating speech functions into the systems design.
- Extensive hardware development and deployment, based on existing technology, is inappropriate. Exploratory hardware efforts, however, are vital for gaining practical knowledge about applications and for establishing the limitations of existing technology.
- Standardization should be established to quantify the performance of automatic speech recognizers and to permit comparisons among algorithm philosophies and environments. Common data bases and prescribed procedures for assessing performance should be made generally available.
- Sophisticated computational capabilities are required to support continued advances in speech recognition work. A program for advanced research in speech recognition should have appropriate interfaces with government-sponsored work on high-speed processors and strategic computing.
- A substantial, sustained, and coordinated program of research and development is required to realize the potential of speech recognition within the U.S. government. The program should be built around long-range goals, with the acquisition of fundamental knowledge as a central thrust. This objective is especially crucial to advancing continuous speech recognition and to achieving talker independence with large vocabularies. A focus of responsibility and accountability as well as a means for coordinating the program is necessary.

*Abstracted from detailed recommendations presented in Chapter 6.

INTRODUCTION

HUMAN-MACHINE INTERACTION BY VOICE

For some years, the potential of voice communication with computers has been recognized as a means for making machines easier for humans to use. In many cases it would be more convenient and faster to speak commands to a computer than to use a keyboard, typically in a specialized symbolic format. Similarly, it is frequently desirable to obtain information from a computer by spoken output, especially when visual channels are occupied or when visual displays are not available.

The capability for computer voice input is traditionally called automatic speech recognition (ASR). The capability for spoken output is called speech generation or speech synthesis.

Ideally, humans would like to be able to converse with computers as fluently as with one another. Current capabilities are far short of this and will remain so for many years. But there are many situations where voice-interaction of a restricted nature is not only possible, but highly useful. Research into these areas has been progressing, especially in the commercial sector where applications range from voice control of toys, television sets, and microwave ovens to computer terminals that permit manipulation of data files and information access by spoken commands.

In large measure, these initial applications are characterized by benign environmental conditions, fault-tolerant tasks, and forgiving users. This commercial sphere is consequently a good proving ground for fledging voice-interactive techniques.

In contrast, military and other government applications are frequently characterized by severe environmental conditions, such as high levels of acoustic noise, equipment encumbrances to normal speech, and physiological and psychological stresses associated with combat conditions. Further, the military sector is typified by life-critical tasks and demanding (though disciplined) users of speech and computer equipment.

Nevertheless, the attractions of voice input/output are just as strong, because of the heavy information-processing loads often imposed on humans in complex defense situations. A question, therefore, is whether, at this early point, the fledging techniques of voice-interaction can be used successfully by the U.S. military.

ISSUES IN THE TECHNOLOGY OF VOICE INPUT/OUTPUT

Speech Recognition

Automatic speech recognition has at least three dimensions: vocabulary size, fluency, and speaker dependence. Vocabulary size relates to the number of items that a machine can recognize. Fluency concerns the mode of speaking to a machine, ranging from single isolated words to continuous conversational speech. And, speaker dependence means that an individual must "train" the machine to respond to his or her own voice. By contrast, speaker independent systems can respond to a variety of voices.

For benign environments, the current state of speech recognition technology permits reliable recognition of individual words spoken distinctly and in isolation. Word recognition accuracy exceeding 95 percent is achievable. For a speaker-trained (speaker-dependent) system, vocabularies of a hundred words or more can be accommodated. For speaker-independent operation, comparable accuracy can be achieved for isolated word vocabularies of about a dozen words (Rabiner and Levinson, 1981).

The computational complexities for these two conditions are comparable. Both require about 1-10 million instructions per second (MIPS). Research is progressing on connected and continuous speech, but performance is not yet adequate to support broad applications.

Speech Synthesis

Speech synthesis also has at least three dimensions: quality of voice output, versatility of message generation, and complexity of the synthesis algorithm. The most direct and most limited method for computer voice output is to record the speech sound waveform digitally for a variety of human-spoken words and phrases, to store these elements in a computer memory, and to concatenate them into appropriate messages as needed. Such systems have been in use for some years for automatic voice announcements of a restricted variety. While the voice quality is high (i.e., comparable to human speech), so too is the expense of waveform storage. Also, the versatility with which vocabulary elements can be acceptably rearranged is low. Contextual variations in speech prosody (sound pitch, duration, intensity) cannot be made directly.

Economy of storage and control of prosody in assembling messages from human-recorded words and phrases can be achieved by encoding the speech at reduced bit rates. A popular method is linear-predictive coding (LPC) used for coding rates of 1.2k to 9.6k bits per second. Because the speech is represented in parametric form, the possibility exists for concatenating words and phrases smoothly, and in a variety of sequences. Control of prosody appropriate to context is likewise possible. The price paid for the increased economy and flexibility is a moderately synthetic-sounding quality. While versatility is increased, the technique is still limited to the word and phrase vocabulary selected for storage.

The ultimate in flexibility is synthetic speech that is unrestricted in vocabulary and context. To accomplish this, the machine must have internal knowledge of the distinctive speech sounds of the language (the phonemes), and it must have algorithms for assembling the phonemes and the prosody appropriate to the context of the desired message. With this capability, the machine can, in principle, synthesize unrestricted speech from a discrete phonetic description of the message.

Methods for accomplishing this capability differ in their internal representations of the distinctive speech sounds. One method stores the frequency values of the vocal resonances measured in human speech sounds. It interpolates between these values in producing sound sequences. Another method uses a library of stored sound elements corresponding to fractional parts of syllables. These elements are excerpted from human speech, parametrically encoded by LPC, and stored in the machine. Still another method uses no vestige of human speech, but computes, from a programmed model of the human vocal tract, the trajectories of the resonant frequencies corresponding to a specified sound sequence.

Additionally, if the machine is provided the computational capability for converting the written symbols of language into the phonetic equivalent, it may then convert unrestricted printed text or data into the spoken equivalent. The letter-to-sound conversion for this "text synthesis" can be accomplished by algorithm, or by a stored pronouncing dictionary, or, more typically, by a combination of these tools. The price paid for this great versatility is the processing power required, the storage necessary for dictionary and data, and the less-than-human quality of the synthesis.

Some commercial systems for text synthesis are now emerging, but the problem of achieving natural sounding synthesis from unrestricted text remains very much in the research stage.

System Design and Human Factors Integration

Voice-interactive systems, by definition, involve humans and human performance directly affects the utility of a particular speech feature. Human factors issues are therefore central to successful

deployment of voice-interactive systems. Heretofore, emphasis has been given largely to advancing the individual "component boxes" of speech recognition and synthesis. The time is now appropriate to advance the application of speech technology through a "total systems approach" that recognizes the human as a component of the integrated system.

Environmental Factors

To date, the most successful applications of speech technology have been with cooperative and forgiving users in benign environments. These conditions are typical of initial applications for commercial products.

Government and military applications are unique because of the range of environmental conditions under which high reliability and high performance must be achieved. Environmental factors can have at least two significant effects. First, the speech signal can be contaminated by acoustic noise and distorted by physical obstructions before it reaches the speech recognition device. Second, environmental stress and noise can elicit variable speech from a human talker. The same environmental factors, acting on the human listener, can interfere with accurate perception of speech that is synthesized by machines. These environmental factors have not figured prominently in speech research.

BASIC CONCEPTS IN SPEECH TECHNOLOGY

WHAT IS AUTOMATIC SPEECH RECOGNITION?

An automatic speech recognition (ASR) device accepts a speech signal as input and produces as output a string of symbols corresponding to a spoken message. Figure 1 shows the three components of a generic ASR process, namely, a human speaker, an ASR device, and an application host. The human speaker encodes his or her intentions via speech, the ASR device translates the spoken words into a format that can be interpreted by the host, and the host accepts the output and takes appropriate action.

When an ASR device is integrated into a total system to facilitate user/machine interaction, task-specific knowledge may be incorporated in the device to improve recognition performance. In such cases, the output symbol string from the device will represent the system's "understanding" of the input message. In commercially available systems to date, no speech understanding per se is attempted, and the output symbol string is simply a sequence of recognized words or utterances.

The ASR process operates in an environment that is acoustic and task oriented. Both the acoustic characteristics of the environment and the nature of the tasks being performed affect the system's performance by affecting each of its components (the human speaker, the device, and the application host). Proper application of automatic speech recognition should address each of these components and their interactions.

AUTOMATIC SPEECH RECOGNITION PARAMETERS

Table 1 lists some of the parameters that characterize the capabilities of ASR devices. The three types of parameters are: device specific, task specific, and environmental.

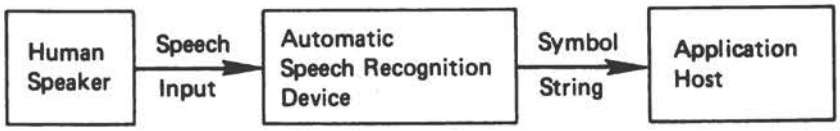

FIGURE 1 A generic automatic speech recognition system.

Device-specific parameters include the speaking mode and machine training or enrollment. The speaking mode may be in the form of (1) isolated words, where pauses are introduced between words; (2) connected words, where the words are concantenated without pauses; or (3) continuous speech, where the words flow smoothly, as in natural speech.

In speaker-dependent systems, speech samples are required from each user. This process is referred to as "speaker enrollment" or "training the system." Speaker-independent systems, by contrast, require no speaker enrollment. Some speaker-independent systems are speaker-adaptive, i.e., the system adjusts to the talker's speech patterns as he or she interacts with the system.

Task-specific parameters include vocabulary size and syntax. The vocabulary size refers to the number of words that the system recognizes. The acoustic similarity between words can affect system performance because words that sound similar are more difficult to distinguish from one another.

The syntax of a specific task is the (artificial) grammar that the system accepts for that task. The simplest type of syntax can be specified in terms of a finite-state network, where the words that are allowed after each state or node are given explicitly. More general syntax structures approximating natural language are specified in terms of a context-sensitive grammar (Woods, 1975).

One measure of syntax complexity is the average branching factor, which is roughly defined as the average number of words that are allowed at each node of the grammar. The branching factor is an important parameter that affects system performance considerably.

TABLE 1 Parameters that Characterize the Capabilities of Speech Recognition Devices

TYPE	PARAMETER	RANGE
Device Specific	Speaking mode	Isolated word to continuous speech
	Training (enrollment)	Speaker-dependent to speaker-independent
Task Specific	Vocabulary	Small (<20 words) to very large (>20,000 words)
	Syntax	Finite state to context sensitive
	Branching factor	Small (<10 words) to very large (>100 words)
Environmental	Signal-to-noise ratio	High (>30 dB) to low (<10 dB)
	Speaker stress	Low to high

While there are many environmental parameters, signal-to-noise ratio (S/N) is perhaps the best known. It is the ratio of the average signal power to background noise power, typically measured in decibels (dB). The difficulty of the recognition task increases as the signal-to-noise ratio decreases. The type of noise can also be very important (e.g., narrow vs. broad spectrum, impulsive vs. continuous, and stationary vs. time varying). Other environmental factors include microphone position, speaker variations (such as rate of speech, speaking level, and speaker fatigue), and psychological and physiological stress (elicited, for example, by combat dangers or by G-loading during flight in high-performance aircraft).

SPEECH GENERATION

Speech generation, or synthesis by machine, is the voice output function complementary to speech recognition for voice input. The fundamental approaches to speech synthesis have already been outlined on pages 6 and 7. Additional discussions are available in the literature (Sherwood, 1979; Butler et al., 1981). Machine synthesis of speech, coupled with automatic speech recognition, represent the constituent ingredients for voice-interaction between human and machine.

SPEECH RECOGNITION AND GENERATION IN A VOICE-INTERACTIVE SYSTEM FRAMEWORK

A human-machine system includes a human and a machine operating within an environment to accomplish a set of system goals. Figure 2 illustrates the machine, human, and environmental components of such a human-machine system. The right side of the figure includes the machine components. "Controls" permit the human to alter the machine status, and "displays" convey the machine status to the human.

On the left side of the figure are the human components of the system. These include sensing information from the displays, information processing and decision making, and responding, through action, to activate the controls. Note that both the machine and human elements of the system operate within a specific work environment.

In a voice-interactive system, the human action would be in the form of speech, while the control device would consist of a microphone and an ASR device. The information could be presented either visually (visual displays) or by synthesized speech (speech displays). The human would sense information visually or auditorily. System performance is a function of not only the capabilities and limitations of the controls and displays hardware, but also the capabilities and

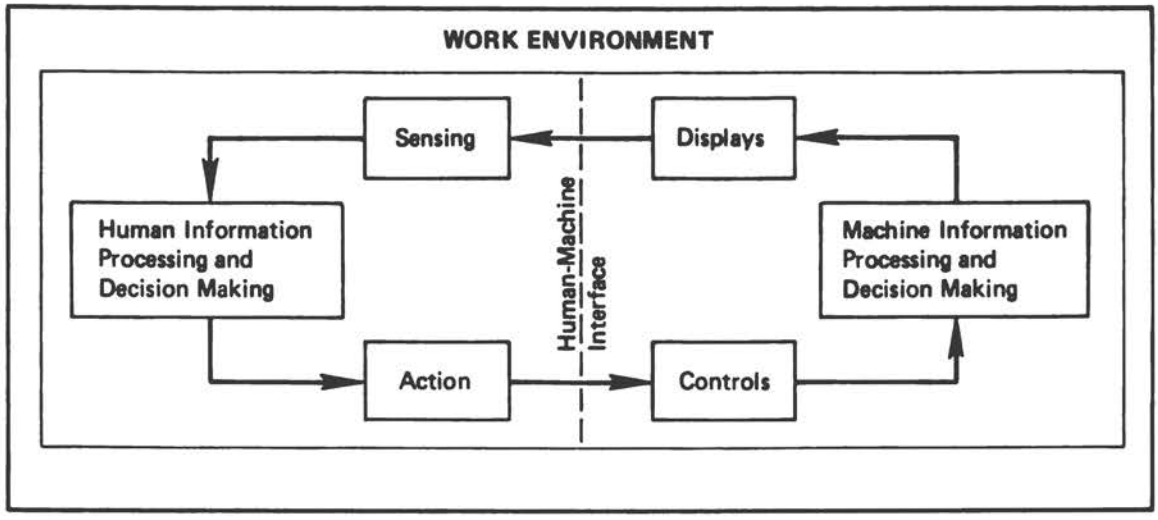

FIGURE 2 Human, machine, and environmental components of a human-machine system.

limitations of the human operator(s), the constraints of the environment and task, and the communication channel between the hardware and the human operator.

When automatic speech recognition is incorporated into a system, any of these factors (machine, human, and environment) can influence the acoustic signal transmitted to the speech recognizer and can affect recognition accuracy. This, in turn, affects the overall performance of the system. Therefore, speech recognition algorithms must take into account these sources of variability.

However, developing more robust speech recognition algorithms (i.e., ones that are less sensitive to variations in speech input) is only a partial solution for problems encountered in using speech controllers in an operational system. Selecting tasks where speech is a compatible and efficient form of data entry is important. If a speech recognizer is used as the control device (voice control) for a task where speech is not the appropriate communication channel (for the task or work environment), even a perfect speech recognition system will not result in optimum system performance.

In addition, the perceptual and motor capabilities of the human directly affect system performance, and human skill in production and understanding of language must be considered. When speech is used to present feedback or other information (speech displays), high intelligibility in speech output is necessary but not sufficient. The vocabulary, syntax, and content of voice messages must be designed to facilitate rapid and accurate comprehension. The naturalness of the speech output, e.g., its similarity to human speech, might or might not be a benefit, depending upon the application.

Inadequate design of the dialogue for communicating with automatic speech recognition systems may result in reduced system performance, independent of the accuracy of the speech controller or the speech generation system. For example, if recognition vocabulary selection is intended solely to provide minimal inter-word confusion for the recognizer, the words selected may not be those typically used to perform the task. In this case the human may have difficulty remembering the "legal" vocabulary. If speech displays give ambiguous messages, the operator may misinterpret them, resulting in degraded performance. Therefore, successful application of speech technology depends both on the performance of the recognizer and on the adequacy of the integration of speech into the voice-interactive system.

System improvements have sometimes been attributed to automatic speech recognition when, in fact, the improvements were due, at least in part, to one of the following: (a) the automation of functions previously performed manually, (b) the reduction in complexity of a data-entry task, or (c) the provision of machine understanding of the operator's intent. To assess the unique value of speech input/output technology to any application, the improvements resulting from the use of speech must be separated from the gains attained from other aspects of the implementation.

PERFORMANCE CRITERIA

In judging the value of ASR devices, at least two dimensions of performance need to be considered. These include the criteria associated with the ASR device itself and the performance criteria associated with the entire voice-interactive system. These two components will be discussed separately, since their constituent measures are different.

ASR Performance

No standards currently exist for testing ASR devices and assessing their performance, although some efforts are being directed toward developing testing standards for limited applications. The National Bureau of Standards is taking an active role in this effort (Pallett, 1982).

The most common ASR performance measure today is word accuracy, defined as the number of correctly recognized words divided by the number of input word tokens. The word error rate is then obtained by subtracting word accuracy from unity.

