ARPA ORDER NO.: 189-1
3P10 Distributed Information Systems

R-1386-ARPA
January 1974

The Use of Speech for Man-Computer Communication

Rein Turn

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

The research described in this Report was sponsored by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
3P10 Distributed Information Systems

R-1386-ARPA
January 1974

The Use of Speech for Man-Computer Communication

Rein Turn

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

Published by The Rand Corporation

PREFACE

This report is part of a Rand study of "Voice Data Processing Capabilities Applied to Defense Requirements." The project is designed to augment the current speech understanding research (SUR) of other Defense Advanced Research Projects Agency contractors by investigating applications of this research to military systems. Among the various components of the project are:

- Analysis of the nature of speech as a man-computer communication channel.
- Identification of military man-computer interfaces where the use of speech would be operationally attractive.
- Study of the acoustic signal processing aspects of speech understanding systems.
- Study of natural language and linguistic aspects of speech understanding systems.

This report focuses on the nature of speech as a man-computer communication channel. It discusses various intrinsic characteristics of speech that may be attractive, or cause problems, in man-computer communication.

The material in this report should be of use to planners, designers, and implementers of man-computer interfaces, and to researchers in speech recognition and understanding. The Information Processing Technology branch of ARPA, in particular, should find this report useful in their larger study of speech understanding by computer.

SUMMARY

This report investigates the intrinsic characteristics and the associated attractive features and problem areas of speech as a man-computer communication channel. Among the attractive features of speech and auditory channels are their independence of visual and manual channels, the omnidirectional nature of speech propagation, the ability to communicate simultaneously with men and machines, and the potential for using a telephone instrument as a complete computer terminal.

The problem areas include incomplete knowledge of linguistic and semantic aspects of speech processing, lack of effective techniques of acoustic signal processing, and the need for large amounts of digital processing. It is expected, however, that the results of the current large speech understanding research projects and the advances in digital technology should, in a few years, permit economically attractive implementation of speech-based man-computer interfaces.

CONTENTS

I. INTRODUCTION

Many contemporary computer applications require continuous interaction between men and computers. Typically, men communicate to computers new material in the form of programs and data, requests for processing or retrieval of previously stored data or data entered from external sources, and other information required for processes performed by the computer. In turn, computers communicate to men the requested information, results of completed processes, and any other information they are programmed to produce.

The principal man-computer communication channels are manual, visual, and audio channels. In this report, the manual channel is considered to include all mechanically operated computer-input devices, not just those operated by hand. The visual channel includes displays and signals for visual sensing by man and electro-optical sensing by computers. The audio channels include computer equipment and systems for recognizing spoken utterances and equipment for producing synthetic speech.

Most of the present interactive computer applications employ manual channels for man-to-computer and visual channels for computer-to-man communication. The use of the speech channel for these purposes is still in its infancy. However, recent advances in designing speech synthesis equipment and the current research efforts in the design of techniques for computer recognition of speech are likely to make speech communications between man and computer technically and economically feasible in a few years.

The choice of man-computer communication channels depends on numerous operational, human, and economic factors. Most important among these are the ease of use in the context of the tasks performed, the interaction language used, and the operational environment; the ability to maintain required interaction rates; the implications on processing speed and memory capacity; and the cost-benefit advantages over other, competing channels.

For manual and visual channels these factors have been thoroughly analyzed and are widely available in the literature [1]. In the case of the speech channel, however, this information is scarcer [2,3]. The purpose of this report is to provide additional design information by identifying the attractive features and problem areas associated with the use of speech as a man-computer communication channel.

II. SPEECH AS A MAN-TO-COMPUTER COMMUNICATION CHANNEL

It is a natural activity for a person to mentally encode his requests, observations, and ideas into sentences of a natural language—one that he uses in his daily communication with other persons—and express these in spoken form. Natural languages have evolved over long periods of time and, characteristically, permit great flexibility in expression and enormous variety in shades of meaning. That is, the mapping of mental images into natural language expressions is a many-to-many process. The resolution of the uncertainty inherent in natural language statements is done by the receiver on the basis of the context—the receiver's knowledge of the speaker's characteristics, the circumstances associated with the communication, and so on. Often the uncertainty cannot be resolved at all and the receiver must request additional information.

The expression of a given natural language statement in speech is another many-to-many transformation—the generated acoustic signals differ from speaker to speaker as functions of their voice tract physiology, sex, accent, dialect, physical condition, and emotional state. Further, all natural languages contain homonyms, which can be resolved only in context.

The understanding of spoken natural language expressions is a complex process that must draw upon a great deal of the receiver's accumulated experience and knowledge and may require further clarifying communications with the speaker. Attempting to understand a spoken utterance without the use of context and previous knowledge is similar to the problem of understanding a spoken expression in a foreign language by looking up each word in the dictionary. First one would have to hypothesize the spelling and handle the homonym possibilities, and then resolve the multiple meanings.

The use of unconstrained natural language utterances for speech communication with computers is beset with the difficulties outlined above. Since it is not practical to provide a computer with all the contextual information required to resolve the ambiguities inherent in unconstrained natural language, some restricted form of the language must be used. For example, the vocabulary may be limited to a few hundred words that are used with unique meanings, and rigid syntactical rules may be imposed. Further, contraints may be placed on the speakers (it may be required that isolated-word speech, rather than continuous speech, be used—each word would be uttered separately with a pause after each word). Despite the loss in expressional power and flexibility that such restrictions entail, there are situations

where speech may be attractive for man-to-computer communication even if severely constrained languages must be used.

The following sections discuss the intrinsic characteristics and the associated attractive features and problem areas of the use of speech as a man-to-computer communication channel. A part of this discussion is based on material that has previously appeared in the literature [2-5].

For ease of reference, the following code system is used to designate each characteristic, attractive feature, and problem area: the letter C indicates a characteristic, the letter A an attractive feature, and the letter P a problem area. For example, the first characteristic is designated by C-1, its first attractive feature by A-1.1, and its first problem area by P-1.1.

