Speech as a man-computer communication channel

by REIN TURN
The Rand Corporation
Santa Monica, California

INTRODUCTION

Many computer applications require continuous interaction between men and computers. Typically, men communicate to computers data and programs, requests for processing and information retrieval, and other information required for the performance of computer-aided tasks. In turn, computers communicate to men the results of processing operations, the requested information, and any other messages they are programmed to produce.

The principal means for man-computer communication are manual, visual, and audio channels. The manual channel includes all mechanically operated input devices. The visual channel consists of printouts, displays and signals for visual sensing by man and electro-optical sensing by computers. The audio channels are the computer equipment and systems for recognizing spoken utterances, as well as the equipment for producing spoken output.

The choice of man-computer communication channels depends on numerous operational, human, and economic factors. Among these are the ease of use of the channel in the context of the man-computer tasks, the nature of the interaction language, the ability to maintain desired interaction rates, and the effects of the operational environment. The processing and storage requirements of the communication channel, and its cost-benefit advantages or disadvantages over competing channels, are important economic factors. An ideal channel is easy and natural to use, compatible with the total system, provides operational advantages and is cost-effective.

Most of the present man-computer communication channels are manual for man-to-computer communication and visual for computer-to-man communication. Their characteristics and design factors have been thoroughly analyzed and are widely available.¹

The use of the speech channel is still in its infancy. However, the first generation speech synthesis equipment is becoming commercially available² and the current research in computer recognition of speech³,⁴ is likely to make speech communication between man and computer technically and economically feasible in a few years. Limited capability isolated-word recognition systems are already being tested for simple control applications.⁵ Several continuous speech understanding and recognition systems are being developed in the research laboratories.

Speech has the potential for becoming a versatile man-computer communication medium. This paper discusses its attractive features, problem areas, and application criteria for this purpose. A description of specific implementations of speech recognition systems, however, is beyond the scope of this paper.

SPEECH CHARACTERISTICS

It is a natural activity for a person to mentally encode his observations, ideas, and requests into a natural language—one that he uses in his daily communications 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 statements is a many-to-one process.

The expression of a given natural language statement in speech is another many-to-one transformation—the generated acoustic signals differ from speaker to speaker as functions of their voice tract physiology, age, sex, dialect, physical condition and emotional state.

The receiver of a spoken utterance must resolve the inherent uncertainties on the basis of context and his accumulated experience and knowledge. He uses his mental "model" of the speaker, the circumstances associated with the communication, and his "world model." Various non-verbal signals by the receiver also enter the understanding process. Sometimes the uncertainty cannot be resolved at all and further clarifying communications with the speaker are necessary.

The use of natural language utterances for speech communication with computers is beset with the difficulties outlined above. Since it is not practical to provide the computer with all the contextual information required to resolve the ambiguities, some restricted form of the language must be used. For example, the vocabulary may be limited to a few

¹ Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors.
hundred words that are used with unique meanings, and rigid syntactical rules may be imposed. Further constraints may be placed on the speakers (e.g., it may be required that isolated-word speech, rather than continuous speech, be used). Despite the loss in expression power and flexibility that such restrictions entail, there are situations where speech is attractive for man-computer communication.

The following sections discuss the intrinsic characteristics and the associated attractive features and problem areas of speech input to computers. A part of this discussion is based on material which has previously appeared in literature.6,8

Message generation and encoding

The constant use of speech has made humans very skillful in communicating with others through this channel—speech can be produced effortlessly, spontaneously, at a high rate, and under almost all environmental conditions. Hence, the first characteristic of speech:

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

The associated attractive features from the point of view of man-computer communication are:

- The use of speech is familiar and convenient when the interaction language is similar to the speaker's native tongue and is easy to pronounce.
- Speech is highly suitable and the preferred channel for spontaneously generated utterances.
- Speech is potentially the highest rate versatile communication channel for computer input.
- Using speech for man-computer communication may permit the "participation" of a computer in human discussions and teleconferences.

The speech input channel loses some of its attractiveness as the language departs more and more from natural language (e.g., when words are artificial and difficult to pronounce and when abbreviations, special characters, and punctuation marks are used). Some applications are not at all suitable for speech input, such as entering graphic data. Clearly it is more natural to trace out a curve on a graphic input tablet than to read the coordinates.

The potential speed advantage of spoken input is illustrated in Table 1. However, it must be borne in mind that a high data rate is not necessarily a high information rate.

The possibility of simultaneous communication with both men and machines has interesting implications. For example, a computer and its data base may become an active participant in a conference.

Interaction with other channels

The next speech characteristic pertains to its interaction with the other channels available for man-computer communication:

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. Hence, an attractive feature is:

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

This is a very important feature of the speech channel. In numerous situations communication with the computer is not the only task. A standard example is piloting an aircraft while attempting to interact with the onboard computer.

Speech propagation

Speech propagates in the atmosphere in the form of pressure waves. These are reflected from and around objects. They can be easily changed into electrical form. The related speech characteristic is:

3. Speech propagation is omnidirectional. No free line of sight is required.

This leads to the following attractive feature:

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

Microphones with various "fields-of-view" and sensitivities can be constructed and a computer input console would need not be user-centered, but could be "stretched out" to allow optimal placement of various input-output devices and displays. The user can walk around while entering information.