There are at least four primary kinds of word errors: substitution, insertion, rejection, and deletion. A substitution is the recognition of one word for another. An insertion is the recognition of a word when none is intended. (This may occur in response to extraneous sound, such as a cough, or to improper speech such as illegal syntax.) A rejection occurs when the device declines to recognize a detected input, usually because of inadequate match between input and legal vocabulary. A deletion is the failure of the recognizer to detect that an input utterance has been spoken.

The relative importance of the four types of errors depends on how the system is to be used. For many applications, rejection and deletion are less objectionable than substitution and insertion. For example, in one task, the relative costs of these types of errors might be assigned as: rejection, 1; deletion, 2; insertion, 5; and substitution, 10. Besides the relative importance of these classes of errors, specific errors may be critical for a given application. For example, substituting "eject" for "reject" or "fire" for "five" would be critical in a fighter cockpit. Unless otherwise specified, the term "error rate" in this report will include the sum of all four types.

A measure of word error rate is meaningless unless it is accompanied by a specification of system parameters and conditions under which the tests were performed. As a minimum, the specification should consider the system parameters shown previously in Table 1. Success in application often depends on how well these parameters can be controlled in order to obtain an acceptable error rate. Typical control techniques include the use of noise-cancelling microphones and

user training to promote consistent speech input. How well a recognizer performs with variations in the speech input, which are attributable to environmental and speaker factors, is often referred to as the "robustness" of the recognizer.

System Performance

The term "system performance" embraces the entire human-machine system. ASR performance and total system performance represent different levels of analysis. If the first is poor, the second is likely to suffer also (McCauley et al., 1982). But, even when recognition accuracy is adequate, system performance can still suffer due to failures in other subsystems or to overloading of the operator. It is important, therefore, to evaluate not only ASR performance, but the effect of the entire human-machine interface on total system performance.

The influence of the speech interface design can have a substantial and often subtle effect on total system performance. Features of the speech subsystem--such as vocabulary design, error correction methods, enrollment, and the presentation of recognition feedback--can be more important to system performance than overall ASR accuracy.

Recognition errors can significantly affect system performance because they force users to monitor accuracy and correct errors. Thus, throughput may be reduced and operator workload increased. System design should improve performance by minimizing the burden of this monitoring/feedback/correction task.

Measures of system performance include, but are not limited to, time and accuracy of performance of voice controlled tasks, operator performance of concurrent tasks, and measures of overall mission success.

3

CURRENT STATUS

This chapter presents the current status of automatic speech recognition (ASR) technology in the framework of a speech-interactive system.

SPEECH KNOWLEDGE

Speech is generated by closely coordinated movements of several groups of human anatomical structures. Muscles attached to the ribs and diaphragm build up pressure below the larynx, providing energy for speech. Voiced sounds are produced by positioning the vocal folds inside the larynx such that they are set into vibration by the airflow. Unvoiced sounds are produced by generating turbulence somewhere inside the vocal tract. (Examples of voiced sounds include the vowels and such consonants as /r,m,n,b,d,v,z/, while unvoiced sounds include such consonants as /p,t,f,s/.) Detailed acoustic characteristics of speech sounds result from changing the configuration of the vocal tract using anatomical structures such as the tongue, jaw, and lips.

As shown in Figure 3, the long-term power spectrum of the speech signal shows the presence of a significant amount of energy below 1 kHz, above which the spectrum falls off rapidly with frequency (Dunn and White, 1940). Figure 4 shows that, generally, the speech signal contains useful linguistic information in the frequency range of 250 Hz to 7 kHz (French and Steinberg, 1946). The detailed acoustic characteristics of speech differ from sound to sound.

Voiced sounds typically have a harmonic spectrum with sharp spectral prominences that correspond to the resonant frequencies of the vocal tract, while voiceless sounds tend to have a continuous spectrum with spectral peaks that are less pronounced.

Studies of language organization suggest that underlying the production and perception of speech there exists a sequence of basic discrete segments that are concatenated in time. These segments, called "phonemes," are assumed to have unique articulatory and acoustic characteristics (Chomsky and Halle, 1968). While there are

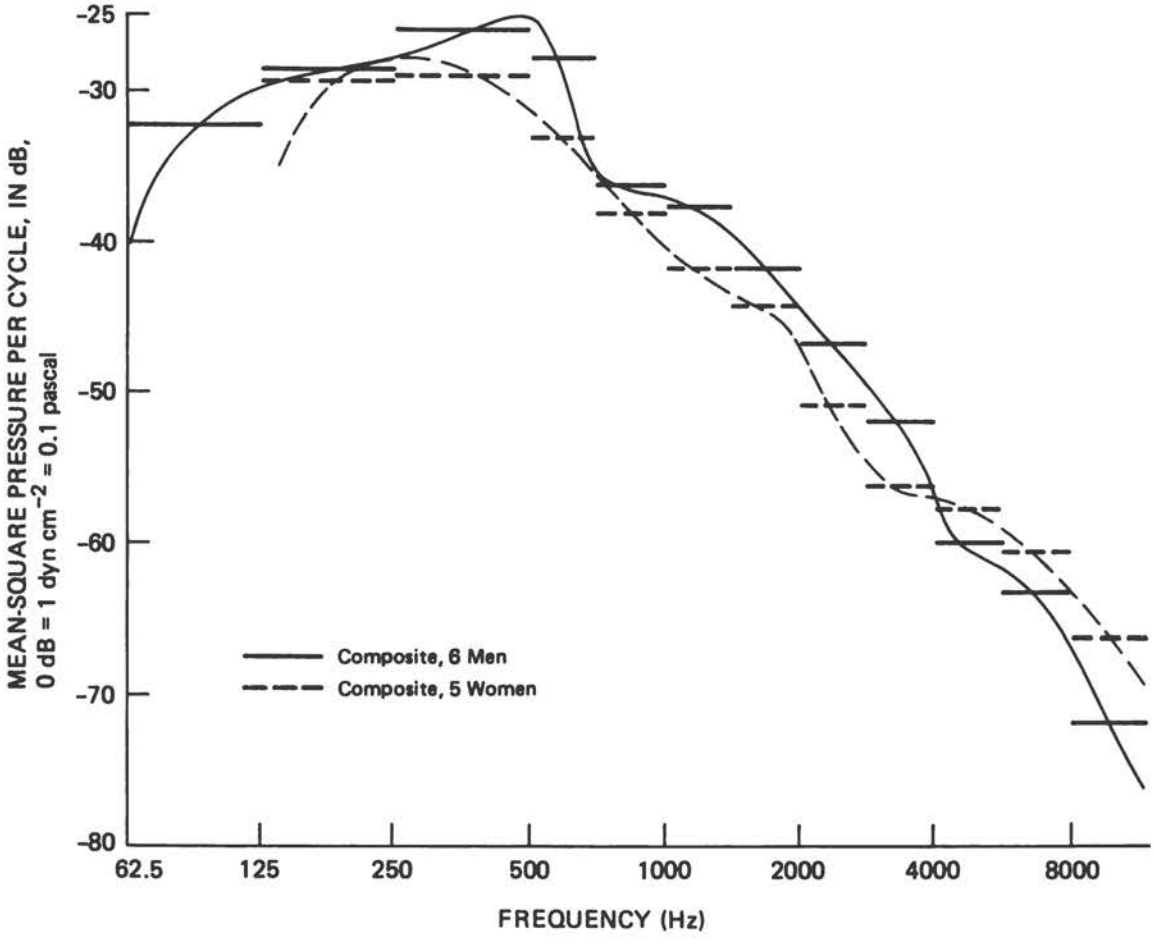

FIGURE 3 Long-term power density spectrum of continuous speech. The power spectrum shows a significant amount of acoustic energy below 1000 Hz. The spectrum falls off at a rate roughly inversely proportional to the square of frequency above about 800 Hz (Dunn and White, 1940).

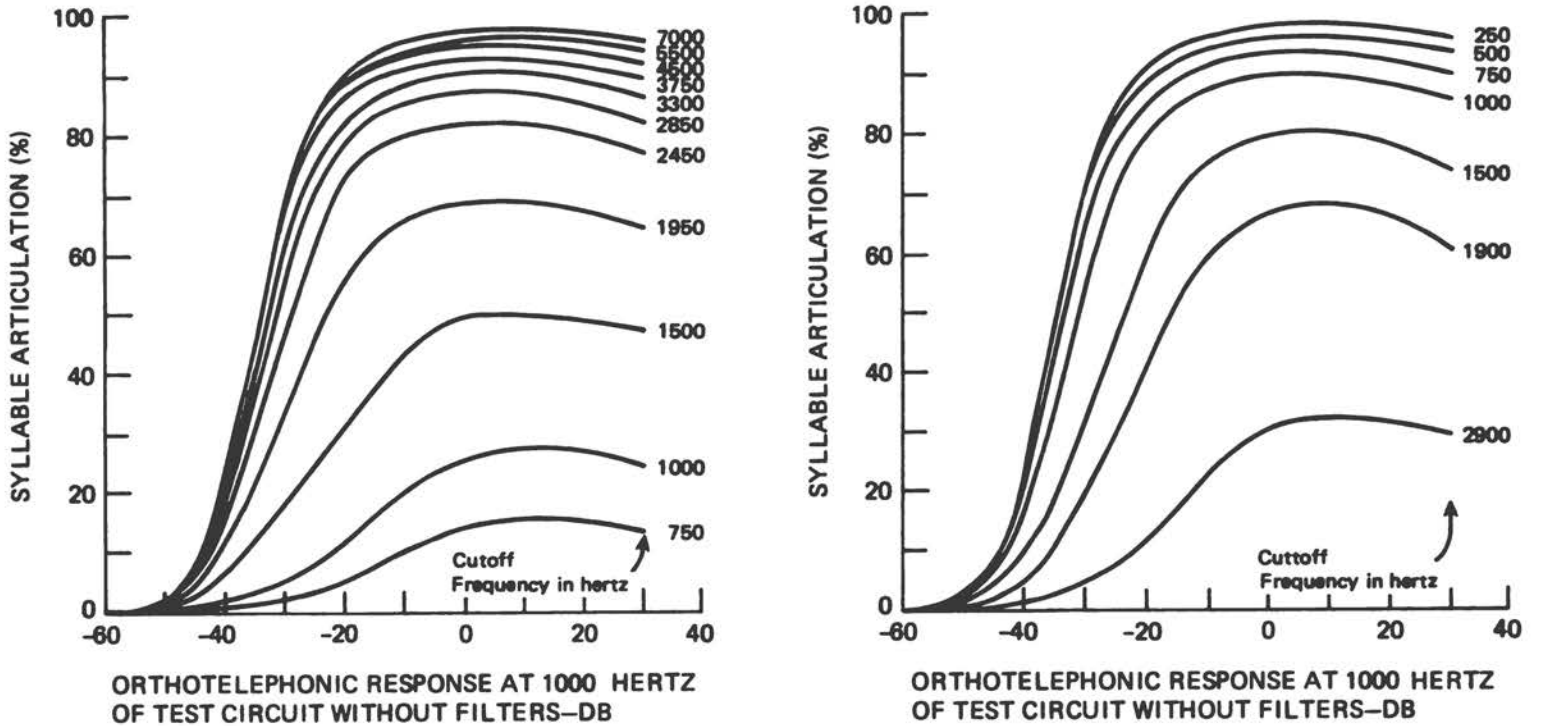

FIGURE 4 Typical speech intelligibility measurements for syllables, measured as a function of the bandwidth of the signal, after the signal has been high-pass (right panel) or low-pass (left panel) filtered. Syllable intelligibility (articulation) in percent is plotted on the ordinate, and relative signal level is plotted on the abscissa. Cut-off frequency of the filter is the parameter. The figure suggests that near perfect intelligibility can be obtained if the speech signal is band-limited to 250 Hz to 7000 Hz (French and Steinberg, 1946).

essentially an infinite number of articulatory gestures that can be produced by the human vocal apparatus, the inventory of the basic sound units, or phonemes, is remarkably limited.

American English, for example, uses some 16 vowels and 24 consonants. Each one of these sound units has contrastive acoustic characteristics and is combined with others to form larger units such as syllables and words. Knowledge about the acoustic differences among these sound units is essential to distinguish one word from another, such as "bit" and "pit." Over the past three decades, the acoustic theory of speech production has been formulated (Fant, 1960; Flanagan, 1972) and refined to the extent that the properties of speech sounds in restricted environments are fairly well understood.

When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment. This is due to the interaction among various anatomical structures and their different degrees of sluggishness. The result is an overlap of phonemic information from one segment to the other in the acoustic signal. For example, the same underlying phoneme /t/ in words such as "tea," "eaten," "steep," "beater," and "tweed" has drastically different acoustic characteristics. This effect, known as "coarticulation," can occur within a given word or across a word boundary. The word "this," for example, will have very different acoustic properties in phrases such as "this car" and "this ship."

Over the last decade, certain advances in articulatory phonetics and acoustic phonetics have been made. Our knowledge of the acoustic properties of speech sounds has increased from isolated consonant-vowel syllables to include words and sentences (Cole et al., 1980).

Normal speech can be degraded by additive noise (both acoustical and electrical). Abnormal, distorted speech may result from psychological and physiological factors of stress and acceleration, or from acoustic environmental factors (such as speaking into the back pressure of an oxygen mask or high-intensity auditory masking).

The only degradation for which there is significant quantitative understanding is normal speech with additive noise (Kryter, 1970; Lim, 1983). As such, the understanding is primarily in the relationship between the general properties of the noise and the intelligibility of the degraded speech. This relationship can be obtained directly from articulation tests in which the intelligibility of speech is measured for varying noise conditions, or from a quantity called "articulation index," which measures the amount of the speech spectrum that lies above the masking noise (Kryter, 1962).

Figures 5 and 6 show for human listeners some typical results of the relationship between intelligibility and signal-to-noise ratio (SNR) for additive broadband noise (Miller et al., 1951). The intelligibility at a given signal-to-noise ratio depends on the type of speech material and the size of the vocabulary. Similar or related behavior might be expected in the performance of recognizers operating

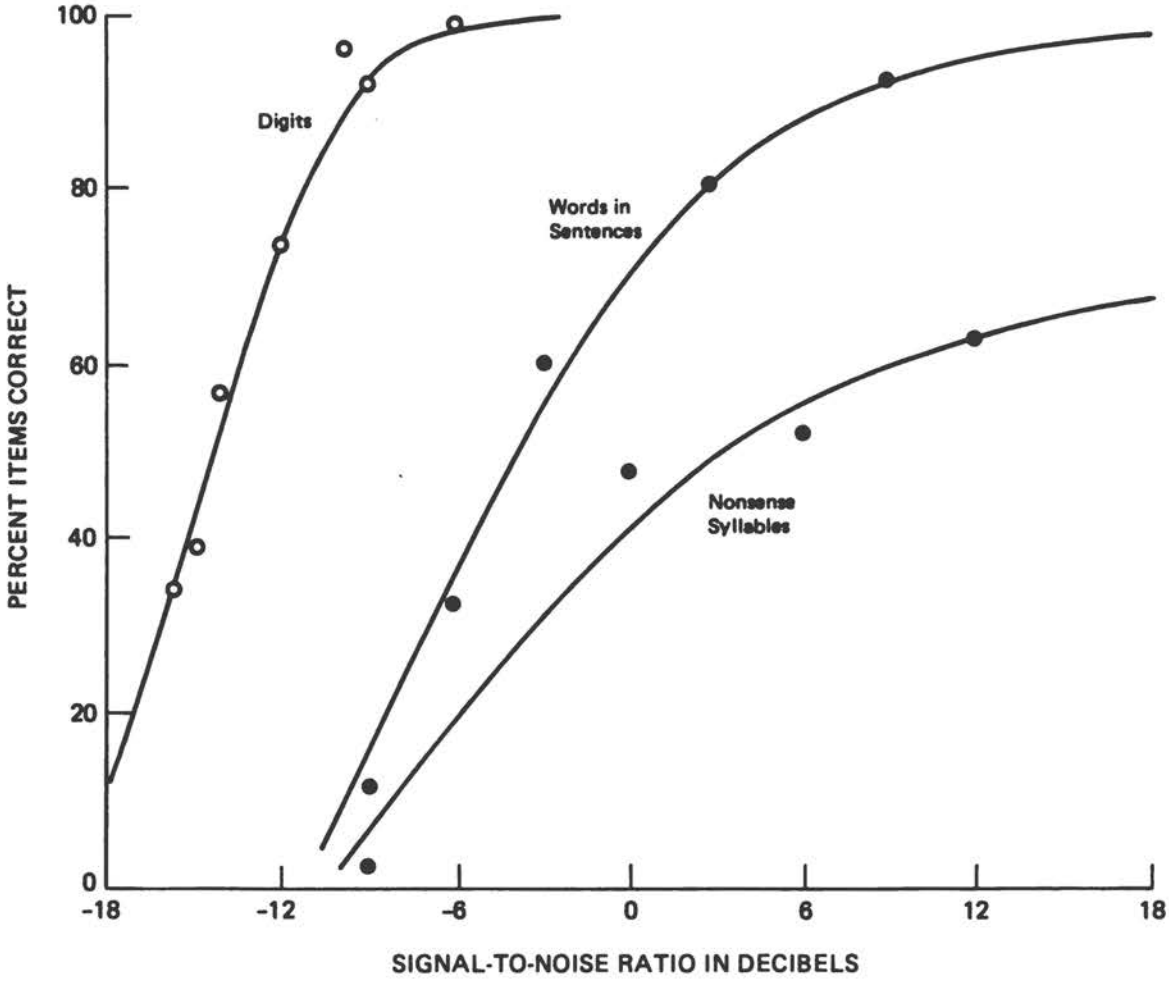

FIGURE 5 Intelligibility scores for different types of spoken material as a function of signal-to-noise ratio. Speech is produced under quiet conditions and masking noise is added to the original. (Miller et al., 1951).