MESSAGE GENERATION AND ENCODING

The constant use of speech has made humans very skillful in communication with others through this channel. Speech can be produced effortlessly, spontaneously, at a high rate, and under almost all environmental conditions. Speech is the principal way humans communicate with each other. Hence, the following characteristics of speech can be identified:

C.1 Speech is man's natural and primary communication channel.

The first attractive feature of this characteristic is:

A-1.1 The use of speech is familiar and convenient when the language is similar to the speaker's native tongue and is easy to pronounce.

The speech channel loses its attractiveness as the language departs more and more from natural language, i.e., when words are artificially composed and are difficult to pronounce, requiring character by character spelling; when the syntax is rigid; and when abbreviations, numeric data, special symbols, and punctuation marks must be included. Military travel orders are representative of a language that is unattractive to read aloud. Although any person can be trained to become fluent in some special language, departures from familiarity certainly diminish the attractiveness of speech.

A-1.2 Speech is highly suitable and the preferred channel for spontaneous generation of messages.

Among such channels may be emergency messages and orders to change some action. It has been claimed that under normal circumstances speech generation has lower reaction time than moving a hand or even a finger to operate a pushbutton. Situations requiring emergency inputs into a computer may arise in connection with human monitoring of computer controlled processes or equipment and computer monitoring of human performance (for example, when the human controller of some processes or equipment becomes physically incapacitated). One would expect, however, that such emergency commands would be items of a very limited vocabulary (such as "Stop" and "Help") and would involve only a few words.

A-1.3 Speech is potentially the highest capacity versatile communication channel for man-to-computer input.

Data about the communication rates possible by using speech and the various manual channels are summarized in Table 1. These data show that speech is a high capacity communication channel naturally available to all humans (who would be likely to be interacting with computers) without the need for additional training.

Considerably higher data input rates are possible with special keyboards where a complex statement can be entered by operating a single pushbutton reserved specifically for the statement. Although this arrangement is faster than typing or speaking, such a pushbutton arrangement is not very flexible and it requires training; similar speed advantages are also possible in the speech channel by using codes.

Table 1

DATA RATES FOR MAN-TO-COMPUTER COMMUNICATION

Communication Mode	Rate (Words/sec.)	Remarks
Oral reading [9]		
Random words	2.1 - 2.8	Selected from 5000 word dictionary
Random words	3.0 - 3.8	Selected from 2500 most familiar monosyllable words
Nontechnical prose	3.9 - 4.8	
Repeating the same word	8.0 4.0	One syllable Two syllables
Silent reading	2.5 - 9.8	
Spontaneous speaking	2.0 - 3.6	
Handwriting [4]	.38 - .42	
Handprinting [4]	.22 - .53	
Typing [10]		
Skilled	1.6 - 2.5	Text (100 wpm - 150 wpm)
Inexperienced	.2 - .4	
Stenotype (chord typewriter) [11]	3.3 - 5	Typically 1/3 of the strokes of the typewriter
Operating touch-tone telephone [4]	1.2 - 1.5	10 buttons
Operating thumb-wheel input device	1.8 digits/sec.	Sequence of 10 digits [12]
Rotary dialing	1.54 digits/sec.	Sequence of 10 digits [12]

A-1.4 Using speech, simultaneous communication with both men and computers is possible.

This advantage over conventional man-computer communication devices can be used in the design of systems where computers monitor and assist in decision processes and planning. With the help of expected future developments in linguistic processing, decision theory, heuristic search, and other topics of artificial intelligence, data bases and programs could be developed for analyzing human conversations and statements for logical consistency and factual content, point out overlooked implications, and the like. Such systems are not likely before the 1990s, but they offer intriguing possibilities [13].

INTERACTION WITH OTHER CHANNELS

The next characteristic pertains to the interactions of the speech channel with other communication channels available to humans:

C.2 The speech channel is independent of the visual channel or human voluntary motor activities (other than those required for speech production).

The only muscles required for speech production are those that operate the vocal cavity, tongue, jaw, and lips and that control breathing. Other muscles and other bodily activities interfere only insofar as they affect breathing or require conflicting mental activities.

A-2.1 Communication using speech can take place simultaneously with other visual or manual tasks, when the speaker is walking, and in total darkness.

This is a very important feature of the speech channel. In numerous situations, especially in military systems, communication with computers is not the only task. A standard example is piloting an aircraft while interacting with other equipment through a computer.

Other situations where the user's eyes and hands are occupied but information must be entered into the computer include the following:

- Computer-aided troubleshooting of equipment, performing experiments, medical diagnosis.
- Source data input in taking inventory, in making field observations, in tracking tasks.
- Operating computer-graphics equipment—graphic input tablet and stylus, examining reconnaissance photographs.
- Monitoring of computer control of processes and equipment.
- Control of teleoperator systems.
- Data fusion, as in intelligence work.

Most of these applications involve well-defined tasks where a speech interface would require only a small vocabulary, and isolated-word speech recognition would be adequate. For example, a proposed voice-operated radio channel selector [14] has a vocabulary of 12 words to be spotted in continuous speech.

SPEAKER CHARACTERISTICS

The acoustic characteristics of the generated speech signals depend on the structure of the speaker's vocal tract (a function of the speaker's sex and age) and its dynamics. The latter is a function of the native language of the speaker (if not English, a foreign accent) or his geographic background (a regional accent). Infections and other pathological conditions in the vocal tract or nasal cavity also affect the speech quality. Articulation and timing are influenced by fatigue. Unusual emotional conditions can change the normal speech characteristics (change the pitch, cause tenseness, change breathing rate).

C.3 Speech contains a great deal of information about the speaker: his physiological characteristics; physical condition; emotional state; and geographic, national, and cultural background.

This leads to two attractive features and two problem areas in the application of speech for man-to-computer communication.

A-3.1 The use of speech input allows checking the speaker's identity for access control purposes.