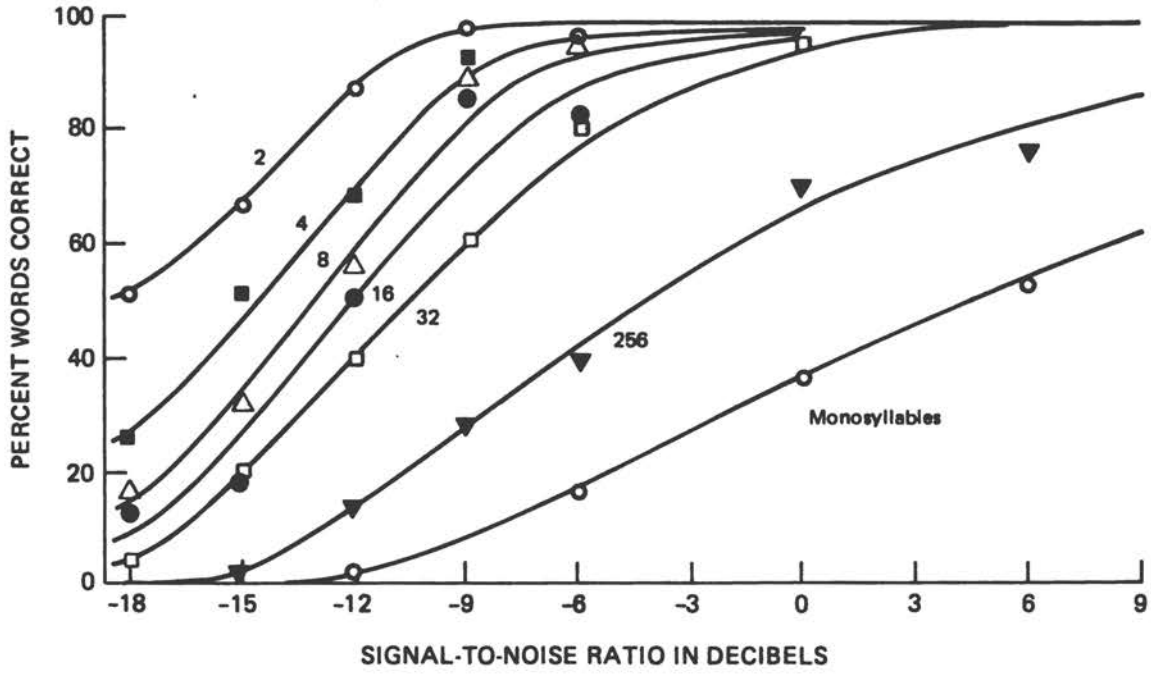

FIGURE 6 Effects of vocabulary size upon the intelligibility of noise-masked monosyllabic words (Miller *et al.*, 1951).

with noise-contaminated inputs. Under such conditions, any microphone or signal enhancement system that improves signal-to-noise ratio would be desirable.

By comparison, data on intelligibility and physical spectra of speech distorted by stress, acceleration, and noisy acoustic environment are exceedingly meager. Some of the data which are available have been collected within the context of recognizer performance evaluation (see, for example, Williamson and Curry, 1984).

ALGORITHMS

Algorithms for limited speech recognition are sufficiently well understood that some devices are available commercially. These commercial devices, with minor exceptions, typically deal with the recognition of a small set of isolated words spoken by a known talker. Within these limits, current devices span a wide range of performance, cost, and size. The more advanced systems, however, are still in the laboratory demonstration stage. For more details on types of algorithms, the reader is referred to Bahl et al. (1983), Dixon and Martin (1979), Lea (1980), Ney (1984), and Rabiner and Levinson (1981).

Isolated Word Recognition

Today's isolated word recognition (IWR) systems share several characteristics (see Figure 7). They generally require distinct pauses (typically 200 ms) between words and they usually operate in a speaker-dependent manner. Treating each word as a whole pattern, recognition is performed by matching the parameters of the input signal to stored templates for the vocabulary items. The word whose stored template best matches the input is selected as the recognized word.

The matching algorithm typically uses time-alignment procedures; the most successful is dynamic time warping (Itakura, 1975; Sakoe and Chiba, 1971), which is designed to account for the inherent temporal variability of the speech signal. These IWR systems usually operate on a small vocabulary of 10 to 200 acoustically distinct words or short phrases.

One of the important characteristics of present IWR technology is the need for user enrollment. Most IWR systems require each new speaker to provide at least one exemplar for each word in the vocabulary. The performance varies greatly from speaker to speaker. Moreover, the need to train a system for each user limits the size of the vocabulary that can be accommodated. As the vocabulary increases to, say, 10,000 words, the training procedure becomes so time-consuming as to render such systems impractical.

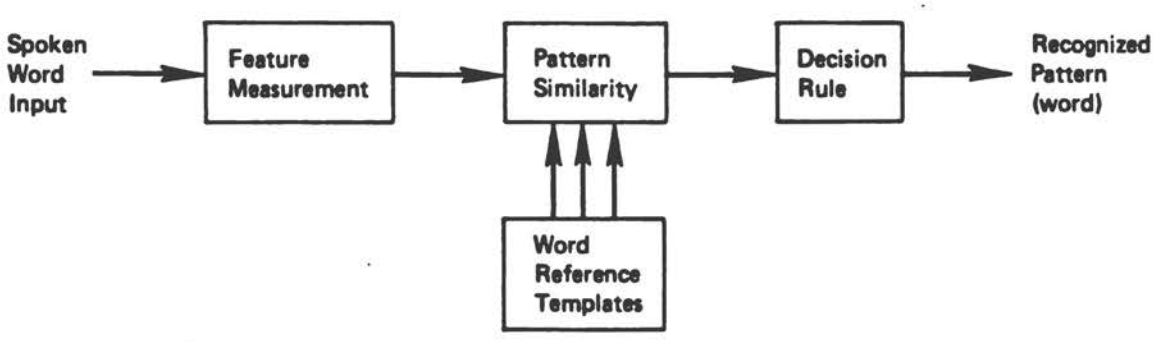

FIGURE 7 Block diagram of a typical isolated word recognizer.

IWR systems can be made speaker independent to a useful extent if the stored templates are representative of a wide variety of talkers. One of the more successful techniques for this is pattern "clustering" and the use of multiple patterns for each vocabulary item (Rabiner et al., 1979). However, these systems have been demonstrated typically for relatively small vocabularies, such as the spoken digits.

Some IWR systems utilize statistical recognition methods, with hidden Markov models playing a central role (Bahl et al., 1984; Baker et al., 1984). One major advantage of hidden Markov models is the existence of an automatic training procedure that has been used effectively to optimize performance for a given set of training data. Compared to template matching methods, some statistical methods require substantially more training data but less computation for recognition to achieve good performance. In principle, given sufficient training data, statistical methods are capable of achieving superior performance. However, for certain applications, and especially for speaker-dependent systems, it may not be practical to obtain sufficiently large amounts of training data to achieve the desired performance.

Current IWR devices typically rely on general pattern recognition algorithms that use little or no speech-specific knowledge. In contrast, some researchers (for example, Cole et al., 1983) have used feature-based recognition (Schwartz, 1982) as an alternative to spectral template matching.

The idea behind feature-based recognition is to identify the acoustic features that are needed to define speaker-independent patterns for linguistic events. Upon acquiring this knowledge, algorithms are created to extract the features from speech. Then, the features are put into a classifier designed to integrate cues into a unified decision. While feature-based recognition has been studied for over a decade, its success relative to pattern-matching algorithms has not been fully demonstrated.

Connected Word Recognition

Most commercial IWR systems require that consecutive words be delineated by pauses. This is necessary because in continuous speech the acoustic signal is often altered at word boundaries, thus making it difficult to determine where a word ends and another begins. For example, the word-final and word-initial /s/ in phrases such as "gas station" usually merge into a single sound. Because pausing between words is unnatural, systems that require such a speaking mode are unlikely to be used for many tasks.

There have been attempts to generalize IWR algorithms to deal with connected words. The approach usually involves scanning across the entire utterance for all possible word matches, with a relaxation of the matching criteria at word boundaries (Kato, 1980; Myers and Rabiner, 1981). Syntactic constraints have also been imposed in some

algorithms to prune potential word candidates. Because these connected-word recognition systems generally operate in a speaker-dependent mode, the size of the vocabulary and the length and complexity of the sentences are often quite restrictive.

Connected word recognition (CWR) systems deal with a restricted class of continuous speech phenomena. They slightly extend the techniques developed for isolated word recognition to deal with a class of connected word sequences in which the words are pronounced essentially the same regardless of their position in the sequence. Digit strings and lists of items are usually pronounced in this way. However, since natural English sentences do not have these restricted properties, the techniques used for connected word recognition are not adequate for handling unrestricted, natural speech. Complicating phenomena in naturally spoken English include variations in energy level, pitch, and duration associated with sentential intonation, different word pronunciations depending on syntactic function and position within a sentence, and complex context-sensitive constraints on the sequences of words that constitute acceptable inputs.

Continuous Speech Recognition

Continuous speech recognition (CSR) systems attempt to deal with the wider range of phenomena found in natural speech described earlier. They use linguistic knowledge (syntax and semantics) to characterize the range of acceptable sentences (Hayes-Roth, 1980; Woods, 1975). They typically have large vocabularies (1000 words or more) and often use phonetic, phonological, and/or syllabic representations. They analyze a sentence by formulating, extending, and evaluating alternative hypothetical word sequences that can be matched to the acoustic input and checking them for syntactic, semantic, and pragmatic acceptability as well as degree of acoustic match.

Research in continuous speech recognition received significant impetus in the early 1970s when the Defense Department's Advanced Research Projects Agency* (ARPA) initiated an ambitious five-year, multi-site effort to develop speech understanding systems.** The project sought to develop systems that would accept continuous speech from many speakers, with minimal speaker adaptation, and operate on a 1000-word vocabulary, artificial syntax, and a constrained task domain (Newell et al., 1973).

---

*ARPA is now called the Defense Advanced Research Projects Agency (DARPA).

**The term "speech understanding systems" was used by Newell et al. (1973) to refer to systems where the recognition errors that count are not errors in speech recognition per se, but errors in task accomplishment.

Under the ARPA project, from 1972 to 1976, several systems were developed, aiming at the 1000-word vocabulary level, with differing aspirations towards the inherent language complexity and real-time issues. The grammars that were used ranged from finite language grammars to more ambitious context-sensitive grammars with discourse capabilities. Acoustic processing ranged from template matching techniques to feature-based acoustic-phonetic methods. For detailed reviews of the ARPA project, see Klatt (1977) and Lea and Shoup (1980).

Since the termination of the ARPA project in 1976, CSR research in the United States has been quite limited. One notable exception is the effort (Dixon and Tappert, 1973) at the IBM T. J. Watson Research Center. The goal, here, has been to develop a speech transcription system, with no current plans to include a speech understanding capability (Bahl et al., 1983). The work uses a statistical language model, including a hidden Markov phonetic model.

Because research on continuous speech understanding is still very much in its infancy, no practical systems are available. Considerable research will be required in this area before practical applications can be expected in other than the simplest of applications.

Continuous speech understanding systems have so far dealt almost exclusively with speech that is read or pronounced clearly. A difficult challenge for the future is understanding spontaneous speech, where speakers are composing ideas as they speak. Besides the complexities of continuous speech, spontaneous speech involves the phenomena of false starts, hesitation pauses, prolonged syllables and voiced pauses, and frequently ungrammatical sentences. Progress here will require advances in mechanical inference and computational linguistics as well as in speech processing, acoustic phonetics, and other speech-specific disciplines.

Keyword Recognition

Thus far in this report, the range of speech allowed as input to an ASR system has been assumed to be limited chiefly by vocabulary size. Such "closed-set" recognition requires cooperative speakers who restrict their speech to the given vocabulary. There are applications where it is desirable to recognize a limited set of words from essentially unrestricted continuous speech. This type of "open-set" recognition is known as "keyword recognition" and is often referred to as "word spotting." The vocabulary items are the key words that are searched for by the recognizer. The system rejects all other words and extraneous acoustic sounds.

Major applications of keyword recognition include situations in which one has no control over what a speaker utters (unknown or uncooperative speaker) or where one wants to free a cooperative speaker from the constraints of a fixed vocabulary. In fact, many isolated and connected word recognition systems are actually

implemented as keyword recognition systems. A word is recognized by the system only if it sufficiently matches a certain stored template; otherwise the input is rejected.

When mentioned in this report, we shall restrict the use of the term "keyword recognition" to applications where the speaker is unknown or uncooperative. In essence, one is dealing with speaker-independent recognition of certain key words in continuous, unrestricted speech. The recognition algorithms employed generally use some form of template matching, as in connected-word recognition.

Computational Requirements

In each ASR system, there are various factors that affect computational requirements. The nature and relative importance of the factors depend upon the specific recognition algorithm. In general, however, computational load increases with vocabulary size. Depending on the specific algorithm employed, IWR and CWR systems with vocabularies of moderate size (less than 100 words) require machines that operate in the range of 100,000 to 10 million instructions per second (Mips) for real-time recognition, where each mathematical operation such as multiply or add is considered a single instruction. (The definition of "real-time" can vary depending on the task, but recognition delays of up to 300 ms are considered acceptable for many applications.)

Continuous speech recognition systems with large vocabularies require significantly larger computational capabilities. Current systems for CSR research require on the order of 200 Mips or more for real-time recognition, and future, more ambitious systems are expected to require machines with equivalent speeds of about 10,000 Mips. Research on these advanced recognition systems is impractical in today's laboratory, where a typical machine has a speed of 1 Mips, which can be increased to about 10 Mips with the addition of an array processor. Laboratory computers with speeds on the order of 100 Mips will be needed if research on advanced systems is to take place in a meaningful way.

HUMAN FACTORS INTEGRATION

Human factors research on the design of voice-interactive systems seeks to identify appropriate applications for automatic speech recognition and to apply human factors methodology to effectively integrate speech subsystems into the overall task of the operator. The current state of the art in human factors research provides some, but not all, of the procedures with which to attain these goals (McCauley, 1984). The task is challenging because ASR technology is continually evolving, and guidelines for using it will depend on many interrelated variables. These include the characteristics of the

users, the physical and task environments, the ASR system design, and the task itself.

Human factors research on the design of integrated voice systems is limited and reports are spread among conference proceedings, government technical reports, and journal articles (see for example, Cotton and McCauley, 1983; McCauley, 1984, Pallett, 1982; and Simpson, in press-b). Unfortunately, none of the suggested design guidelines contained in these reviews appear in standard references used by system design engineers (e.g., Department of Defense, 1981; Van Cott and Kinkade, 1972; Woodson, 1981).

Studies of ASR System Design Issues

Task Selection

Incorporating speech into complex systems has the potential for significantly reducing the visual and manual information workload. While this reason is sufficient to investigate a task's appropriateness for speech, the decision to use voice for a particular task requires weighing the advantages and constraints of the visual/manual modes versus the speech/audition modes in the context of the task to be performed. Two approaches have been taken to research on task selection for ASR. Some researchers have aimed to develop and apply methodologies for selecting appropriate tasks for speech (North and Lea, 1982) and user-preference questionnaires for application of voice recognition and speech generation (Cotton et al., 1983; Kersteen and Damos, 1983). Other researchers have investigated human speech data entry performance when simultaneous verbal and manual tasks are required. For example, one study determined that speech is faster primarily for complex tasks requiring cognitive and/or visual effort. Simple tasks involving the copying of numeric data were accomplished more quickly and accurately with keyboards than with voice entry (Welch, 1977).

Basic time-sharing research using speech recognition and synthesis provides some evidence that speech is better than manual input and visual output for some types of tasks (Wickens et al., 1983). When subjects performed two tasks simultaneously, one spatial and the other verbal, spatial task performance was better when the verbal task was accomplished using speech than when both tasks had to compete for the manual and visual channels. Wickens' (1984) theory of Stimulus-Central Processing-Response Compatibility offers a potentially useful framework for selecting appropriate tasks for speech.

In summary, the results of these several studies suggest that the benefits to be derived from voice input and output are highly task dependent and that more research is needed, using realistic tasks (Aretz, 1983; Mountford et al., 1983), to investigate further modality interactions.

User Characteristics

User characteristics can affect ASR system performance. Successful applications of ASR to date usually involve a small number of carefully selected talkers who have been trained to speak distinctly and use the equipment correctly. One study (Doddington and Schalk, 1981) reported that three-fourths of the talkers tested had better-than-average recognition scores, indicating that a few of the talkers experienced the majority of the problems.