There is considerable interest in using speech differences as a means for authenticating a person's identity. Carefully chosen speech samples can be analyzed and a set of speech parameters computed and stored. To authenticate his identity the person speaks a predetermined sentence. The speech parameters determined from this sample are compared with stored parameters. Considerable work is being done on this topic [15, 16]. Hence, using speech as an input channel allows checking of the user's identity as a by-product.

A-3.2 The speech communication channel has the potential for monitoring the physical and emotional state of the user.

The capability stems from the effects on speech of fatigue, illness, and emotions, as mentioned above [17]. For tasks requiring an operator's full attention and sound judgment, the speech channel may allow checking his condition.

A problem area associated with characteristic C.3—the person-to-person variability of speech signal and its dependence on physical and emotional condition—complicates the speech processing task and requires knowledge of the speaker's characteristics. These can be obtained beforehand or "learning sessions" must be arranged for the new speaker before he can operate the interface.

P-3.1 The variations of speech signals with individual characteristics and conditions can greatly increase the processing and storage

space requirements for speech understanding or recognition and increase recognition error rate.

Another problem related to this is the variation in pronunciation and speaking habits of speakers from different geographic, national, or cultural origins.

P-3.2 An acceptable recognition rate with a particular speaker requires the determination and storage of his speech characteristics. Hence, spontaneous replacement of one operator with another may be difficult and a training session may be required.

Efforts are underway to minimize the tuning required. In some applications where the speakers are uncooperative (as in monitoring of voice communications for intelligence purposes) this is a considerable problem.

Both of the above problems can be expected to arise in systems where the speakers are likely to have heterogeneous backgrounds. Operators could be specifically chosen from a more or less homogeneous group, but the possibility for easy replacement with a speaker not in the group becomes more difficult.

SPEECH PROPAGATION

Speech propagates in the atmosphere in the form of pressure waves. It also propagates through liquid and solid media, but these introduce attenuation and distortion. Pressure waves are reflected from and around objects. They can be easily changed into electrical form and back again.

C.4 Speech propagation is omnidirectional. No free line of sight is required.

This leads to the following attractive feature of speech for use as a man-to-computer communication channel:

A-4.1 For speech input, the speaker can be in an arbitrary orientation relative to the microphone, at considerable distance from the microphone, or behind a barrier.

Microphones with various "fields-of-view" and sensitivities can be constructed. The user may move around relative to the microphone in performing a task. The computer input console need not be user-centered, but may be "stretched out" to allow optimal placing of various output devices and displays side by side. The user can walk back and forth while inputting information.

There is also a problem area here—interference by the ambient acoustical noise:

P-4.1 The omnidirectional nature of speech propagation allows interference by other acoustical signals generated in the same room or in the general environment.

Interference may be due to users of another speech interface in the same room or to operational noise of the computer system or other equipment, or from outside. The noise problem will be discussed in more detail in a subsequent section. It is mentioned here to show that the nature of speech signal propagation creates the interference potential.

C-5 Speech is easy to convert into electrical form for long distance transmission. Transducers are inexpensive and small and can provide high fidelity.

Consequently, the attractive features for man-to-computer communication are:

A-5.1 Speech communication with computers is compatible with existing voice communication networks and systems. This allows remote input from locations where no special computer-related equipment is available.

A-5.2 The use of lightweight, portable microphones or microphones built into other equipment allows considerable freedom of movement by the user.

These two features also show that the existing voice communication system can be used for speech input to a computer if it meets certain minimum quality standards.

The following is a problem area in implementation of speech interfaces with computers:

P-5.1 The electrical form of speech input is subject to electrical noise and distortion in the telephone system or in radio communications.

Certain speech sounds (such as fricatives) resemble white noise, which also occurs in telephone and radio transmissions and, thus, confuse the recognition system. Other common types of noise and distortion are burst noise, echo, crosstalk, frequency translation, and clipping. All of these can increase understanding recognition error rates [18,19].

Finally, there is one more problem area caused by the nature of speech propagation:

P-4.2 Speech communications can be overheard directly by others in the vicinity, or by using acoustic pickup devices. Hence, another dimension has been added to the security problem in man-computer communications.

This acoustic emanation problem is added to the existing electromagnetic emanation problem, and to all the other data security threats that exist independently of the mode of the man-to-computer communication interface [20].

A speech signal propagating through the atmosphere is a transitory phenomenon. A speech input into the computer, likewise, does not leave an easy-to-perceive hard copy. An acoustic tape recording can be made, but this is troublesome to consult.

C-6 Speech propagation is transitory and volatile.

The associated problem area is

P-6.1 No hard copy is produced of speech input as a natural by-product. A magnetic recording can be made but is inconvenient to use.

ENVIRONMENTAL INFLUENCES

Speech generation and speech propagation are both affected by environmental conditions. Certain of these (such as temperature, humidity, or cramped condition of the speaker) affect speech generation or propagation only indirectly (for example, through accelerating the onset of fatigue and emotional conditions); others have more direct effects (such as mechanical forces on the speaker, composition of atmosphere, need to wear special equipment).

C-7 Speech production is affected by mechanical forces on the speaker.

The mechanical forces may be in the form of vibrations, acceleration forces, or other steady state or random forces caused by motion. The same forces also affect operation of manual input devices and, in some cases, the effect is more pronounced. In the case of speech, mechanical forces mainly affect breathing and controlling of the jaw, the organ with the greatest mass in the speech production system.

In comparison with the use of conventional man-computer interfaces, speech has the following attractive features when subjected to various environmental conditions.

A-7.1 Speech is unaffected by weightlessness.

There is no evidence that speech is affected by weightlessness (at least as far as short duration space flights are concerned) and artificial gravity. Although the movement of hands and fingers is also not appreciably impaired under these conditions, the operator may have to be strapped to a conventional input terminal.

Regarding susceptibility of speech to other mechanical forces, speech generation and voice characteristics may be affected by body resonances and sudden jolts:

P-7.1 Speech generation is affected by vibrations, high levels of accelerations, and other mechanical forces.