Enrollment

Enrollment is a critical element in speaker-dependent ASR systems. The most successful enrollment techniques seem to be those that avoid any systematic bias in the speech samples. For example, recognition accuracy is better when the several tokens of each vocabulary item are sampled randomly instead of collecting all tokens of a vocabulary item in sequence (Poock, 1981a). Further, recognition performance is enhanced when enrollment occurs in an acoustic environment similar to that of operational conditions (Coler, 1982; Kersteen, 1982).

System Feedback to the User

Feedback by the system to the user typically enhances performance. One study of the level and mode of feedback revealed that performance improved when subjects, who were not accustomed to feedback, were presented with some type of feedback. Conversely, performance was degraded by reducing the feedback to which a user had already been accustomed (Poock et al., 1983). However, various approaches to presenting feedback are not equally effective (Schurick et al., 1984). In the absence of feedback, a user may assume, incorrectly, that a sequence of voice commands was executed properly by the system. Although there is general agreement about the need for feedback, the problem is how to best provide that feedback without interfering with the operator's primary task.

Error Correction

System performance can be improved by two types of error correction--automatic and user activated. In the first, the system can be designed to detect an illegal input sequence automatically, change it to the most likely legal sequence, and then, if desired, present the correction to the user for verification. For example, with syntactically constrained dialogues the parser can iteratively evaluate both the first- and second-choice vocabulary items returned from the recognizer (Spine et al., 1983).

In addition to automatic error correction, provision can and should be made for error correction by the user. Three documented types of user errors include: (a) failure to remember the vocabulary set, (b) failure to follow the speech cadence (input timing) restrictions, and (c) conversing with co-workers with an active microphone. Vocabulary errors involve speaking words outside the vocabulary, including synonyms; cadence errors include using connected speech with discrete word recognizers (McCauley and Semple, 1980).

Comparative Input Modality Studies

Differing results have been obtained in research comparing speed and accuracy of voice versus manual keyboard input, depending on the unit of input (alphanumerics or functions) and other task specific variables (McSorley, 1981; Poock, 1981b). But, as noted previously under task selection, cognitively and/or manually demanding tasks may be performed faster with voice input, while simple data entry tasks that do not involve concurrent visual and manual tasks may be accomplished faster with manual input (Welch, 1977).

Simulation of ASR Devices

Research using simulations of speech displays and controls originated with studies of how people communicate to solve problems (Chapanis, 1975). Problem solution occurred most rapidly whenever the voice link was available. An analysis of the linguistic output of the communicators revealed a consistent failure to follow grammatical rules. Using the voice channel, communications were extremely wordy (five times as many words), the vocabulary was four times larger, and the words were communicated ten times faster than with typewritten or handwritten communications. These data characterize human voice communication independent of hardware constraints. This study did not restrict the speech channel in vocabulary, syntax, or permissible speaking cadence, as is done with current ASR capabilities, but the results illustrate the power of voice communication for problem solving and emphasize the importance of further development of ASR technology.

Several attempts have been made to study system performance and acceptability when the speech channel is restricted in various ways to simulate the use of ASR hardware. One experiment simulated a "listening" typewriter where speech was constrained either in terms of vocabulary size or speech pause requirements (Gould et al., 1983). Shortcomings of the simulation included slow response time, failure to simulate substitution errors as well as rejection errors, and inconsistent restriction of discrete data entry when the spelling mode was used to enter words not in the vocabulary. Another study (Poock and Roland, 1982a) demonstrated the difficulty of designing a good

simulation of ASR. The study attempted to evaluate user acceptance of various levels of recognition accuracy. All levels of ASR accuracy tested in the simulation were judged acceptable by the subjects, probably indicating simply that they liked the concept of voice input. Because the simulation was believable for participants, it serves as a foundation for developing techniques to simulate speech recognition for human factors research.

Speech Display Design

Speech displays (the presentation of information via computer-generated speech) may be appropriate for several different functions. These include, but are not limited to, feedback from the ASR system to the user, responses to user queries, annunciation of system status, and warnings. Research on message design (i.e., wording, speaking rate, voice pitch, and voice type) for voice displays has been performed almost entirely on voice warnings (Simpson, in press-b). As yet, little systematic study has been directed toward voice message design for system feedback and responses to queries. The choice of machine-like or human-like voice quality appears to be application dependent. When human voice communications are simulated, as in training systems, human-like voice is preferred (Cotton and McCauley, 1983). Conversely, when a machine displays information using voice, as in cockpit warning systems, a machine-like voice is preferred for its distinct quality and for the cue it provides regarding the identity of the speaker (Simpson, in press-b).

Task Environment

The major environmental factor that has been studied to date is the effect of background noise on recognition accuracy. Little is known of the effects of environmental noise on human performance while using ASR devices. Some relationships between psychophysiological state and voice parameters have been investigated, including changes in laryngeal tension, rise in the fundamental frequency, pitch perturbations, and breath noises (Hecker et al., 1968; Huttar, 1968; Kuroda et al., 1976; Lieberman, 1961; Williams and Stevens, 1969). But, the complete picture of the effects of stress on speech production remains to be established. A major question is how to quantify and reproduce known levels of stress while speech measurements are being made.

Because stress-related changes in speech can take many forms and are not consistent among people or tasks, ASR performance may vary dramatically. Currently, successful applications of ASR do not involve severe time constraints or life-threatening situations.

PERFORMANCE

It is not possible to assess comprehensively the performance of current speech recognition technology. This is due to the empirical nature of the technology (no underlying theory of predictive value exists) and to the dependence of performance on a vast array of factors. The only practical method for determining the performance of a recognizer today is to implement such a system and measure actual performance. However, the predictive power of this technique is limited by the sample sizes that are required to account for performance variability among users. Oftentimes, tests are performed with a small sample of experienced users. As a result, a recognition technology demonstrates poorer performance than expected when taken from the laboratory into the field.

System performance is often measured differently from ASR performance, in the sense that a system input may comprise several or many words to be recognized. Thus a recognition error rate of 1 percent (per word; i.e., one word out of 100) may escalate into a system input error rate of 10 percent (e.g., one 10-word message out of 10 messages) or more, depending on the composition of the input message.

The performance of current recognition technology depends strongly on basic recognizer specifications such as speaker dependence, vocabulary size, and speaking mode. Performance also depends strongly on environmental factors such as noise level and stress on the user.

IWR and CWR in Benign Environments

Best recognizer performance is achieved for isolated-word recognition in a speaker-dependent mode in benign environments. It is common for such systems to claim a recognition error rate of less than 1 percent. A sampling of six such systems evaluated on a 20-word vocabulary using a common data base demonstrated a range of performance from 12 percent error down to 0.2 percent error (Doddington and Schalk, 1981). Performance roughly correlated with the price of the recognizer, which ranged from a few hundred dollars to several tens of thousands of dollars. Recently, a number of low-cost, high-performance, small vocabulary IWR systems have begun to emerge (Baker et al., 1984).

Recognition performance for speaker-independent speech recognition is often significantly poorer than for speaker-dependent recognition; an error rate of about 5 to 10 percent is typical of laboratory results for small vocabularies. For digits spoken in isolation, error rates as low as 2 percent have been reported in laboratory experiments (Rabiner et al., 1979). Operational usage of speaker independent recognition needs to take account of the performance across large populations. Such characterization has not yet been fully explored.

It is commonly accepted that recognition performance typically degrades as vocabulary size increases, but few studies have been performed to quantify the relationship between performance, vocabulary size and similarity of the vocabulary items (Spine et al., in press). Some studies, in fact, have shown no remarkable difference in performance as a function of vocabulary size (Kaneko and Dixon, 1983). In operational systems, recognizer performance is often maintained at a high level by syntactic control of the active recognition vocabulary. Most applications currently use less than 100 words in the active vocabulary.

Perhaps even more important than vocabulary size is the acoustic distinctiveness of the various vocabulary words. Most application development efforts analyze this distinctiveness and modify the vocabulary as appropriate.

The performance of connected word recognition is necessarily poorer than that of isolated word recognition. This is directly attributable to several additional problems that pertain to connected word recognition. The most important is that the acoustic variability of the speech signal is greater in connected word recognition. This increased variability is caused by the effect of coarticulation mentioned earlier.

With few exceptions commercial and laboratory efforts in connected word recognition do not, as yet, attempt to model this variation. To the extent that words spoken in connected speech exhibit a greater acoustic variation, their recognition will be accompanied by a correspondingly higher error rate. It is, therefore, difficult to offer a simple characterization of the performance of connected word recognition technology without knowing, for example, the rate of speech. Typical usage involves experienced users speaking words in connected utterances but with careful articulation. Word error rates of 1 to 25 percent have been reported in small-vocabulary task domains (Bell, 1982; Levinson and Rosenberg, 1978).

Continuous Speech Recognition

The performance of CSR systems can be specified in terms of word, sentence, and semantic accuracy. Sentence accuracy is the percentage of recognized sentences with no word errors. Semantic accuracy, sometimes referred to as task accuracy, is the percentage of sentences understood and tasks performed correctly by the system.

In specifying the performance of a CSR system, it is important to determine the difficulty of the specific task to be executed by that system. One relatively objective measure of difficulty of a task is its "branching factor," which can be interpreted roughly as a measure of the number of words that can follow each word in a sentence. Two definitions of the branching factor have been used: one is known as the static branching factor (Goodman, 1976); the other is perplexity (Bahl, et al., 1983), which is defined using information-theoretic

concepts. For any given task, perplexity is always less than or equal to the static branching factor, but the ratio of the two can vary greatly depending on the task.

Another measure of task difficulty is the average sentence length (in number of words). Generally, a longer sentence is likely to have more errors. Sentence error rate can be viewed as a function of word error rate, branching factor, and average sentence length. In general, an increase in any of these three factors increases the sentence error rate.

Table 2 shows a comparison of a representative sample of CSR systems. The Carnegie-Mellon University (CMU) HARPY (Lowerre and Reddy, 1980) and Bolt Beranek and Newman Inc. (BBN) HWIM (Wolf and Woods, 1980) are speech understanding systems while the IBM (Bahl et al., 1983) is a transcription system for which semantic accuracy is not applicable.

The HARPY and IBM systems were both tested on retrieving technical abstracts. The task was specified in terms of a finite-state grammar with a relatively small branching factor. The BBN HWIM system had a travel management task, which was specified in terms of a context-sensitive grammar with a large branching factor. The IBM laser patents task, described probabilistically, had a relatively large perplexity.

All tasks shown in Table 2 had vocabularies of about 1000 words, but vocabulary size was not a consistent predictor of performance. The branching factor, perplexity, and average sentence length were far more important in predicting performance. Note that larger branching factors and perplexity resulted in decreased performance, as expected.

The IBM and HARPY experiences (cf. Table 2) have shown that finite state grammars with small branching factors can be used with a modified dynamic programming algorithm, or a variant of some statistical decoding algorithm, to recognize continuous speech efficiently. The HWIM system explores a completely different set of techniques for a more ambitious task, although its potential was never fully explored.

Keyword Recognition

The work on keyword recognition has been minimal. Thus, performance results for keyword recognition are meager. Generally, however, the performance of speaker-independent keyword recognition has been far below the reported performance of continuous speech recognition systems. Typical word detection probability for a 10-20 keyword vocabulary is usually in the range of .20-.60 in an acoustically benign environment, depending on a specified false-alarm rate (Bamberg et al., 1981; Golibersuch, 1983; McCullough, 1983).

TABLE 2 Performance of Representative Continuous Speech Recognition Systems. All tasks had vocabularies of about 1000 words each. These data are taken from Bahl *et al.* (1983), Lowerre and Reddy (1980), and Wolf and Woods (1980). (Where the entry space is blank, information for entry was unavailable from the literature.)

SYSTEM	CMU HAPPY	BBN HWIM	IBM	
TASK	ABSTRACT RETRIVAL	TRAVEL MANAGEMENT	ABSTRACT RETRIVAL	LASER PATENTS
STATIC BRANCHING FACTOR	33	196	33	-
PREPLEXITY	4.5	33.8	4.5	24.1
AVERAGE SENTENCE LENGTH (in words)	6	6	6	20
NUMBER OF SPEAKERS	2 MALE 3 FEMALE	3 MALE	1 MALE	1 MALE
TRAINING	SPEAKER-DEPENDENT	SPEAKER-NORMALIZED	SPEAKER-DEPENDENT	SPEAKER-DEPENDENT
WORD ERROR RATE	2%	-	0.1%	9%
SENTENCE ERROR RATE	9%	59%	1%	-
SEMANTIC ERROR RATE	5%	56%	NOT APPLICABLE	NOT APPLICABLE

Integrated Systems

When an ASR system is integrated into a larger human-machine system, performance must be measured in terms of overall system performance. This includes both recognizer and human error as a function of the task environment. In applications where the operator is involved in critical, high-workload tasks, performance analysis should include the range of performance as well as the average performance attained.

General methods for performance measurement are available for the different levels of human-system performance. Techniques are known for measuring human performance in isolation, in the context of specific tasks, and in the context of particular ASR subsystems. Commonly accepted measures include response time and accuracy, and performance on secondary tasks (Ogden et al., 1979).

Measuring the performance of ASR in a task is more difficult and less well understood. In addition to speed and accuracy of operator performance of complex tasks, it is necessary to measure such other variables as operator workload, ability to deal with novel situations and emergencies, and conflicts between speech controls and displays and other controls and displays. General methodologies for measurement at this level remain to be developed. In addition, metrics are needed to predict recognizer performance across the range of several variables simultaneously (Spine et al., in press).

Performance in Noisy/Stressful Environments

A number of environmental factors can distort the speech signal and, thus, adversely affect recognizer performance. These include high ambient noise, microphone and oxygen face mask combinations, acceleration, and vibration, as well as talker disorientation, stress, and fatigue. As we proceed down the list, our understanding of the effect of these factors becomes more vague and anecdotal.

Clearly, changes in the speech signal do occur when a human is under stress or performing extremely difficult tasks. As noted earlier, speech under stress may vary in terms of amplitude, frequency, pitch contour, precision of articulation, and spectral slope. These factors all act to degrade speech recognition performance. In the following paragraphs we consider the robustness of current recognizers against these environmental factors and the limitations of current techniques.

Mental loading has also been shown to degrade performance even in nonstressful environments (Armstrong and Poock, 1981a). This is presumably attributable to the diversion of effort from the speech input task. As the difficulty of the mental task or the stress increases, so does the recognition error rate. Other dependencies, such as with task duration or fatigue, have also been demonstrated (Armstrong and Poock, 1981b). These studies suggest that current

recognition technology is rather fragile in the sense that various secondary factors have significant effects on voice characteristics and, ultimately, on recognition performance.

Experiments indicate that machine recognizers are considerably more sensitive to noise than are humans. In one study, ASR error increased by an order of magnitude when signal-to-noise ratio decreased from 30 to 10 dB (Neben et al., 1983). The data also indicated that the best recognition performance was achieved when the talker was enrolled in the same acoustic environment as that of the anticipated operating conditions (Simpson et al., 1982). However, it is often impractical to train a system for a large variety of noise environments.

Speech enhancement, or noise stripping, has been shown to be effective in reducing the perception of communication channel noise, thus increasing speech understanding by human listeners. These techniques also have been demonstrated to improve the performance of ASR systems in wideband random noise and aircraft cockpit noise (Cupples, 1984; Neben et al., 1983). Effective wideband reduction is on the order of 12-14 dB. Further research is needed to develop recognition techniques that can adapt effectively to varying noise conditions.

Several recent studies indicate the behavior of speech recognition systems in noisy environments. In two studies, high performance was achieved in a helicopter environment only when enrollment was conducted in the same noise environment as subsequent recognition tests (Kersteen, 1982; Simpson et al., 1982). Contrasting results were found in a subsequent study using a different recognizer (Coler, 1982) where the recognizer made only one error in 3200 spoken input words despite a noise level of 100 dBA. Unfortunately, the signal-to-noise ratio was not measured in this study.

Another, more extensive study evaluated the effects of noise, oxygen mask, and G-force loading on operational voice data entry in an advanced F-16 cockpit (Werkowitz, 1984). In this project, two vendors supplied speech recognition units for inflight testing. During these tests, the best performing system exhibited an error rate of about 10 percent over all noise and G-force conditions, with no significant degradations up to 110 dBA noise and 5 G's.

In a subsequent, controlled test of performance in a simulated F-16 noise environment (Williamson and Curry, 1984), a 70-word vocabulary data base was collected from five subjects. The data were collected in noise levels up to 112 dBA, with the pilot using an oxygen mask and M101 microphone. Signal-to-noise ratio for the 112 dBA condition was approximately 20 dB. Speech recognition devices from five potential vendors were evaluated using this data base. Evaluation was performed using three vocabularies of 25, 25, and 20 words. The best performing system demonstrated an average error rate of 7 percent (3 percent substitution and 4 percent rejection) with test data collected in noise at 97, 106 and 112 dBA. Recognition error was about twice as great at 112 dBA than at 97 dBA. Some talkers achieved much better performance than others; error rates ranged from 2 to 16 percent among the talkers who participated.