However, these effects are not very substantial. For example, a set of experiments in a centrifuge [21,22] showed that the speech recognition accuracy of a specific isolated-word recognition system was changed about 5 percent when vertical sinusoidal vibration was increased a nominal .05 to .3 g. Sustained acceleration reduced the recognition accuracy by 10 percent when the subject received 4 g. acceleration. The main problems here were difficulty in maintaining normal breathing, increased breathing noise, and straining of facial muscles.

The effect of vibration on tactile input devices decreases the input rate and increases errors. For example, an experiment involving operating pushbuttons,

rotary dials, and thumbwheels, [23] showed almost negligible change in performance at .3 g. vibration, but about 10 percent degradation at .8 g.

Changes in the atmospheric pressure and composition also affect speech generation and voice characteristics:

C-8 Speech production and propagation are affected by the composition of the atmosphere and the ambient pressure.

There are no attractive features associated with this speech characteristic. There are some problems, however.

P-8.1 Speech intelligibility and the natural voice characteristics of the speaker are affected by atmospheric composition and pressure.

This causes problems in submarine systems, especially in voice communication from divers [24,25], as well as in other manned systems. Elevated pressure, likewise, affects speech intelligibility [26]. Another problem arises in the use of breathing equipment by pilots and astronauts:

P-8.2 Special breathing equipment, such as an oxygen mask, produces breathing noises that affect speech recognizability.

The noise spectra of inhaling and exhaling are broadly spread and mask many of the important speech sound frequencies [22].

AMBIENT NOISE

In almost every environment there are ambient acoustic signals due to people, equipment in operation, or natural phenomena that potentially interfere with speech inputs:

C-9 A propagating speech signal is subject to interference by any other acoustic signal.

There seem to be no strong attractive features of the speech channel due to this characteristic, although certain information about the speaker's environment may be extracted from the interfering ambient noise. For example, if the background noise represents operation of some equipment being monitored by the speaker, a change in the background noise spectrum may be a signal of approaching malfunctioning of the equipment. In some other situation, the speaker himself may be at a location that is subject to intrusion or danger. The intrusion noises here may alert the central control and permit quick dispatching of assistance. Thus, a weak attractive feature might be claimed:

A-9.1 Interference of speech signals by other ambient acoustic signals permits extraction of information about unusual activities at the speaking location.

The problem area is an obvious one:

P-9.1 Interference of the speech signal by ambient noise can greatly reduce speech recognizability.

Depending on the nature of the ambient noise (its frequency spectrum, intensity, frequency of occurrence), the reliability of the speech channel may be sporadic. In certain applications, use of speech as a computer input may be entirely ruled out because of the ambient noise. Indeed, even man-to-man speech communication is impossible in many high noise environments.

The ambient noise problem may be quite acute in environments containing equipment in operation (aircraft engines, teletype terminals, and so on) or other speakers [27]. Among the techniques available for alleviating this problem are noise-cancelling microphones, special vocabulary designs, and signal processing techniques [28].

IMPLEMENTATION OF SPEECH INTERFACES

Implementation of the speech communication channel for the man-to-computer interface requires considerable equipment and processing: a microphone for speech-to-electrical signal conversion, analog equipment for speech feature extraction and for analog-to-digital conversion, and a digital computer for the recognition and understanding of the utterance [5,6,29,30]. However, only the microphone need be in the same location as the speaker; the rest of the equipment is usually at the site of the computer, and the necessary processing may be performed by the same computer.

C-10 A microphone is the only speech interface equipment that must be in the same enclosure as the speaker.

The consequent attractive feature is:

A-10.1 The speech interface can be implemented without using any space on a terminal or console panel.

This has important connotations in systems where many displays and controls are packed on a terminal or control console panel.

Another characteristic of the speech interface pertains to speech processing:

C-11 A digital computer is an essential element in the implementation of a speech interface for computer input.

Compared with the manual channel, the speech interface involves more processing and equipment. Indeed, it is unlikely that the speech interface can ever compete with keyboard, pushbuttons, and the like on a strict equipment and processing cost basis. This produces the problem area:

P-11.1 The speech interface requires special analog equipment and digital processing. The latter depends on the constraints placed on the

interaction language (vocabulary size, the amount of pausing between words, syntactic rules) and on the nature of the tasks involved.

Although all types of speech interface implementations require equipment for initial processing of the acoustic speech signal [29], the requirements for linguistic processing [30,31] depend on the various characteristics of the interface. For example, very modest amounts of linguistic processing may be required in isolated-word, syntactically constrained, small vocabulary speech recognition systems. However, the linguistic processing required for unconstrained natural language understanding and recognition systems is still beyond the capabilities of contemporary computer science.

The isolated-word speech interface systems have already become a reality [32], but more work is required to develop sufficiently capable continuous speech understanding systems [33]. The present research efforts in this area are expected to lead to practical, continuous speech man-to-computer interfaces in the late 1970s [4]. However, these systems will continue to place restrictions on the syntax and vocabulary size.

III. SPEECH IN COMPUTER-TO-MAN COMMUNICATION

Unlike the use of speech for computer input, automatic synthesis of spoken messages by computers is now practical. This is indicated by a recent survey of the state of the art [34] and by the number of firms producing voice response and speech answer-back equipment [35,36]. Hence, only a brief discussion of the attractive features and problem areas of the use of speech for computer-to-man communications is presented to complement the more extensive discussion above of its use for man-to-computer communications.

ATTRACTIVE FEATURES AND PROBLEM AREAS

The following attractive features of speech as a computer-to-man communication medium ensue directly from the general characteristics of speech discussed in Section II. The coding system is continued here:

A-1.5 Speech is the natural way for humans to receive communications from others. It is compatible with the use of speech as the computer input channel.

Humans can maintain a high level of vigilance for acoustic signals and are capable of detecting the expected verbal messages despite high levels of ambient noise. It may be possible to listen to more than one spoken message at a time; special alerting and emergency messages can get immediate attention:

A-1.6 Several spoken messages could be received and understood simultaneously.

A problem area here is the speed of spoken computer-to-man communication compared with that of the visual channel; the visual channel is normally many times faster:

P-1.3 The rate of receiving spoken messages is much slower than the rate of receiving messages through the visual channel.

Just as in the case of speech generation, the human auditory input channel is independent of the visual channel and of most of the human motor activities. How-

ever, there is considerable interaction between the auditory channel and speech generation.

A-2.2 Spoken messages can be received without interrupting the use of the visual channel or any motor activities and in total darkness.

As pointed out in Section II, human speech contains a great deal of information about the speaker, as well as non-verbal clues (prosodic features), which may be used by the speaker to augment the verbal message. These are analyzed by the listener and used to resolve ambiguities (for example, the final inflection can change a sentence from a statement to a question, or the lack of it can change a statement phrased as a question into an exclamation). Although the present state of the art of speech synthesis is not yet sufficiently advanced, the ability to synthesize prosodic features may make speech output from computers more effective than visual messages.

A-3.3 Speech output has the potential for highly effective computer-to-man communication.

The omnidirectional nature of speech propagation in the atmosphere has several attractive features for the use of synthesized speech for computer-to-man communication:

A-4.2 In computer-to-man communication through speech, the human listener can be in arbitrary orientation, some distance from the computer, or behind a barrier, and he may be in motion.

A-4.3 Any number of listeners can receive the spoken message from the computer simultaneously.

The attractive features A-5.1 and A-5.2 regarding the compatibility of speech with existing voice communication systems also apply to computer-produced spoken messages. Indeed, given a computer system with both speech input and speech response capability, the ordinary telephone instrument becomes a computer terminal.

Speech is a transient and volatile phenomenon and requires special efforts for converting into readily accessible hard copy form, making problem P-6.1 equally applicable to computer-to-man communication.

Of the environmental factors, only ambient noise has effects on the human ability to receive spoken computer messages when these are sent in a broadcast manner. However, even when spoken messages are broadcast in moderately noisy environments, the human auditory system has the ability to concentrate on a specific message and ignore other messages or noise (this is the so-called "cocktail party situation"). Individual headsets can be used even in extremely noisy environments.

A-7.2 Speech reception by humans is not appreciably affected by weightlessness, vibration, or mechanical forces.

Finally, regarding the implementation of the speech output interface from a computer, the listener needs no other equipment than a speaker or headsets. The attractive feature A-10.1 also holds here; the speech output equipment does not

complicate the terminal equipment or require additional panel space (except for a simple volume control).

APPLICATIONS

The present applications of speech as the computer-to-man communication channel are mainly in banking and in credit checking industries, where simple, well-formatted responses can be used [34]. However, there is a great deal of interest in achieving general capabilities for converting text into synthesized speech [37]. Construction of reading machines for the blind is a research area of special interest [38].

IV. INTERFACE ANALYSIS FOR SPEECH APPLICATIONS

The design of an effective yet economical man-computer interface is a complex process that must take into account the nature of the tasks being implemented at the interface; the human roles, capabilities, and shortcomings in task performance; and the environment in which the interface is used. These are discussed below in terms of the speech channel characteristics and their associated attractive features and problem areas, which are summarized in Table 2 at the end of this section.

ROLES OF HUMAN OPERATORS

The most demanding role for a human operator in a man-computer system, military command-control systems in particular, is that of the *decision maker.* The role of the computer in this situation is to provide the necessary information and assistance for decision making support and to be instrumental in the dissemination of the decision. There are several dimensions that characterize military decision making [39] and place demands on the man-computer interface:

- *Criticality* of the consequences and outcomes. The interface must be reliable and permit unambiguous and secure communications. It must help the human operator to interact dependably under high levels of psychological stress.
- *Diversity* of the population of decisions to be made. The interface must be flexible.
- *Dynamic nature* of the decisions. Most of the decisions remain valid only for short periods and must be frequently modified. The interface must be natural, flexible, and easy to use.
- *Diversity* of decision makers. In the military, the operators come from heterogeneous populations and have different cultural backgrounds, knowledge of the interface capabilities, and decision making strategies. The interface must be simple to operate, flexible, and helpful.
- *Effectiveness* criteria for decisions are varied and intricate and depend on the operational context.

Man-computer communication through speech can contribute to the design of interfaces that are flexible and natural to use (characteristics C-1 and C-5). However,

speech communication is not necessarily suitable in all situations. For example, it is more natural to identify graphically presented information items by *pointing* at an item than by speaking the item's name or the coordinates of its location. The use of speech channel also places additional requirements on the *reliability* of the interface (P-3.1, P-3.2, P-9.1, and P-11.1). On the other hand, it also provides capabilities not readily achievable by using other man-computer communication channels (A-3.1, A-3.2, and A-9.1).

Among the other roles of human operators in man-computer systems are the following:

- *Sensor or transducer:* the human task is to input information into the computer system's data base. He may actually acquire the data directly (such as through surveillance of some activities of interest) or act merely as a transducer (such as converting printed text into computer readable form). This is essentially an open-loop operation, although feedback may be provided to assure the correct input of the data. A speech interface enhances this role through characteristics C-4, C-5, and C-10.
- *Retriever or inquirer:* the human task is to request information from the data base. The process is a simple question and answer operation within some well-defined task area (such as literature search or obtaining factual statements about force status). The naturalness and flexibility of the speech communication channel (characteristic C-1) can increase the operator's effectiveness in this role.
- *Controller:* the task is to order discrete state changes in the state of equipment or processes (such as using the computer to go through a checkout process one step at a time). The independence of the speech channel on the manual and visual channels (characteristic C-2), the propagation characteristics (C-4), environmental effects (C-7, C-8, and C-9), and interface implementation aspects (C-10 and C-11) influence the effectiveness of the speech interface.
- *Monitor:* the monitor observes an automated control operation performed by computer, where the human role is to oversee the process and intercede if necessary. The role here is to provide additional reliability as well as flexibility in handling unusual situations. Monitoring is a vigilance task, which often involves long periods of passive observation of data displays. The speech channel can both provide alerting information from the system and allow rapid responses by the operator to emergency situations (A-1.2, C-2).
- *Problem solver:* the problem solver is a participant in a computer-assisted task, where the computer contributes evaluations and data allowing the human partner to proceed. Examples here are computer-aided tracking of objects, diagnosis of malfunctions, and pattern recognition. Here the man-computer interface provides tight coupling and a great deal of feedback. Another form of problem-solving activity is performed by a trainee in a computer-assisted instructional system. Once again, speech characteristics C-1, C-2, C-4, and C-11 are applicable.