The recognition system tested in the Coler (1982) helicopter study was also used in the F-16 study. Although it performed extremely well in the helicopter test at 100 dBA, performance was more modest in the F-16 test, with a substitution rate of 6 percent for the lowest noise level (97 dBA). The two tests, however, used different versions of the vendor's algorithm, different vocabularies, and different talkers. This strongly reinforces the argument that the plasticity of the speech input medium makes performance forecasting difficult.

The "rejection accuracy" of speech recognition devices was also tested in the F-16 study. Rejection accuracy was defined as the probability of rejecting an out-of-vocabulary word (taken from one of the two sub-vocabularies not currently active). This is often an important parameter when occasional extraneous acoustic input or syntax errors occur. For the best performing system, the average rejection accuracy was measured as 73 percent. That is, about three quarters of the time the incorrect word was properly rejected by the recognition device. The remainder of the time the incorrect word was mistakenly accepted as one of the active vocabulary words.

The value of testing exploratory hardware under actual field conditions cannot be emphasized enough, especially for severe environments where laboratory duplication and control are difficult if not impossible. These field experiments with prototype hardware often delineate problems which can then be studied in the laboratory.

Summary of Limitations of Present Techniques

Current speech recognition technology is fragile; that is, recognizer performance, which can often be demonstrated favorably in the laboratory, may degrade significantly under the effects of acoustic noise, user stress, and operational conditions. Thus, recognition depends not only on detailed system specifications but on a myriad of subtle factors as well. In the laboratory, speaker-dependent, IWR systems often can recognize up to 100 words with about a 1 percent error rate, but the performance of operational systems often falls far short of this. Extension to speaker independence, connected words, difficult vocabularies, or noisy/stressful environments may increase error rates even more.

While the human factors literature includes research that supports certain principles of ASR system design, this knowledge has not yet been formulated as design guidelines. Human factors methodology is sufficiently developed to permit comparison of task-specific ASR systems experimentally, but does not yet have the tools required for the generation of generic ASR system design guidelines. For the near term, simulation of ASR capabilities in conjunction with the development of improved system performance measures should be a productive methodology for accomplishing this work.

4

APPLICATIONS

INTRODUCTION

For two decades, voice-interactive system (VIS) technology has been considered as a possible human interface for various applications. The ability to talk to a computer in a manner similar to talking to a human is intuitively appealing both in military and non-military settings. The ultimate voice interface will require a sophisticated mixture of voice recognition and synthesis, speech understanding, and a capability for natural language query in an application system. This capability is beyond the reach of current technology. It is important, therefore, to determine what applications show the most promise considering the current and projected future technology. A description of various VIS applications is presented here with primary emphasis on military environments.

FUNCTIONAL AREAS

For the purpose of this discussion, current VIS applications are divided into four major categories: data base management systems (DBMS), command and control of weapon systems (CCWS), training systems, and other applications. Each area uses voice as an input/output mode to allow human interaction in a specific application. The characteristics of the VIS depend upon the specific requirement of each application category. Table 3 lists typical characteristics associated with the first three categories. It is based on committee observation of data, general knowledge, and experience in the field.

Data Base Management Systems

Entry and retrieval of information from a computerized data base management system is one functional area with excellent potential for VIS applications. A major thrust in current Department of Defense

TABLE 3 Typical Voice-Interactive System Characteristics by Functional Application

	DATA BASE MANAGEMENT SYSTEM	COMMAND AND CONTROL OF WEAPON SYSTEMS	TRAINING
Vocabulary size in words	1000-5000	Less than 100	50-500
Recognizer type	Connected utterance	Discrete with connected digit capability	Discrete with connected digit capability
Speaker-dependence	Speaker-independent	Speaker-dependent	Speaker-dependent
Enrollment time	Not applicable	Varies	Less than 5% of total training time
Typical noise	Quiet	Moderate to high	Quiet
Typical operator stress	None to moderate	Moderate to high	Moderate to high
Operational requirements[a]	Less than 5% error	Less than 1% error	Less than 3% error
VIS response time	Less than 5 seconds	Less than 1 second	Less than 2 seconds
System integration requirements	Moderate	Critical	Important
Physical constraints (size, weight, power, cooling)	Minimal	Severe	Minimal

[a] There are certain safety and survivability conditions that mandate minimal error tolerance for portions of the vocabulary.

(DOD) research is in the collection of data and presentation to decision makers of timely and accurate information of current force status and the tactical situation. This communication process requires the development of computer networks and large data base management systems, along with an easy and flexible technique to enter data and retrieve information. Much of the originating data are not collected automatically but are collected by human sources. Currently, data must be entered manually into an automated system for eventual transfer to a centralized data base within the information flow network. VIS technology is being considered as a means for data entry by the person collecting the information. This would reduce by at least one, and possibly more, the links within the information network. Subsequent users of the data base can use voice systems instead of a keyboard to retrieve information. Voice access would be highly useful for DBMS applications, but the requirements for a sophisticated voice-interactive DBMS, as indicated in Table 3, are well beyond the capability of automatic speech recognition (ASR) technology in the foreseeable future.

Command and Control of Weapon Systems

Many of today's very sophisticated weapon systems are enhanced by on-board computers using multi-functional, interactive display systems. VIS can improve human-machine interaction in the management and control of weapon systems, especially when an operator is involved in tasks requiring hands and eyes to be busy.

Current research is evaluating which command and control functions can benefit most from using VIS technology. An ultimate goal is to shift tasks from an operator's hands and eyes to voice and ears, especially during times of high workload. A very high payoff potential of a VIS is the ability to provide the operator with information or functional control that is not available through traditional command, control, or input/output methods. This may be possible only in conjunction with other sophisticated automation features and functions in the context of meeting the requirements of the total system.

Training Systems

Automated training systems can use computer speech recognition and generation to simulate operational two-way voice communications. The combination of voice-interactive systems with other technologies, such as automated instruction and performance measurement, can standardize instruction and ease instructor workload. This could result in the economic benefits of reduced training staff requirements (Cotton and McCauley, 1983). To date, research on voice-interactive training systems has focused on tasks that require voice communication, such as air traffic control.

Other Applications

There have been attempts to use voice-interactive systems for voice identification and verification, keyword spotting, and other applications. Voice identification and verification is a security procedure to insure that only authorized personnel have access to information or a specific physical area. The unique characteristics of the human voice permit it to be used to identify people (Doddington, 1983).

Keyword recognition, when performed successfully by machine, can be an enormous labor-saving technique in the sorting of voice messages (Woodard and Cupples, 1983). Listening to radio broadcasts is a time-consuming, labor-intensive, tedious task for military operators. The signal quality is often poor, which increases operator fatigue and reduces effectiveness. Typically, there are a large number of voice channels of interest and only a relatively small number of operators available. To enhance the effectiveness of the limited number of operators, it becomes crucial to have the operators listen only to the most important channels. The importance of a particular channel can be specified usually by the occurrence of certain important speech in the conversation. Therefore, speech recognition could be used to detect given key words automatically. Once a prescribed number of keywords has been detected, a human operator could be alerted and switched to that channel.

CASE STUDIES

The following case studies are examples of research conducted within the functional application areas mentioned above. They serve as a basis to understand better the requirements and characteristics of voice-interactive systems. Observations from these case studies will be discussed in the next section.

Data Base Management Systems

Case Study: TACFIRE

Poock and Roland (1982b) reported on a study conducted between 1974 and 1976 that investigated the possibility of using voice-interactive systems to enter data into a tactical artillery fire control system called TACFIRE. The forward observer normally enters these data manually, but that person also has to carry a weapon and observe enemy locations and movement, at times through binoculars; both hands and eyes are busy. In addition, environmental conditions are not always conducive to manual data entry. For example, night operations and wearing protective clothing, such as gas masks and bulky gloves, make

it difficult to enter data quickly and accurately. The study ended in 1976 without a successful field demonstration of a voice-interactive system because the technology was neither sophisticated nor reliable enough to be useful in the forward observer's arduous and stressful environment.

A subsequent study investigated the implementation of VIS to the data entry process of the TACFIRE system in a much less stressful environment (Poock and Roland, 1982b). This study focused on the input and retrieval of data from the TACFIRE computer located in a van at a division artillery headquarters. The computer data base contained information on requested fire missions, available firing units, and the commander's criteria for target priorities.

The TACFIRE operator retrieved information or entered data by filling in the blank data fields on an order template. A commercially available recognizer was used to test the environmental impact on speech recognition performance. The study concluded that:

- Connected digit entry is mandatory because much of the required information is numeric in nature; for example, coordinates.
- The vocabulary size required by many message templates exceeds the 250-utterance capacity of the equipment that was used.
- The simulated system was well received by TACFIRE system designers, but the recognizer that was installed did not operate effectively because of a cumbersome hardware interface between the recognition unit and the TACFIRE computer.

Case Study: PHOTOGRAPHIC INTERPRETER

In military intelligence, photographs are viewed through magnifying stereoscopes. Collection and correlation of data from photographic interpretation and other sources are done, in part, algorithmically on computers. Thus, there is a need to enter the data into a data base management system.

Normally, photo interpreters either write the information on a record sheet for later entry into a computer system or enter the data directly using a keyboard. Both forms of data entry require looking away from the stereoscope. This is a classic eyes-hands busy task since the operators' hands are busy focusing and moving the picture, and their eyes are busy interpreting it.

One experiment examined the use of voice-interactive systems for data entry by photo interpreters (Jay, 1981). This experiment used a discrete-utterance recognizer and involved a screen viewing area within the stereoscope in order to provide visual verification of the data entered. The results showed that through the use of the voice-interactive system, the task was completed faster and more accurately than with any other tested method of data entry.

Case Study: INTEGRATED INFORMATION DISPLAY

Each major Navy Fleet Command Center can access current information on the location of friendly or known enemy ships in an operating area. Most of this information is updated automatically from inputs sent directly from U.S. ships concerning their present capability, position, and planned movements or from other automated intelligence sensors. This information can be accessed, updated manually, and displayed tabularly or graphically using the integrated information display system.

At present, VIS equipment is going through an initial testing phase in the integrated information display system (Poock and Roland, 1984). The Navy is considering its use as a tool to reduce training time for new or infrequent users. This will permit the more expert users to undertake more complicated tasks such as correlating tracks and ensuring that the information in the system is correct.

The conclusions are similar to those of the TACFIRE study, namely:

- A connected alphanumeric capability is needed to input track numbers.
- A 250-utterance vocabulary may be adequate for a limited military exercise, but more than 1000 utterances would be needed for daily command post operations.
- The hardware interface developed does not allow certain terminal operations to be implemented by voice command. This detracts from the effectiveness of the recognition equipment.

Although the system is in its initial test phase, both users and managers are enthusiastic about it.

Case Study: ELECTRONIC PARTS TRACEABILITY

Voice data entry has been used to document the source of parts in the manufacture of high-reliability electronic subsystems (Lerman, 1980). Traceability data are necessary to establish the cause of any subsequent failure in components. In a survey of the operations at Lockheed's Missile Systems Division, Lerman determined that 30 percent of operator time was spent recording traceability data. Alternative methods for entering these data into the computer were investigated, including light pencil, scanner, keyboard entry, and voice data entry. Voice data entry proved better than the other methods. It has been used by Lockheed since 1979 in a hybrid circuit assembly section and, for the last three years, in an electronic assembly area.

Lockheed reports that voice data entry has increased productivity, improved data accuracy, and reduced costs. Now, 48 operators produce the same number of circuits as 62 had originally done. The error rate for data entry has been cut from 27 percent to near zero. This

virtual elimination of errors derives not from perfect recognition accuracy, but from automatic error checking and from the requirement that operators verify the data on a visual display before entry into the data base.

Additionally, the voice entry system captures data at point of origin, frees the operator's hands for manufacturing operations, and eliminates the need for typing or keyboard skills. Lockheed reports that operator acceptance of the system is more than 90 percent. Management has determined that the economic benefits of using the voice data entry justify its use.

Case Study: MAINTENANCE AIDING

The Defense Advanced Research Project Agency (DARPA) has sponsored work on an experimental system for maintenance aiding based on voice-interaction and video disk display (Klass, 1982; Vestewig and Propst, 1982). The system was designed to allow a maintenance technician to have both hands free and unencumbered by manuals while executing maintenance procedures.

The operator wears a headset that contains a small television-display tube projected onto an eyepiece. The operator can access, via the speech recognition system, a sequence of photos showing how to perform various maintenance operations. With a limited vocabulary of approximately 20 words, the operator can, in effect, turn the pages of the maintenance manual. The experimental system has met with sufficient success that an operational version of this voice-interactive maintenance-aiding system is being considered.

Command and Control of Weapon Systems

Case Study: ADVANCED FIGHTER TECHNOLOGY INTEGRATOR (AFTI) F-16 FLIGHT TEST

The Air Force, Navy and National Aeronautics and Space Administration (NASA) sponsored a program to flight test voice-interactive systems specifically designed for the harsh environment of modern fighter aircraft. The program consisted of extensive stand-alone and integration tests of the voice recognition systems prior to flight tests. Considerable improvements in the recognition reliability rate were noted during these phases of the program. A speaker-dependent, isolated-word recognition scheme with a small (34 word) vocabulary was defined for the flight test portion. Results for these systems were outlined in Chapter 3 under the section on Performance.

Many valuable lessons learned in this program are being used in the operational utility evaluation phase presently being conducted. Aircrew attitudes concerning a voice-interaction system in the cockpit

of a modern fighter are important and can affect system performance. Improvements in the recognition reliability helped to secure aircrew acceptance during the flight tests (Moore and Ruth, 1984).

Case Study: VOICE RECOGNITION AND SYNTHESIS (VRAS)

The Naval Air Development Center sponsored the development of a sophisticated software system that used voice recognition and synthesis for training crew members who operate airborne anti-submarine monitoring stations. The ultimate goal was to determine whether voice-interactive systems could be used in actual anti-submarine monitoring operations (Stokes, 1982). The system had a complex syntactic structure, which could be reconfigured for specific applications. Although the recognition reliability was too low for successful operational use, many lessons were learned concerning the nature and capabilities of using a voice-interactive system in an operational environment.

Many human factors issues concerning system integration and the use of voice-interactive systems were addressed and found to be inadequate. For example, the language structure was too confining, and crew members had a difficult time remembering the required sequence modes. Nevertheless, voice recognition and synthesis proved a useful tool and is presently being used to develop the syntax processor in a Navy project to incorporate voice-interactive systems in the FA/18 aircraft.

Training Systems

Case Study: PRECISION APPROACH RADAR TRAINING SYSTEM AND AIR CONTROLLER EXERCISER

The Naval Training Equipment Center has sponsored development of prototype training systems for precision approach radar controllers and the air intercept controllers (Breaux, 1977; Grady, 1982). Both prototype systems demonstrated the feasibility of eliminating the need for a person to act as a pseudo-pilot, thus reducing training-support personnel. Air controller tasks are amenable to speech recognition because they involve highly structured speech as the primary output of the trainee.

These training systems were among the first to be designed around speech technology rather than retrofitted with a speech capability. In retrospect, it was perhaps unfortunate that the system development program was so ambitious, including automated instruction, automated performance measurement, adaptive syllabus control, and modelling of pilot behavior and environmental variables. All of these subsystems were directly interactive with the speech recognition subsystem and, therefore, any errors in speech recognition were amplified by

subsequent system functions (McCauley and Semple, 1980; McCauley, Root, and Muckler, 1982).

The recognition accuracy of both the Precision Approach Radar Training System (PARTS) and the Air Controller Exerciser (ACE) fell short of providing graceful interaction between trainees and the system. PARTS used a discrete utterance recognizer while ACE used a connected speech recognizer. Both were insufficiently tolerant of trainees speech variability under the stress of simulated operational conditions. Trainees are at a disadvantage in the use of speech recognition systems because they generally do not have extensive prior experience with voice communications and because they are required to learn both the new job and voice control techniques.

Nevertheless, these prototype training systems demonstrated the potential for speech interaction in real-time simulation and the power of combining it with several automated technologies. The system evaluations indicated that slight increases in recognition accuracy would lead to considerably more effective training.

Observations from Case Studies

Review of these case studies lead to some observations about VIS technology and system requirements:

- Recognition accuracy was one of the main limitations.
- The variability in human speech under stressful conditions contributed to unacceptable performance.
- The success of voice-interactive systems in most applications arose from its integration with other procedures or automation features.
- Projects designed from inception to incorporate a voice-interactive system had a greater probability of success than when the capability was added to an existing system.
- The importance of an approach that stresses human factors and integrated systems design is starting to be recognized.
- Highly connected systems that depend on accurate speech recognition for input tended to amplify the effects of recognition errors.
- A staged process of VIS development, including regular checks and tests by users, was more likely to lead to successful systems.
- Speaker enrollment (in a speaker-dependent system) was sometimes more effective when conducted in the context of the operational task.
- Lack of sufficient task analysis and design specification tended to result in a mismatch between system requirements and VIS capabilities.
- Other voice communication functions in the task environment sometimes interfered with the speech recognition task.