In all of the above roles of human operators in man-computer systems, speech communications with computers promise operational advantages. The principal drawbacks are the increased demands on interface reliability and the need for additional processing for the speech interface. The implementation of the speech

interface as an isolated-word (rather than continuous-speech) recognition system reduces the reliability problems, but it also reduces the attractiveness of speech communications from the point of view of naturalness and flexibility (characteristic C-1).

APPLICATION CRITERIA

In each of the above roles, the human operator performs a task or a set of tasks. The following characteristics of these tasks provide a checklist for applicability of the speech interface for their performance:

1. Nature of the task (routine, critical, time urgent).
2. Time characteristics of the task (continuous, periodic, sporadic).
3. Variability of the task.
4. Intensity level of task performance (high level interaction, vigilance task, routine interaction, monitoring).
5. Response requirement for task performance (time-critical, leisurely).
6. Input loading of task performer (the number of information sources and the need for their correlation for task performance).
7. Output loading of the task performer (the number of different responses that he may need to generate, different input mechanisms he may need to operate).
8. Operator's physical state when performing his task (sitting, standing, moving, prone).
9. Operator's physical safety and other stress conditions when performing task.
10. System's state when operator performs task (fixed stationary, continuous motion, erratic motion).
11. Operator's level of isolation in performing task (alone, part of a group, members of other groups) at the station.
12. Environmental condition (climatic, acoustic, mechanical, pressure, atmospheric).
13. Training and skill level requirements of the operator.
14. Nature of the interaction language and formats.
15. Requirements for security.

The speech understanding and recognition systems used to implement a speech interface are also characterized by a series of design features that must be taken into account when considering the use of speech for a given man-computer task performance. These are discussed in detail in [2] and need not be repeated here.

Answers to the above checklist provide information for the implementation of a speech interface in a particular man-computer task performance application and allow determination of the expected operational benefits. Equally important are the questions on the *costs* of implementing the speech interface: required processing power and memory capacity. Depending on the specifics of the proposed application, conditioning of the communication links and the volume, weight, and power consumption of the speech interface equipment may also be important.

AN EXAMPLE

As an example of the analysis of a man-computer interface for potential use of the speech communication channel, consider the computer-aided control of avionics functions and equipment in fighter aircraft: communications, flight control of the aircraft, navigation, fire control, electronic countermeasures, and test and fault location systems. At present, these systems tend to be autonomous and possess their own independent controls, processors, and displays. In the future, however, they will be parts of an integrated avionics information system [40].

The emphasis in this application is on the reduction of the pilot's present manual workload by using the speech channel (A-2.1). The commands to the avionics control system to change communication channel frequencies, present displays, or report equipment status can use the vocabulary and syntactical structure of the requests the pilot would issue to his copilot for the same purpose. Hence the interaction would be natural and rapid (A-1.1, A-1.3).

The interaction takes place in a limited context. A relatively small vocabulary (50-100 words) and a constrained syntax can be used without greatly affecting the interaction. The simplicity of the transducer and its panel equipment (C-5 and C-10) allows the pilot to enter commands without the need to concentrate on the manipulation of the interface devices.

The major problems arise from the environmental effects. The aircraft engine, the equipment in the cockpit, and the pilot's oxygen mask produce acoustical noise, which may interfere with the operation of the speech interface (P-4.1, P-8.2, P-9.1). Electrical interference and crosstalk also affect the speech interface (P-5.1). Special microphones, filtering equipment, or digital processing may be required for adequate reduction of the noise problems.

The principal effect of noise interference is the reduction of speech recognition accuracy. In most of the avionics control tasks, the pilot cannot be expected to offer a command more than twice. The need to repeat a command should arise only a small number of times. Any need to engage in a longer dialog to recognize a command will defeat the advantages of the speech interface. Hence, reliability is an important design criterion. All other interface design factors [4] can be used for achieving high reliability: The vocabulary and syntax can be selected and structured to minimize recognition ambiguity, considerable system tuning may be permitted, and user training may be made a part of the general flight training program.

Finally, the speech interface equipment—analog and digital processors and associated memory units—must be added to the aircraft's equipment load at the expense of space, weight, power consumption, and, possibly, other operational features that could have been incorporated in lieu of the speech interface. Whether these costs are acceptable in view of the benefits gained—relieving the pilot of manual control tasks that interfere with the performance of his primary missions—depends on the specifics of the situation.

Table 2

SPEECH INTERFACE CHARACTERISTICS

Characteristic	Attractive Feature	Problem Area
C-1 Speech is man's natural and primary communication channel	A-1.1 Use of speech is familiar and convenient when the language is similar to natural language	P-1.1 Artificial syntax, restricted vocabulary, etc. tend to mitigate the naturalness of speech
	A-1.2 Speech is highly suitable and the preferred channel for spontaneous messages	P-1.2 The use of feedback for clarification may reduce the channel capacity of speech to a considerable extent
	A-1.3 Speech is potentially the highest capacity, most versatile man-computer communication channel	P-1.3 For man, the rate of receiving spoken messages is much slower than receiving messages through the visual channel
	A-1.4 Using speech, simultaneous communication with both men and machines is possible	
	A-1.5 Speech is a natural way for men to receive communications from others. It is compatible with the use of speech as a computer input channel	
	A-1.6 Several spoken messages can be simultanously received and understood by man	
C-2 The speech channel is independent of the visual channel and of human motor activities	A-2.1 Computer input using speech can take place simultaneously with other visual or tactile tasks, or when the speaker is walking	
	A-2.2 Spoken messages can be received by man without interrupting the use of the visual channel or any motor activities, and in total darkness	
C-3 Speech contains information about the speaker	A-3.1 Speech interface allows checking the speaker's identity for access control purposes	P-3.1 Differences of individual characteristics increase implementation costs
	A-3.2 The speech communication channel allows monitoring the physical and emotional state of the speaker	P-3.2 Spontaneous replacement of speakers cannot be made--a training session is required to tune the recognition programs
	A-3.3 Speech output by computer has the potential for highly effective computer-to-man communication	

Table 2 (*Continued*)

Characteristic	Attractive Feature	Problem Area
C-4 Speech propagation is omnidirectional. No free line of sight is needed	A-4.1 For speech input, the speaker can be in an arbitrary orientation, some distance from the microphone, or behind a barrier	P-4.1 The omnidirectional nature of speech propagation allows interference by other acoustical signals
	A-4.2 The human receiver of spoken messages can be in any arbitrary orientation, some distance from the terminal, behind a barrier, or in motion	P-4.2 Speech communications can be overheard by anyone in the vicinity or by using eavesdropping devices, thus providing additional security threats
	A-4.3 Any number of listeners can receive spoken messages simultaneously	
C-5 Speech is simple to convert into electrical form	A-5.1 Speech communication with computers is compatible with existing voice communication networks and allows input from remote sites	P-5.1 The electrical form of speech is subject to electrical noise and distortion
	A-5.2 Use of simple, lightweight microphones allows freedom of movement	
C-6 Speech is transitory and volatile		P-6.1 No hard copy is produced as a byproduct of operation of the speech interface
C-7 Speech generation is affected by mechanical forces on the speaker, but less than the manual channel	A-7.1 Speech generation is not appreciably affected by weightlessness	P-7.1 Speech generation is adversely affected by vibration, g-loads, and other mechanical forces on the speaker
	A-7.2 Speech reception by man is not appreciably affected by weightlessness, vibrations, or mechanical forces on the listener	
C-8 Speech generation and propagation are affected by composition and ambient pressure of the atmosphere		P-8.1 Speech intelligibility and natural voice characteristics are adversely affected
		P-8.2 Breathing equipment produces noise interference
C-9 A propagating speech signal is subject to interference by other acoustic signals	A-9.1 Interference permits extraction of information about events at the speaker's location	P-9.1 Interference by other acoustic signals greatly reduces speech intelligibility
C-10 A microphone is the only transducer required at the speaker's location	A-10.1 The speech interface does not complicate the terminal equipment. It is simpler than for any manual channel	
C-11 A digital computer is an essential element in speech input processing		P-11.1 Special equipment and processing is required. This increases the cost or limits the interaction language

V. CONCLUDING REMARKS

The use of speech as a man-computer communication medium offers several attractive features over the conventional manual and visual channels. The most important among these are independence of the speech and auditory channels, which permits the performance of other manual or visual tasks while communicating with the computer; the omnidirectional nature of speech propagation, which permits the operator to communicate with the computer while he is in motion or remote from the input/output transducers; the ability to communicate simultaneously with both computers and humans; and the potential for using a telephone instrument as a complete computer terminal.

The current problem areas in implementing continuous speech input systems are theoretical, technical, and economic. Theoretical problems have to do with the present incomplete knowledge of linguistics and semantics and lack of efficient algorithms for automatic understanding of natural language utterances. Technical problems deal mainly with the acoustic signal processing of continuous speech utterances—word boundaries, basic speech elements, prosodic features, speaker-independent processing techniques, and the like. Economic problems stem from the need for special signal processing equipment and general purpose digital processing beyond the requirements of the current manual or visual interfaces.

None of these problems appear to be insurmountable in applications where constraints on vocabulary, syntax, and speakers are acceptable. The current ARPA speech understanding research (SUR) projects [4] and the research projects sponsored by other government agencies and private industry [34] are aiming to produce substantial continuous speech understanding capabilities in a few years. Isolated-word recognition systems are already being tested in "real life" applications and environments [14,21].

Despite the attractive characteristics of speech and auditory channels described in this report, their implementation in a particular man-computer task situation makes sense only when their use is *natural* for performing the task and *compatible* with the environment. The nature of information involved in a man-computer task must be thoroughly analyzed before committing to the use of a speech interface. Together with other modes of man-computer communication, the speech-based interfaces can help an operator concentrate on the task he is performing rather than on operating the interface.

REFERENCES

1. Meadow, C. T., *Man-Machine Communication,* John Wiley & Sons, New York, 1970.
2. Lea, W. A., "The Impact of Speech Communication with Computers," *Proceedings, Sixth Space Congress,* Cocoa Beach, Fla., March 1969, pp. 15-19 through 15-31.
3. Lea, W. A., "Establishing the Value of Voice Communication with Computers," *IEEE Transactions on Audio and Electroacoustics,* June 1968, pp. 184-197.
4. Newell, A., et al., *Speech Understanding Systems, Final Report of a Study Group,* AFOSR-TR-72-0142, Carnegie-Mellon University, May 1971.
5. Hill, D. R., "Man-Machine Interaction Using Speech," *Advances in Computers,* Vol. 11, Academic Press, New York, 1971, pp. 166-222.
6. Otten, K. W., "Approaches to Machine Recognition of Conversational Speech," *Advances in Computers,* Vol. 11, Academic Press, New York, 1971, pp. 127-163.
7. Barmack, J. E. and H. W. Sinaiko, *Human Factors Problems in Computer-Generated Graphic Displays,* Study S-234, Institute for Defense Analyses, Princeton, N. J., April 1966.
8. Quastler, H. (ed.), *Information Theory in Psychology,* The Free Press, Glencoe, Ill., 1955.
9. Pierce, J. R. and J. E. Karlin, "Reading Rates and the Information Rate of a Human Channel," *Bell System Technical Journal,* Vol. 36, 1957, pp. 497-516.
10. Hershman, R. L. and W. A. Hillix, "Data Processing in Typing: Typing Rates as a Function of Kind of Material and Amount Exposed," *Human Factors,* October 1965, pp. 483-492.
11. Seibel, R., "Data Entry Through Chord, Parallel Devices," *Human Factors,* April 1964, pp. 189-192.
12. Deininger, R. L., "Rotary Dial and Thumbwheel Devices for Manually Entering Sequential Data," *IEEE Transactions on Human Factors in Electronics,* September 1967, pp. 227-230.
13. Firschein, O., M. A. Fishchler, L. S. Coles, and J. M. Tenenbaum, "Forecasting and Assessing the Impact of Artificial Intelligence on Society," *Third International Conference on Artificial Intelligence, Advance Papers of the Conference,* Stanford University, Calif., 20-23 August 1973, pp. 105-120.