- The speed of command entry was not necessarily the primary measure of effectiveness when the user was engaged in simultaneous manual tasks. Performance on the primary manual task was sometimes facilitated with the use of voice on the secondary task even though the secondary task was accomplished at a slower rate.
- For externally paced tasks, the timing of the task sequence was disrupted by either long system-response time or recognition errors.
- The lack of an appropriate recognition feedback mechanism tended to confuse operators regarding the status of the system.
- The probability of system success was enhanced with the presence of a person who advocated the incorporation of the technology; i.e., a "champion."

FUTURE APPLICATIONS

If it is assumed that the capability of ASR technology will continue to grow, the future looks bright for available applications that can use it. Each functional area discussed above contains numerous examples of potential military applications that would benefit from advances in VIS technology. Some promising additional areas for future applications are discussed below.

Natural Language Retrieval

Research is continuing on the development of query systems that use a natural-language data base. A natural language interface to a data base would allow users to ask questions of the system in a manner similar to normal conversation. From the users point of view, efficiency would be increased by the addition of an ASR device capable of continuous speech recognition for a large vocabulary. This would enable verbal interchanges between users and machines to emulate interpersonal communication. In effect, the information held within the system would be available for the asking.

For example, a military commander could request specific information on the exact location and status of forces. Or, all personnel management records could be retrieved, changed and updated through voice commands. The retrieval of logistics and support capability information also would be available. And, most importantly, data could be entered into a data base as they are being gathered. To fully realize this goal, however, would be one of the most challenging applications for both natural language and speech research.

Nap-of-the-Earth Flying

Over the past 40 years military helicopter pilots have had to adjust to changing cockpit controls and displays, to new and faster airframes, and to more demanding missions, such as night, adverse weather, and nap-of-the-earth flying (i.e., following the contours of the earth at very low altitude). The military solution to these changing requirements has been more and better training for pilots. It appears that pilots, rather than the hardware, have now become the limiting factor in mission success (Voorhees and Kersteen, 1983).

Current Army plans call for a single-pilot, scout helicopter to become operational in the 1990s. While a single-pilot helicopter eliminates problems of crew coordination and makes a smaller, lighter aircraft possible, the now heavy visual and manual workload may reach intolerable levels when shifted to one pilot. In an effort to solve these problems the Army has begun to emphasize human factors research on improved designs for visual displays coupled with voice displays and voice controls. The goal of such research is to improve the information transfer rate in the cockpit (Voorhees and Kersteen, 1983).

Commercial Operations Under Instrument Flight Rules

While the Army is developing helicopters for nap-of-the-earth flight, NASA's Ames Research Center has begun to investigate the use of voice technology (both recognition and synthesis) for civilian helicopter operations that are now associated with high visual and manual workload. These include, but are not limited to, search and rescue, off-shore drilling operations, forestry, and crop dusting.

NASA's Langley Research Center has also recently embarked on a research program for general aviation aimed at applying voice controls and displays to alleviate high visual and manual workload levels of the single pilot flying under instrument flight rules (North, 1984).

Technology for Future Applications

Applications of speech technologies to many of these areas will require advances not only in speech technologies per se, but also in language processing and computerized deduction, and especially in the interaction of the three areas. That is, for many applications, the system must not only recognize the spoken sequence of words but must also recognize the structure and meaning of the utterance and in many cases deduce (rather than simply retrieve) the appropriate response to make. Such applications will need to make use of relevant results from the fields of computational linguistics and artificial intelligence, where current studies are concerned with the problems of

computer analysis of natural language as well as computerized deduction, planning, and problem solving. Advances will be required in these fields as well as in speech recognition to make such systems practical.

While not of direct interest to the sponsors of this report, ASR technology also has broad potential application in aids for the physically handicapped. Numerous projects using current technology are regularly featured in the publication Communication Outlook.

5

DIRECTIONS FOR FUTURE WORK

Progress in the development and use of speech input/output (I/O) technology requires a commitment to research in both speech recognition and system integration. This connection can be achieved by identifying critical shortcomings of current technology and fostering efforts to overcome them. In this process, it is important to maintain a perspective on application needs. For example, both connected speech and speaker independence are key capabilities that must be developed to meet the demands of the applications discussed in the previous chapter.

AUTOMATIC SPEECH RECOGNITION TECHNOLOGY

Two paths can be followed in advancing speech recognition technology. One is to improve and enhance current techniques and algorithms. The other is to pursue new algorithms and approaches. Near-term applications may benefit from improvements to current recognition technology. Breakthroughs in capabilities, or large advances in recognition performance, will require new approaches to solving speech recognition problems.

Although the distinction between "new algorithms" and "enhancements to current algorithms" is not obvious, for the purposes of this discussion, current technology includes the components depicted previously in Figure 7. This architecture uses signal-processing features, followed by a comparator that makes whole-utterance assessments of the incoming features, followed by a higher level decision process. New algorithms and approaches generally depart in substantial ways from the architecture shown in Figure 7.

As its most important priority, speech recognition research should focus on identifying and accommodating various sources of variability in the speech signal. Current automatic speech recognition (ASR) technology often performs well under laboratory conditions or in some

field tests but, then, exhibits far poorer performance in other comparable tests. In the past, work has often been directed more toward controlling the environment, talker, and speech signal than toward developing algorithms that accommodate normal speech signal variations. This will have to change if there are to be significant advances in ASR technology.

The most common example of controlling the speech signal is the requirement that speech recognition be done in a speaker-dependent mode. Better performance is obviously achieved this way. But when this restriction is carried into the laboratory, research is frequently diverted from this important source of speech signal variation and from the opportunity to improve robustness. In fact, the study of speech recognition independent of speaker offers a rich opportunity to make technological advances that may improve performance under other degrading conditions such as stress and noise. Productive studies of speaker-independent recognition may be expected to improve speaker-dependent recognition as well.

Extension of Current ASR Technology

As noted in Chapter 3, current ASR technology is generally agreed to be fragile; i.e., recognizer performance is susceptible to significant degradation under the effects of acoustic noise, user stress, and operational distractions. Thus, efforts to extend current technology need to focus on improving the robustness of algorithms to these degrading factors. Additionally, there is a need to study and develop speech features and comparison techniques that are less sensitive to noise and to uncontrolled changes.

Some of the common acoustic variations that combine to degrade recognition performance include those due to differences in vocal effort and in allophonic variations. These variations might be better accommodated through algorithms that normalize gross spectral differences, which occur with changes in vocal effort, and through improved end-point detection. Alternatively, they can be achieved through algorithms that do not require end-point locations. However, even with concentrated effort on enhancing current recognition technology, the current limitations of small vocabulary, speaker dependence, and isolated utterance probably will not be substantially relaxed.

Other improvements in system performance may be gained through attention to and control of the acoustic environment, the user, the task environment, and the application host. Such improvements will, however, be limited by the performance of ASR devices themselves.

To make meaningful advances, development efforts will have to be performed in the context of carefully defined speech data bases. Such data bases should represent the important sources of speech signal variation in an economical manner. A tradeoff between adequacy of

representation (for example, number of speakers, size of vocabulary, levels of stress/noise) and size of the data base is necessary because manageability is an important attribute to acceptance and use.

An important objective is to elicit and represent realistic variations in speech by a given speaker. This might be achieved through explicit variations in level of speech effort or through natural reactions to imposed conditions such as noise, task stress, or G-forces. Availability of such data bases would enable systematic quantification of progress in algorithm development.

Directions for Advanced ASR

Specific research goals for advanced ASR can be suggested by viewing the speech recognition process as a hierarchy of information representations, including phonetic units, syllables, words, and sentences. Available technology for adequately modeling this process is currently insufficient or nonexistent, even at the lowest levels of phonetic representation.

Promising directions for advanced research can be divided into two areas; namely, research into phonetic representation (low-level perceptual units) and research on higher levels of representation. Because they are interrelated, coordination of research in the two areas is important.

Specific tasks and guidelines for meaningful research should include speaker-independent recognition and connected speech. Only through successful approaches that encompass these inherent characteristics of speech can progress be made toward recognition capabilities that approach human performance. Also, advanced speech recognition technology must build on an understanding of the causes for the fragility and limited capabilities of present ASR technology. For example, the deleterious effects of the acoustic environment on the signal must be systematically quantified. Techniques are needed to measure speech degradation due to physiological and psychological stress. Above all, research should identify those aspects of the speech signal that convey linguistic information as opposed to those that convey talker and environmental information. In other words, fundamental research needs to be conducted on the acoustic manifestations of phonetic contrasts and on the discovery of parameters and rules that describe such contrasts.

As previously indicated, current speech recognition systems focus largely on recognizing independent utterances (usually words), with little dependence on context and little internal syntax. However, for most applications, speech recognition takes place in an ongoing sequence of interactions that reflect syntactic, semantic, and even pragmatic constraints. In applications where the desired input utterances are meaningful English sentences, problems of discourse

structure and sentence structure become important. Ideally, a speech recognition system should be able to perform as a full conversational partner.

The development of these capabilities will likely require coordination with work in computational linguistics and artificial intelligence that is currently addressing problems of dialog structure and the theory of communicative actions. More exploration is needed on models of grammar and semantics, and on the influence of the goals and objectives of speakers. Several relations are poorly understood. These include the interactions between the prosodics of an utterance, its syntactic structure, and the communicative intent of the speaker. If speech recognition systems are to support conversational interactions, research should be conducted into these related aspects.

The establishment of research and evaluation data bases is even more important to the success of advanced ASR research than to the extension of current technology. These data bases are needed to focus effort on key recognition problems, as well as to measure progress. Important data base features should address speaker independence and continuous speech, as well as variables such as stress, acoustic noise, and transmission channel characteristics.

Progress in advanced systems can be gauged against human performance in speech recognition. Ideally, the recognition technology should perform as well as humans who are limited to the same information. Although recognition accuracy requirements for a specific task environment may differ substantially from human capabilities, humans provide the existence proof that speech can be understood with high accuracy under a variety of degraded conditions, including stressful ones. Quantification of human recognition performance can provide a reference for evaluating machine performance. For example, a recent experiment on human recognition of spoken digits suggests that high performance (0.01 percent error rate) is possible with existing (LPC) parameteric representations of speech (Leonard, 1984); however, recognition performance requirements for a specific task environment may differ substantially from human performance capabilities.

The importance of advances in VLSI and computers to progress in speech recognition cannot be overstated. Speech recognition technology depends on computer technology for research as well as implementation. Only through the availability of significantly increased computing capability is it likely that major advances in automatic speech recognition will be made. This includes order-of-magnitude increases in computer power, coupled with similar increases in memory capacity and decreases in costs. Improvements in laboratory computing facilities are important because the ability to conduct appropriate research depends strongly on the existence of adequate computer support.

HUMAN FACTORS INTEGRATION

If the benefits of speech technology are to be realized, a significant effort is needed in human factors integration. This includes task selection, determination of performance requirements, speech display design, environmental effects, and performance assessment. Human factors research is needed to develop procedures for selecting appropriate tasks for the voice mode and for integrating voice controls and displays into the total system design. There is no single area that can be chosen for particular emphasis. However, all of the research should emphasize the total system in which speech recognition and generation are to be incorporated.

Task Selection

Automatic speech recognition can more appropriately be applied for some tasks than for others. Criteria include the extent of workload reduction, increase in throughput, and/or the balance of costs and benefits. Speech is not an attractive substitute for manual data entry when the latter is being performed successfully unless cost savings can be realized as in the Lerman (1980) project described in Chapter 4. Voice input is likely to improve system throughput only in complex tasks involving high cognitive, visual, and manual loading. More work is needed on task/modality compatibility to support decisions about selecting appropriate tasks for speech interactive systems.

Purely analytic procedures such as user questionnaires and task analyses are not likely to be sufficiently accurate to enable detailed specification of speech system requirements. Simulation techniques hold promise in this regard and may help to establish the speech system requirements early in the system design process.

Speech Subsystem Performance Requirements

Human performance using speech controls and displays can be assessed through laboratory simulation of ASR hardware. This strategy enables various levels of speech recognition capability to be controlled and evaluated experimentally. Important issues to be addressed include the following: speed and accuracy required, criticality of errors by type, appropriate forms of error correction, the need for speaker independence and connected or continuous speech, the effects of vocabulary size, and human abilities to constrain speech in terms of vocabulary, syntax, and speaking patterns.

System performance measures are needed that integrate recognizer, human, and system performance, task workload, system utility, and user acceptance. Data from simulations can be used to identify candidate

tasks for speech and to assess the level of performance required for successful use of speech data bases. Simulations also can provide samples of speech produced under various task conditions, such as noise, mental workload, stress, and various levels of recognition error rate. Finally, simulation provides a research environment for developing general guidelines on how speech data entry should be integrated into different task environments.

Human Factors Integration to Incorporate the Speech Modality

Task Design

Successful ASR performance for a particular task will not guarantee that the total system will perform successfully. Basic limitations of human memory and information processing should be considered in the design of any human-machine interface. Many of the problems of today's complex control and display systems may be solved by better design of the overall system. The possible role of speech controls and displays in those solutions can be determined only after considerable analysis or simulation to compare speech with alternative modes of control and display.

Certain unique features of the speech mode preclude a one-to-one mapping of individual manual controls to speech controls and of visual display elements to speech display messages (Cotton and McCauley, 1983; Simpson, in press-b). Further, certain features of speech constrain the way speech can be used in human-machine systems. Speech may not always provide the most rapid means of interacting with the system. The time required for an operator to execute a voice command is strongly influenced by variables such as vocabulary selection, input syntax, and dialogue design.

To minimize human cognitive load and the time to issue speech commands, the number of words required to complete each command should be small. Thus, more basic information is needed on human memory for constrained verbal material and to determine the effects on system performance of placing such constraints on human users. Information about the effects of harsh environments and stress on verbal versus motor memory would be particularly relevant. Research in this area should lead to guidelines for establishing recognition vocabularies that are flexible, easy to remember, low in acoustic confusion, and that avoid awkward speech styles.

Human-System Dialogue Design

Design of all the interchanges between the human and the system, not just the speech interchanges, have major effects on overall system performance.There are at least two subsets of dialogue design--the dialogue between the users and the ASR system, and dialogue between users and all systems under their control.

The human-machine dialogue should be designed with regard to the total set of control and display options for all systems. Mission scenarios will have to be analyzed for speech and other audio loads, and the likelihood of concurrent interfering speech messages. The potential functions to be controlled by speech should be assessed, along with the priorities of all speech messages within the system. Voice commands and displays should be applied in ways that complement rather than conflict with other controls and displays. Future research and development efforts should address these issues.

To improve system performance and throughput with ASR, dialogue should be designed to facilitate rapid information transfer between human and machine and to minimize both the potential for error and the time required for error correction by users. Not only speech commands, but also features such as prompts, system feedback, and system responses to user queries should be carefully designed and a timeline of the total dialogue evaluated. Experiments are needed to determine the desired type and amount of linguistic redundancy for particular applications. Syntax design should be viewed as an integral part of ASR system design, rather than simply a technique for improving the performance of a marginal system.

Error rates possibly can be reduced if system designers provide aids to users such as tonal prompts for cadence, menus of acceptable entries, consistent feedback, and convenient error correction commands. The best format for these dialogue elements can only be empirically determined.

Additional research is needed to evaluate techniques to capitalize upon syntactic and semantic constraints in the dialogue. Applications software using these techniques should improve recognition accuracy and reduce the user's burden of detecting and correcting recognition errors.

User Characteristics and Training

Better methods are needed for predicting the ASR performance to be expected on the basis of user characteristics. For example, the user's dialect may influence recognition performance. Some talkers consistently demonstrate better speech recognition performance than others. Research is needed on techniques for predicting these differences and on potential methods for remediating low-performance users.

Training users to modify their speech patterns will be difficult because speech is a highly overlearned behavior. The extent to which training can reliably alter speaking habits, particularly under stressful conditions, has yet to be determined. This is an important research area for the types of applications envisioned by the Department of Defense.

Enrollment

If speaker-dependent ASR technology is to be used in a hostile environment, better methods are needed that would permit enrollment in a benign environment. The fatigue and stress that can be induced by enrollment under unfavorable conditions are the major reasons for this requirement. The cost of operating expensive equipment merely for purposes of enrolling the speech recognition system is another important factor. These problems might be reduced if future developments in recognition systems include automatic updating of talker samples.

Speech Display Design

Speech displays, or synthetic voice output from machines, have been used extensively for voice announcements, alarms, and information access. With advances in speech recognition, speech displays take on an expanded role, enabling machines to conduct meaningful interactive voice exchanges with humans. In conjunction with recognition, speech displays (or synthetic voice responses) can variously provide confirmatory feedback, response to human queries, system's status announcements and spoken instructions, or prompts to human operators.

Design issues for speech displays include: selection of voice type, message wording, and syntax for specific tasks. Intonation, speaking rate and other prosodic features can convey information that is additional to the written equivalent of the synthesized message. The effective use of these signal dimensions is not well established, and requires fundamental research.

Because a variety of functionally different speech messages might be used in a given system, a method of assigning priorities must be incorporated in the system design, as well as means for handling concurrent messages. Two special cases of the latter can be mentioned: (1) a human operator may be speaking to the recognition device while the speech display is enunciating a message, and (2) more than one speech message may be triggered at a time (competing messages). These human factors issues represent a layer of design that is separate from the design of speech-generating algorithms and the computational techniques for message synthesis. In the latter arena, research goals center on improving methods for synthesizing voice messages directly from printed text, without dependence upon human recordings, signal analysis, or hand editing.

Task Environment

The task environment comprises a number of factors that need to be studied for their effect on human performance and, therefore, on

speech task design. Physical, physiological, emotional, and workload factors can be expected to contribute to the success or failure of a particular speech system design. Only after the effects of these factors are known can speech systems be designed in ways that will enhance rather than hinder human performance and, thus, system performance.

Performance Assessment

ASR performance should be measured within a realistic task scenario, both within the laboratory and in actual operational settings, including worst-case conditions. Laboratory tests using standard vocabularies, experienced users, and controlled environments, are useful for comparing recognizers, but they are not sufficient for predicting actual performance in operational systems. Methods are needed for measuring both human and recognizer performance under realistic conditions. The importance of performance measurement techniques cannot be overemphasized since they provide the data to be used in making decisions about system design and effectiveness.

Candidate measures of system performance include the number of missions successfully accomplished, time to accomplish the mission, the losses involved, productivity increases, cost savings, and reductions in operator workload. Operator workload is an important measure because it can be used to compare alternative system designs. Currently, there is no single, reliable method for assessing human workload in a variety of tasks (Wierwille and Connor, 1983). Although some research is ongoing in this area, an emphasis on this topic would be valuable, not only for ASR applications, but for many other problems in human-system interface design.

Summary of Research Needs in Human Factors Integration

This section has given a perspective on directions for research and development in the human factors integration of speech-interactive systems. The major areas of concern were described as task selection, human factors integration, dialogue design, user characteristics, speech display design, task environment, and performance assessment. Emphasis has been given to the concept that an interactive speech application involves far more than a speech recognizer. It has been suggested that many previous applications of speech have failed, in part at least, because of the inadequate attention to human factors integration. Because of its importance, considerable progress is needed in developing guidelines for accomplishing human factors design. The research goals and procedures described here aim to support the development of essential human factors techniques and guidelines.

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

Based on its study and prior experience, the committee reached the following conclusions:

- The use of speech for communication between humans and machines (automatic recognition for input, automatic synthesis for output) has distinct potential for aiding humans in the acquisition, organization, and processing of information.
- Current technology for automatic speech recognition--including algorithm development, hardware design, and human factors integration--is not sufficiently advanced to achieve robust, reliable performance in hostile and high-stress environments.
- Current technology is not sufficiently advanced to achieve high performance in applications where large vocabularies and/or continuous spoken input are needed, and where the ability to understand speech from a wide variety of speakers is required, even in benign environments.
- Current technology is, however, mature enough to support restricted labor-saving applications in benign environments, with disciplined use under low-stress conditions. The success of these and of future applications depends on the integration of speech recognition with related automation techniques.
- With the exception ot several experimental data bases, no systematic, standardized techniques exist for evaluating and comparing the performance of speech recognizers.
- No established human-factors methodologies exist that specifically differentiate the benefits of speech input from related automation techniques. Neither are there methods that reveal the optimum human-machine architecture for integrated voice systems or set requisite performance levels for speech recognizers embedded in prescribed operational tasks.

- There is insufficient fundamental understanding of the speech degradation produced by environmental acoustic noise and physiological and psychological stress to design reliable recognition algorithms and to predict their performance over a range of adverse conditions.
- Government-sponsored efforts in speech recognition are currently fragmented, directed toward short-range goals, and insufficient in level and duration to sustain major technology advances. To a large extent, existing efforts are ancillary to non-speech projects and make use of instrumentation dictated by expediency and low cost.
- The development and evaluation of speech recognition algorithms will probably require sophisticated computational resources that are not widely available today.
- Successful and cost-effective field deployment of advanced speech recognition systems will be directly related to and in part dependent on continued advances in integrated circuit technology and computer design.
- Speech synthesis, providing automatic voice response for interactive systems, is an important adjunct to automatic speech recognition.
- No central focus exists in the U.S. government to manage research and development in speech recognition, to set policy, to establish goals, or to allocate funding.

RECOMMENDATIONS

The conclusions of the preceding section lead to corollary recommendations. These recommendations are intended to achieve a speech recognition technology that can provide the utility, accuracy, and reliability required for operations in benign as well as severe environments. The conclusions suggest the necessity of a strategic, coordinated, long-range program of research and development, with a substantial initial commitment over a period of several years. The constituents of the projected program are intended to respond to the major gaps in fundamental scientific understanding.

The Speech Signal

A basic research program is needed to characterize speech and its variabilities. Such a program would include the study and documentation of the acoustic properties of different sounds in various contexts, extending to sentences; the exploration of alternative acoustic representations and features; the development of quantitative models and methods for describing speech events and their variabilities; and the investigation of methods for modeling talker differences.

The deleterious effects of severe environments upon the signal produced by otherwise-normal talkers must be systematically quantified. Specifically, techniques for measurement and modeling of the degradations and variabilities introduced by additive acoustic noise, obstructions to articulation (such as oxygen masks), and physiological and psychological stress need to be established. Ancillary insight and guidelines would derive from measurements on the ability of human listeners to perceive and recognize speech that is degraded by the same processes.

Transducers and Signal Enhancement

Because an automatic speech recognizer's performance is limited by the information delivered to it, new methods should be sought and studied for sound transduction (including microphone systems designed for severe environments) and for electronic signal enhancement. The common aim is to provide the most favorable signal subject to environmental distortions. This includes physical control of the environment, such as sound isolation enclosures and noise control measures with active and passive sound absorption.

Algorithm Design

Significant research efforts are required in the design of algorithms and systems for the recognition of continuous speech in complex application domains, for speaker-independent operation, and for robust performance under conditions of degraded inputs. Alternative recognition strategies and algorithms must be explored to achieve the high performance necessary for many applications. Means for the dynamic adaptation of recognition algorithms to changes in talker, environment, and application domain should be investigated.

A specific effort should characterize the performance of generic algorithms as a function of input degradations. This knowledge would facilitate algorithm design for specific environments, and would allow prediction of performance over a range of applications.

Advanced applications of automatic speech recognition might benefit from incorporating capabilities in natural language processing and computerized deduction. Research advances in these areas should, therefore, be integrated with on-going speech efforts.

While fundamental algorithm research is best done without constraints on processor power and cost (that is, with the sole objective to establish and advance the limits of feasibility), related knowledge on performance as a function of limiting algorithm complexity to achieve low-cost implementation is of direct interest.

Human Factors Integration

A "total systems" approach is needed to decide, for specific applications, whether the available technology can supply a benefit and, for promising applications, to obtain the maximum benefit from existing speech technology. Speed, accuracy, reliability, compatibility with concurrent tasks, consequences of errors, and operator workload are obvious factors in systems design and integration.

Research is necessary to establish methodologies for analyzing human-machine communication tasks. The methodologies should identify and quantify the benefits that speech input/output can contribute compared with competing methods for automating the same task, such as manual keyboard entry and visual alphanumeric displays. They should also establish performance criteria for generic classes of tasks. Techniques for simulating speech system characteristics should be developed to assess design alternatives.

Because research results related to voice-interactive systems are largely scattered throughout the literature and are not readily available in standard references, human-factors design guidelines for voice-interactive systems should be generated.

Hardware Development and Deployment

A major conclusion from the preceding section is that current speech recognition technology is not sufficiently advanced to justify widespread field deployment in severe environments. Extensive hardware development and deployment, based upon existing technology, is, therefore, inappropriate.

Exploratory hardware applications efforts, however, are vital for gaining practical knowledge about applications and for establishing the limitations of existing technology, even though such hardware designs tend to be customized, or task-specific, and can not be easily generalized.

Exploratory hardware might be employed, especially in low-risk, low-stress situations, where routine labor-intensive or fatigue-producing applications could benefit from the existing technology, despite the limitations.

Modest efforts in exploratory hardware are already in progress and should continue with increased support, but the emphasis at present should be to enlarge the fundamental scientific understanding of speech recognition.

Evaluation and Comparison of Recognizer Performance

Standardization is needed to quantify the performance of automatic speech recognizers and to permit comparisons across algorithm philosophies, applications environments, and classes of talkers.

A significant step in this direction would be formulation of a common data base, made broadly available, together with prescribed procedures for assessing performance. A central responsibility is recommended for maintaining and distributing the data base, and for mediating comparison tests. Additionally, this capability would facilitate prediction of recognizer performance under worst-case field conditions.

Computational Capabilities

If significant advances in algorithm design are to be made, projected support should extend to financing the most sophisticated laboratory computation available. Continuous speech recognition algorithms are estimated to require computational capability in the range 10^2 to 10^4 Mips. Machines with this capability are virtually non-existent as yet.

For this same reason, advanced efforts in speech recognition should be coordinated with related government work on high-speed processors and strategic computing. The practicability of deploying sophisticated recognizers in the field will depend directly upon the benefits of huge amounts of low-cost computation. These benefits can be expected from continued advances in VLSI techniques and in architectures for parallel computing.

Design and Administration of a Research Program

The need has been stated for substantial, sustained support of research and development to realize the potential of speech recognition for military and other government users. A viable program first requires a focus of responsibility and accountability, and a means for coordinating different interests and needs across a large variety of organizations. One possibility is an inter-agency consortium, similar to that implemented in the Department of Defense for narrowband speech coding. Another possibility is for an appropriate organization to take the lead, as in the earlier speech understanding research program of the Advanced Research Projects Agency (ARPA).

In whatever way it is administered, there are several vital attributes to a suitable program. Foremost is a commitment to long-range research goals, where the acquisition of fundamental knowledge is paramount. The development activity should be managed to capitalize continually on applications that the evolving and improving technology can support. The program should be recognized as a bonafide speech activity, supported for its own merits, and not merely carried along as ancillary to some more substantial undertaking.

REFERENCES

Aretz, A. J. 1983. A comparison of manual and vocal response modes for the control of aircraft systems, Proceedings of the Human Factors Society 27th Annual Meeting, Norfolk, Va., October 10-14, 1983. Santa Monica, Calif.: Human Factors Society.

Armstrong, J. W. and G. K. Poock. 1981a. Effect of Operator Mental Loading on Voice Recognition System Performance. Technical Report NPS55-81-016. Monterey, Calif.: Naval Postgraduate School.

Armstrong, J. W. and G. K. Poock. 1981b. Effect of Task Duration on Voice Recognition System Performance. Technical Report NPS55-81-017. Monterey, Calif.: Naval Postgraduate School.

Bahl, L. R., S. K. Das, P. V. de Souza, F. Jelinek, S. Kata, R. L. Mercer, and M. A. Picheny. 1984. Some experiments with large-vocabulary isolated-word sentence recognition. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, San Diego, Calif., Mar. 19-21, 1984.

Bahl, L. R., F. Jelinek, and R. L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 5(2):179-190.

Baker, J. K., J. M. Baker, R. Roth, and P. G. Bamberg. 1984. Cost-effective speech processing. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, San Diego, Calif., Mar. 19-21, 1984.

Bamberg, P. G., L. G. Bahler, J. M. Baker, and H. G. Kellet. 1981. Gisting Technique Development. Technical Report RADC-TR-81-355. Griffiss Air Force Base, N.Y. Rome Air Development Center.

Beek, B., E. P. Neuburg, and D. C. Hodge. 1977. An assessment of the technology of automatic speech recognition for military applications. IEEE Trans. Acoustics, Speech, and Signal Processing 25(4):310-322.

Bell, D. W. 1982. Experimental design considerations for evaluating voice I/O. Pp. 215-221 in Proceedings of the Workshop on Standardization for Speech I/O Technology, D. S. Pallett, ed. Washington, D.C.: National Bureau of Standards.

Breaux, R. 1977. Laboratory demonstration of computer speech recognition in training. In Voice Technology of Interactive Real-Time Command/Control Systems Applications, R. Breaux, M. Curran, and E. M. Huff eds. Moffett Field, Calif.: NASA-Ames Research Center.

Butler, F., E. Manaker, and W. Obert-Thorn. 1981. Investigation of a Voice Synthesis System for the F-14 Aircraft: Final Report, Contract No. N62269-79-C-0493, MOD P00001. Prepared for the Naval Air Development Center by Grumman Aerospace Corp., Report No. ACT 81-001.

Chapanis, A. 1975. Interactive human communication. Scientific American 232(3):36-42.

Chomsky, N. and M. Halle. 1968. The Sound Pattern of English. New York: Harper and Row.

Cole, R. A., A. I. Rudnicky, V. W. Zue and D. R. Reddy. 1980. Speech as patterns on paper, Ch. 1 in Perception and Production of Fluent Speech, R. A. Cole, ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cole, R. A., R. M. Stern, M. S. Phillips, S. M. Brill, A. P. Pilant and P. Specker. 1983. Feature-based speaker-independent recognition of isolated English letters. Pp 731-733 in Proc. IEEE. Int. Conf. Acoustics, Speech, and Signal Processing, Boston, Mass., April 1983.

Coler, C. 1982. Helicopter speech-command systems: recent noise tests are encouraging. Speech Technology 1(3):76-81.

Communication Outlook. East Lansing, Mich.: Artificial Intelligence Laboratory, Department of Computer Science, Michigan State University.

Cotton, J. C. and M. E. McCauley. 1983. Voice Technology Design Guides for Navy Training Systems. NAVTRAEQUIPCEN 80-C-0057-1. Orlando, Fla.: Naval Training Equipment Center.

Cupples, E. J. 1984. Speech enhancement for improved recognition. Paper presented at 5th Western Conference and Exposition. Armed Forces Communications and Electronics Association, Anaheim, Calif., 31 Jan.- 2 Feb., 1984.

Department of Defense. 1981. Human Engineering Design Criteria for Military Systems, Equipment and Facilities (MIL-STD-1472C). Washington, D.C.: Department of Defense.

Dixon, N. R. and C. C. Tappert. 1973. Intermediate Performance Evaluation of a Multi-Stage System for Automatic Recognition of Continuous Speech. Technical Report TR-73-16. Rome, N.Y.: Rome Air Development Center.

Dixon, N. R. and T. B. Martin, eds. 1979. Automatic Speech & Speaker Recognition. New York: IEEE Press.

Doddington, G. R. 1983. Voice authentication gets the go-ahead for security systems. Speech Technology 2(1):14-23.

Doddington, G. R. and T. B. Schalk. 1981. Speech recognition: turning theory to practice. IEEE Spectrum 18(9):26-32.

Dunn, H. K. and S. D. White. 1940. Statistical measurements on conversational speech. J. Acoust. Soc. Amer. 11(1):278-288.

Fant, G. C. M. 1960. Acoustic Theory of Speech Production. The Hague, Netherlands: Mouton Co.

Flanagan, J. L. 1972. Speech Analysis, Synthesis and Perception. Second Edition. New York: Springer Verlag.

French, N. R. and J. C. Steinberg. 1946. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Amer. 19(1):90-119.

Golibersuch, R. J. 1983. Automatic prediction of linear frequency warp for speech recognition. Pp. 769-772 in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Boston, Mass., April, 1983.

Goodman, R. G., 1976. Analysis of Languages for Man-Machine Voice Communication. Technical Report. Pittsburgh, Pa.: Computer Science Dept., Carnegie-Mellon University.

Gould, J. D., J. Conti, and J. Hovanyecz. 1983. Composing letters with a simulated listening typewriter. Pp. 295-308 in Communications of the ACM, April, 1983.

Grady, M. W. 1982. Air intercept controller prototype training system. NAVTRAEQUIPCEN 78-C-0182-14. Orlando, Fla.: Naval Training Equipment Center.

Hayes-Roth, F. 1980. Syntax, semantics, and pragmatics in speech understanding systems. Pp. 206-233 in Trends in Speech Recognition, W. A. Lea, ed. Englewood Cliffs, N.J.: Prentice Hall.

Hecker, M. H. L., K. N. Stevens, G. Von Bismarck, and C. E. Williams. 1968. Manifestations of task-induced stress in the acoustical speech signal. J. Acoust. Soc. Amer. 44:993-1001.

Huttar, G. L. 1968. Relations between prosodic variables and emotions in normal American English utterances. Journal of Speech and Hearing Research 11:481-487.

Itakura, F. 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoustics, Speech, and Signal Processing 23:67-72.

Jay, G. T. 1981. An experiment in voice data entry for imagery interpretation reporting. Masters thesis. Naval Postgraduate School, Monterey, Calif.

Kaneko, T. and N. R. Dixon. 1983. An hierarchical-decision approach for large-vocabulary, discrete utterance recognition. IEEE Trans. Acoustics, Speech, and Signal Processing 31:1061-1066.

Kato, Y. 1980. Words into action iii: a commercial system. IEEE Spectrum 17(6):29-30.

Kersteen, Z. A. 1982. An evaluation of automatic speech recognition under three ambient noise levels. Pp. 63-68 in Proceedings of the Workshop on Standardization for Speech I/O Technology, D. S. Pallett, ed. Washington, D.C.: National Bureau of Standards.

Klass, P. J. 1982. Technique benefits noise technicians. Aviation Week and Space Technology, October 11, 1982, pp. 133-135.