14. Scott, P. B. and J. R. Richards, *Speech Controlled Radio Channel Selector,* Technical Report AFAL-TR-71-266, Air Force Avionics Laboratory, Wright-Patterson AFB, Ohio, October 1971.
15. Su, L., *On Speaker Identification,* TR-EE 72-4, Purdue University, Lafayette, Ind., January 1972.
16. Meeker, W. F. and L. R. Simmering, *Foreign Language Speaker Identification,* RADC-TR-71-39, Rome Air Development Center, Griffis AFB, N.Y., February 1971.
17. Williams, C. E. and K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates," *The Journal of Acoustical Society of America,* Vol. 52, No. 4, 1972, pp. 1238-1250.
18. Erman, L. D. and D. Raj Reddy, "Implications of Telephone Input for Automatic Speech Recognition," *Working Papers in Speech Recognition, I.* Department of Computer Science, Carnegie-Mellon University, April 21, 1972, pp. 3.1-3.7.
19. Webster, J. C. and C. R. Allen, *Speech Intelligibility in Naval Aircraft Radios,* Technical Report NELC/TR 1830, Naval Electronics Laboratory Center, San Diego, Calif., 2 August 1972.
20. Carroll, J. M., *The Third Listener,* E. P. Dutton & Co, Inc., New York, 1969.
21. Private communications, Commander R. Wherry, Naval Air Development Center, Warminster, Pa., April 1973.
22. Glenn, J. W., R. N. Gordon, and G. Moschetti, *Voice Initiated Cockpit Control and Integration (VICCI) System Test for Environmental Factors,* Scope Electronics, Inc., Reston, Va., 30 April 1971.
23. Dean, R. D., R. J. Farrell, and J. D. Hitt, "Effect of Vibration on the Operation of Decimal Input Devices," *Human Factors,* Vol. 11, No. 3, June 1969, pp. 257-272.
24. Sergeant, R. L. and T. Murry, "Helium-Speech Processing: Report of a Workshop," *Conference Record, 1972 Conference on Speech Communication and Processing,* Newton, Mass., April 24-26, 1972, pp. 356-359.
25. Nixon, C. W., et al., "Study of Man During a 56-Day Exposure to an Oxygen-Helium Atmosphere at 258 mm. Hg Total Pressure: XVI. Communications," *Aerospace Medicine,* Vol. 40, No. 2, February 1969, pp. 113-123.
26. Murry, T., E. J. Nelson, and E. W. Swenson, "Speech Intelligibility During Exercise at Normal and Increased Atmospheric Pressure," *Aerospace Medicine,* Vol. 43, No. 8, August 1972, pp. 887-890.
27. Neely, R. B. and D. R. Reddy, "Speech Recognition in the Presence of Noise," *Working Papers in Speech Recognition, I.* Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa., April 21, 1972.
28. Drucker, H., "Speech Processing in High Ambient Noise Environment," *IEEE Transactions on Audio and Electroacoustics,* Vol. AU-16, June 1968, pp. 165-168.
29. Hoffman, A. S., *The Role of Acoustic Processing in Speech Understanding Systems,* The Rand Corporation, R-1356-ARPA, October 1973.
30. Klinger, A., *Natural Language, Linguistic Processing, and Speech Understanding: Recent Research and Future Goals,* The Rand Corporation, R-1377-ARPA, October 1973.

31. Walker, D. E., "Speech Understanding Through Syntactic and Semantic Analysis," *Third International Joint Conference on Artificial Intelligence, Advance Papers of the Conference,* Stanford University, Stanford, California, 20-23 August 1973, pp. 208-215.
32. "Spoken Words Drive a Computer," *Business Week,* December 2, 1972.
33. Reddy, D. R., L. D. Erman, and R. B. Neely, "A Model and a System for Machine Recognition of Speech," *IEEE Transactions on Audio and Electroacoustics,* Vol. AU-21, No. 3, June 1973, pp. 229-238.
34. Hill, D. R., "An Abbreviated Guide to Planning Speech Interaction with Machines: the State of the Art," *International Journal of Man-Machine Studies,* Vol. 4, 1972, pp. 383-410.
35. Hornsby, T. G., Jr., "Voice Response Systems," *Modern Data,* November 1972, pp. 46-50.
36. "Talking Computers," *Infosystems,* July 1972.
37. Ainsworth, W. A., "A System for Converting English Text into Speech," *IEEE Transactions on Audio and Electroacoustics,* Vol. AU-21, No. 3, June 1973, pp. 288-290.
38. Allen, J., "Reading Machines for the Blind: The Technical Problems and Methods Adopted for Their Solution," *IEEE Transactions on Audio and Electroacoustics,* Vol. AU-21, June 1973, pp. 259-264.
39. Debons, A., "Command and Control: Technology and Social Impact," *Advances in Computers, Vol. 11,* Academic Press, New York, 1971, pp. 319-390.
40. List, B., "DAIS: A Major Crossroad in the Development of Avionic Systems," *Astronautics and Aeronautics,* January 1973, pp. 55-61.