Klatt, D. H. 1977. Review of the ARPA speech understanding project. J. Acoust. Soc. Amer. 62:1345-1366.

Kryter, K. D. 1962. Methods for the calculation and use of the articulation index. J. Acoust. Soc. Amer. 34:1689-1697.

Kryter, K. D. 1970. The Effects of Noise on Man. New York: Academic Press.

Kuroda, I., O. Fujiwara, N. Okamura, and N. Utsuki. 1976. Method for determining pilot stress through analysis of voice communication. Aviation, Space, and Environmental Medicine 47:528-533.

Lea, W. A., ed. 1980. Trends in Speech Recognition. Englewood Cliffs, N.J.: Prentice Hall.

Lea, W. A. and J. E. Shoup. 1980. Specific contributions to the ARPA SUR project. Pp. 382-421 in Trends in Speech Recognition, W. A. Lea, ed. Englewood Cliffs, N.J.: Prentice Hall.

Leonard, R. G. 1984. A data base for speaker-independent digit recognition. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, San Diego, Calif., Mar. 19-21, 1984.

Lerman, L. 1980. Traceability via voice data entry in hybrid manufacturing at Lockheed Missile Systems Division. Pp. 97-109 in Voice Interactive Systems: Applications and Payoffs, S. Harris, ed. Warminster, Penn.: Naval Air Development Center.

Levinson, S. E. and A. E. Rosenberg. 1978. Some experiments with a syntax directed speech recognition system. Pp. 700-703 in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing. Tulsa, Okla., April 10-12, 1978.

Lieberman, P. 1961. Perturbations in vocal pitch. J. Acoust. Soc. Amer. 33:597-603.

Lim, J. A., ed. 1983. Speech Enhancement. Englewood Cliffs, N.J.: Prentice Hall.

Lowerre, B. and R. Reddy. 1980. The HARPY speech understanding system. Pp. 340-360 in Trends in Speech Recognition, W. A. Lea, ed. Englewood Cliffs, N.J.: Prentice Hall.

McCauley, M. E. 1984. Human factors in voice technology. In Human Factors Review 1984, F. A. Muckler, ed. Santa Monica, Calif.: The Human Factors Society.

McCauley, M. E. and C. A. Semple. 1980. Precision Approach Radar Training System (PARTS). NAVTRAEQUIPCEN 79-C-0042-1. Orlando, Fla.: Naval Training Equipment Center.

McCauley, M. E., R. W. Root, and F. A. Muckler. 1982. Training Evaluation of an Automated Air Intercept Controller Training System. NAVTRAEQUIPCEN 81-C-0055-1. Orlando, Fla.: Naval Training Equipment Center.

McCullough, D. P. 1983. Secondary testing techniques for word recognition in continuous speech. Pp. 300-303 in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Boston, Mass., April 1983.

McSorley, W. J. 1981. Using Voice Recognition Equipment to Run the Warfare Environmental Simulator (WES). Masters thesis. Monterey, Calif.: U.S. Naval Post Graduate School.

Miller, G. A., G. A. Heise, and W. Lichten. 1951. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology 41:329-335.

Meyers, C. S. and L. R. Rabiner. 1981. Connected digit recognition using a level-building DTW algorithm. IEEE Trans. Acoustics, Speech, and Signal Processing 29:351-363.

Mountford, S. J., J. Schwartz, and K. Graffunder. 1983. Evaluation of Speech Technology for Automatic Target Recognition. Proceedings of the Human Factors Society 27th Annual Meeting, Norfolk, Va. October 10-14, 1983. Santa Monica, Calif.: Human Factors Society.

Moore, C. A. and J. C. Ruth. 1984. Use of voice integrated with aircraft cockpit displays. Conference Digest of the 1984 International Symposium, Seminar, and Exhibition, San Francisco, June 4-8, 1984. New York: Society for Information Display.

Neben, G., R. J. McAulay, and C. J. Weinstein. 1983. Experiments in isolated word recognition using noisy speech. Pp. 1156-1159 in Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Boston, Mass., April 1983.

Newell, A., J. Barnett, J. W. Forgie, C. Green, D. Klatt, J. C. R. Licklider, J. Munson, D. R. Reddy and W. A. Woods. 1973. Speech Understanding Systems: Final Report of a Study Group. Amsterdam: North Holland/American Elsevier.

Ney, H. 1984. The use of a one-stage dynamic programming algorithm for connected word recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing 32(2):263-271.

North, R. A. 1984. Application of Speech Technology in Single Pilot Instrument Flight Rules Aircraft. Report prepared by Honeywell, Inc. for NASA Langley Research Center under NASA contract NAS1-17053. Minneapolis, Minn.: Honeywell, Inc.

Ogden, G. D., J. M. Levine, and E. J. Eisner. 1979. Measurement of workload by secondary tasks. Human Factors 21:529-548.

Pallett, D. ed. 1982. Proceedings of the Workshop on Standardization for Speech I/O Technology. National Bureau of Standards. Washington, D. C., March 18-19, 1982.

Poock, G. K. 1981a. To train randomly or all at once...that is the question. Proc. of the Voice Data Entry Systems Applications Conf., Santa Clara, Calif., October 7-8, 1981.

Poock, G. K. 1981b. A longitudinal study of computer voice recognition performance and vocabulary size. NPS55-81-013. Monterey, Calif.: Naval Postgraduate School.

Poock, G. K., B. J. Martin, and E. F. Roland. 1983. The effect of feedback to users of voice recognition equipment. NPS55-83-003. Monterey, Calif.: U.S. Naval Postgraduate School.

Poock, G. K. and E. F. Roland. 1982a. Voice Recognition Accuracy: What is Acceptable? NPS55-82-020. Monterey, Calif.: U.S. Naval Postgraduate School.

Poock, G. K. and E. F. Roland. 1982b. Preliminary Conclusions on the Use of Voice Recognition Input to TACFIRE. NPS55-029PR. Monterey, Calif.: U.S. Naval Postgraduate School.

Poock, G. K. and E. F. Roland. 1984. A Feasibility Study for Integrated Voice Recognition Input into the Integrated Information Display System (IID). NPS55-84-008-PR. Monterey, Calif.: U.S. Naval Postgraduate School.

Rabiner, L. R. and S. E. Levinson. 1981. Isolated and connected word recognition--Theory and selected applications. IEEE Trans. on Communications 29(5):621-659.

Rabiner, L. R., S. E. Levinson, A. E. Rosenberg and J. G. Wilpon. 1979. Speaker-independent recognition of isolated words using clustering techniques. IEEE Trans. on Acoustics, Speech, and Signal Processing 27(4):336-349.

Sakoe, H. and S. Chiba. 1971. A dynamic programming optimization for spoken word recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing 26:43-49.

Schurick, J. M., B. H. Williges and J. F. Maynard. In press. User feedback requirements with automatic speech recognition. Ergonomics.

Schwartz, R. M. 1982. Acoustic phonetic recognition. Proc. 6th Int. Conf. on Pattern Recognition, Munich, Germany, October 19-22, 1982.

Sherwood, G. A. 1979. The computer speaks. IEEE Spectrum, August 1979, pp. 18-25.

Simpson, C. A. In press-a. Voice Displays for Single Pilot IFR. NASA CR-172422. Hampton, Va.: NASA Langley Research Center.

Simpson, C. A. In press-b. Integrated voice controls and speech displays for rotorcraft mission management. SAE 1983 Transactions, Vol. 92, Section Y.

Simpson, C. A., C. R. Coler, and E. M. Huff. 1982. Human factors of voice I/O for aircraft cockpit controls and displays. Pp. 159-166 in Proceedings of the Workshop on Standardization for Speech I/O Technology, D. S. Pallett, ed. Washington, D.C.: National Bureau of Standards.

Spine, T. M., J. F. Maynard, and B. H. Williges. 1983. Error correction strategies for voice recognition. Proceedings of the Voice Data Entry Systems Application Conference, Chicago, Ill., Sept. 1983.

Spine, T. M., B. H. Williges and J. F. Maynard. In press. An economical approach to modeling speech recognition accuracy. Int. J. of Man-Mach. Stud.

Stokes J. M. 1982. VRAS System Guide. NADC Report 1400-19-B. Warminster, Pa.: Naval Air Development Center.

Van Cott, H. P. and R. G. Kinkade eds. 1972. Human Engineering Guide to Equipment Design (Rev. Ed.). Washington, D.C.: U.S. Government Printing Office.

Vestewig, R. G. and F. M. Propst. 1982. Speech input and video disk team up for military maintenance. Speech Technology 1(3):73-75.

Voorhees, J. W. and Z. A. Kersteen. 1983. The role of voice technology in an integrated cockpit system. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, Calif., August 1983.

Welch, J. R. 1977. Automated Data Entry Analysis. RADC TR-77-306. Griffiss Air Force Base, N.Y.: Rome Air Development Center.

Werkowitz, E. 1984. Speech recognition in the tactical environment: the AFTI/F-16 voice command flight test. Paper presented at Speech Tech. '84 Voice Input/Output Applications Show and Conference, New York, April 2-4, 1984.

Wickens, C.D. 1984. Engineering Psychology and Human Performance, Columbus, Ohio: Charles E. Merrill Publishing Company.

Wickens, C. D., D. L. Sandry, and M. Vidulich. 1983. Compatibility and resource competition between modalities of input, central processing, and output. Human Factors 25:227-248.

Wierwille, W. W. and S. A. Connor. 1983. Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. Human Factors 25:1-16.

Williams, C. E. and K. N. Stevens. 1969. On determining the emotional state of pilots during flight: an exploratory study. Aerospace Medicine 40:1369-1372.

Williamson, D. T. and D. G. Curry. 1984. Speech recognition performance evaluation in simulated cockpit noise. Paper presented at Speech Techn. '84 Voice Input/Output Applications Show and Conference, New York, April 2-4, 1984.

Wolf, J. J. and W. A. Woods. 1980. The HWIM speech understanding system. Pp. 316-339 in Trends in Speech Recognition, W. A. Lea, ed. Englewood Cliffs, N.J.: Prentice-Hall.

Woodard, J. P. and E. J. Cupples. 1983. Selected military applications of automatic speech recognition technology. IEEE Comm. Magazine, Dec., pp. 35-41.

Woods, W. A. 1975. Syntax, semantics, and speech. Pp. 345-400 in Speech Recognition: Invited Papers at the IEEE Symposium, D. R. Reddy, ed. New York: Academic Press.

Woodson, W. E. 1981. Human Factors Design Handbook. New York: McGraw-Hill.

APPENDIX A

AUTOMATIC SPEECH RECOGNITION WORK OUTSIDE THE UNITED STATES

At present, about two dozen companies are marketing automatic speech recognition devices in the United States. Their products are based on technology that largely originated in a few industrial laboratories and universities. The majority of this fundamental technology is available in the open technical literature, although some is protected by relatively narrow patents, and some is retained as trade secrets.

A similar fundamental understanding and capability exists in several other countries. In particular, significant expertise resides in Japan, Sweden, France, England, West Germany, and Canada. Of these, Japan has the most advanced effort, possibly equaling or surpassing that of the United States. Their motivation for automatic speech recognition (ASR) is double-edged. First, the syllabic structure of spoken Japanese is amenable to automatic speech recognition. Second, the complexities of the written language (Kanji and Kana characters) make keyboard design and operation burdensome, and make automatic "voice-to-text" conversion highly desirable.

Two significant forces have been at work in Japan to foster early ASR applications. One is the national telecommunications industry, led by Nippon Telephone and Telegraph (NTT) in partnership with its "family" companies (such as Nippon Electric Company, Fujitsu, Hitachi, Mitsubishi, Oki, and, more recently, Toshiba and Matsushita). The other is government stimulation by the Ministry of International Trade and Industry (MITI) of the "fifth generation" computer research, which has a strong component of voice input/output capability. As a result, the Japanese now have 20 to 40 speaker-independent, isolated-word systems working in their public telephone network. Each system, located in a telephone central office, can serve up to 128 calling lines. These small vocabulary systems serve the Japanese banking and financial industry. No comparable field deployment of ASR exists in other countries. Additionally, a number of simpler, lower-cost ASR systems are becoming available for personal computers, factory data entry, and device control. The ASR systems are typically combined with speech synthesis components based on technology that ranges from 32k bits/s Adaptive Differential Pulse Code Modulation (ADPCM) down to 2.4k bits/s Linear Predictive Coding (LPC). These technologies are well understood and in advanced states in Japan.

In France, ASR work is supported by the National Telephone Research Laboratory (CNET) and by the National Center for Scientific Research (CNRS) in their LIMSI laboratory. The results of this national research are typically available under license to industry. Some initial, though not widespread, commercialization has been made of the CNRS output. Applications under development include voice control of systems in the Mirage fighter aircraft and of power windows, locks, and directional signals as options in automobiles.

In Sweden, most recognition and synthesis products stem from research conducted in the Speech Transmission Laboratory of the Royal Technical University (KTH). Small venture businesses are selling text-synthesis and word recognition products. The developments also are being fostered by the Swedish government as communication aids for the handicapped.

Experimental products in England, West Germany, and Canada seem to be following very much the lines of the U.S. work, with British Telecom and Marconi, Siemens and Philips, and Bell Northern, respectively, included among the participants.

Speech recognition work has a long history in the Soviet Union, but the paucity of sophisticated computers and advanced electronics for civilian research seems to have retarded its development. At least on the basis of open publications, their progress seems several years behind the West.

In summary, while the United States is at the forefront in fundamental understanding and electronic implementation, it is not the undisputed leader in speech recognition and synthesis. Much of the developed world has comparable knowledge; Japan, at least, may have superior implementation. The competition is close enough, however, that modest additional investment in fundamental research could give the United States a significant edge.

APPENDIX B

SITE VISITS

Technical briefings provided much of the information collected by the Committee on Computerized Speech Recognition Technologies. However, realizing there is no substitute for first-hand experience, members of the committee participated in a number of field trips. In this way, the committee was able to improve its understanding of the specialized environments and stressful situations in which computerized speech recognition systems must function. Additionally, members interacted directly with more equipment users and researchers than would otherwise have been possible.

These field trips are summarized below:

1. July 21, 1983 - A panel composed of committee members George Doddington, Ellen Roland, and John Ruth, plus staff officer Howard Clark, visited Edwards Air Force Base in California to observe flight tests of an operational speech recognition system in an AFTI F-16 aircraft. The panel received briefings on the program's goals and accomplishments, got a close-up view of the aircraft cockpit, and interviewed three test pilots.
2. September 15, 1983 - The committee visited the System Control Laboratory of the Naval Surface Weapons Center in Dahlgren, Virginia, and received briefings on the problems and potential applications of speech recognition systems in the surface Navy.
3. October 19, 1983 - Committee member Carol Simpson visited Lockheed Missiles and Space Company in Sunnyvale, California, to learn about their use of commercial voice recognition systems for data entry of part numbers, dash numbers, lot numbers, and serial numbers to provide traceability of parts--a service that is required of military suppliers. Lockheed has used voice technology data-entry for five years in its hybrid circuit assembly section and for three years in an electronic assembly area.

4. October 26, 1983 - The committee's staff officer, Howard Clark, toured the Army's AVRADA research facilities in Fort Monmouth, New Jersey. Current research activities there involve a performance evaluation of various commercial speech recognition systems operating in a simulated (noisy) helicopter environment. In another laboratory, Army researchers have installed a speech recognition system in a helicopter simulator and are examining its potential mission applications.
5. November 8, 1983 - The committee toured the Combat Information Center of the guided missile cruiser, U.S.S. Long Beach. Committee members were able to explore the working environment and procedural constraints within which a speech recognition system must function.
6. November 11, 1983 - Committee member Carol Simpson went to sea aboard the U.S.S. Long Beach to witness the operation of the Combat Information Center under more realistic conditions than were afforded during the committee's visit on November 8, 1983.

LIST OF ACRONYMS

ACE	Air Controller Exerciser
ADPCM	Adaptive Differential Pulse Code Modulation
AFTI	Advanced Fighter Technology Integrator
ARPA	Advanced Research Projects Agency
ASR	Automatic Speech Recognition
AVRADA	Avionics Research and Development Activity, U.S. Army
CCWS	Command and Control of Weapon Systems
CNET	National Telephone Research Laboratory, France
CNRS	National Center for Scientific Research, France
CSR	Continuous Speech Recognition
CWR	Connected Word Recognition
DARPA	Defense Advanced Research Projects Agency
DBMS	Data Base Management System
DOD	Department of Defense
IFR	Instrument Flight Rules
IID	Integrated Information Display
I/O	Input/Output
IWR	Isolated Word Recognition
KTH	Royal Technical University, Sweden
LIMSI	Laboratory for Science and Mechanical Engineering, France
LPC	Linear Predictive Coding
MIPS	Million Instructions Per Second

MITI	Japanese Ministry of International Trade & Industry
NASA	National Aeronautics and Space Administration
NBS	National Bureau of Standards
NEC	Nippon Electric Company
NTT	Nippon Telephone & Telegraph Company
PARTS	Precision Approach Radar Training System
SNR	Signal-to-Noise Ratio
VIS	Voice Interactive Systems
VLSI	Very Large Scale Integration
VRAS	Voice Recognition and Synthesis