<table>
<thead>
<tr>
<th>TABLE 1—Representative Data Rates for Man-computer Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication mode</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Oral reading</td>
</tr>
<tr>
<td>Random words</td>
</tr>
<tr>
<td>Spontaneous speaking</td>
</tr>
<tr>
<td>Handwriting</td>
</tr>
<tr>
<td>Handprinting</td>
</tr>
<tr>
<td>Typing</td>
</tr>
<tr>
<td>Typing</td>
</tr>
<tr>
<td>Stenotype</td>
</tr>
<tr>
<td>Touch-tone telephone</td>
</tr>
<tr>
<td>Thumb-wheel</td>
</tr>
<tr>
<td>Rotary dialing</td>
</tr>
</tbody>
</table>
tation through the speech input devices. The number and type of receivers may also vary:

- The speech target audience can vary freely from many (using loudspeakers) to a few (using earphones), in both cases low cost equipment can be used. This feature is important for security and privacy.

The easy conversion into electrical form leads to another attractive feature which has the potential of converting a conventional telephone instrument into a computer terminal:

- 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.

Among the problems associated with these characteristics is the interference of speech communications both by ambient acoustic noise and the electrical noise in the voice communication system. Another problem area is the transitory nature of speech—no hard copy is produced of speech input. A tape recording can be made but is inconvenient to use.

**Speaker characteristics**

The acoustic characteristics of speech signals depend on the structure of the speaker's vocal tract and its dynamics. Infections and other pathological conditions in the vocal tract also affect the speech quality. Articulation and timing are influenced by fatigue. Unusual emotional conditions can change the normal speech characteristics, such as the pitch and speaking rate. Hence the characteristic:

4. _Speech contains a great deal of information about the speaker._

This characteristic leads to two attractive features and two problem areas in the application of speech for man-computer interaction:

- The use of speech allows checking the speaker's identity for access control purposes.
- The use of speech has the potential for monitoring the physical and emotional state of the user.

To implement speaker identification capability, carefully chosen speech samples can be analyzed and a set of parameters computed and stored. To authenticate a person's identity the person speaks a predetermined sentence which is also analyzed and the extracted parameters are compared with the stored ones. Considerable research is in progress on this topic. An ability to monitor the operator's physical or emotional state is important in man-computer tasks where the operator's actions, or inactions, may have drastic consequences, such as in air traffic control applications.

The problem areas associated with these features have to do with the complications in the design of speech recognition systems caused by speaker-to-speaker variability, and the variability in voice characteristics of a given speaker. It may be necessary to train the system to recognize each individual speaker's voice characteristics and to store such information in the system. Affected are the required amounts of storage and processing.

**Environmental influences**

Speech generation and propagation are both affected by the environmental conditions. Some of these, such as temperature, humidity, or insufficient working space, affect the speech generation only indirectly (e.g., by accelerating the onset of fatigue and emotional conditions); others have more direct effects. The associated speech characteristic is:

5. _Speech production is affected by mechanical forces on the speaker and composition of the atmosphere._

Experiments have shown that both vibration and acceleration affect speech intelligibility. Changes in the atmosphere, such as the presence of helium in submarine systems, also affect speech by changing the pitch and intelligibility. However, weightlessness does not appear to have any effects.

Any ambient acoustic noise in the environment will interfere with the speech signal. This condition may be quite acute in systems containing equipment in operation (aircraft engines, teletype terminals) or other speakers. Among the techniques available for alleviating the noise interference problem are noise-cancelling microphones, special signal processing techniques, and the use of specially selected, high-intelligibility vocabularies.

**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 in voice response systems and by the number of firms actively engaged in marketing these systems.

The attractive features of the use of speech for computer-to-man communication include the following:

- Speech is a natural way for humans to receive communications from others. It is compatible with the use of speech as a computer input channel.
- Several spoken messages can be received and comprehended simultaneously.
- Spoken messages can be received without interrupting the use of the manual or visual channels, in motion, or in total darkness.
- In receiving spoken output from the computer, the operator can be in an arbitrary orientation relative to the computer, some distance from the computer, or behind a barrier.
- Any number of listeners can receive the spoken message from the computer simultaneously.
- Speech reception by humans is not appreciably affected...
by weightlessness, vibration, or mechanical forces on the
listener.

There are also some problem areas. For example, the rate
of receiving spoken messages is much slower than through the
visual channel. The transient nature of speech requires that it
be recorded on a tape if hard copy is required, but in this form
it is not readily scannable by the human operator.
The ambient acoustic noise interferes with the reception
and comprehension of spoken messages from the computer
when they are broadcast or sent over a telephone instrument.
However, the human auditory system is rather remarkable in
its ability to select out and concentrate on a specific message
and ignore others (this is the so-called “cocktail party”
effect).

IMPLEMENTATION OF SPEECH INTERFACE

The design of an effective yet economical man-computer
interface is a complex process that must take into account the
nature of the tasks to be performed; the human roles,
capabilities, and shortcomings in performing these tasks; the
task performance environment; and the capabilities of the
interface equipment. The task characteristics and the human
roles determine whether a speech interface is suitable. The
associated system design and performance requirements
determine whether a speech interface will be technologically
and economically feasible.

Design criteria

The principal roles of a human operator in a man-computer
system are: decision maker, problem solver, controller,
monitor, retriever or inquirer, and sensor or transducer. The
most demanding of these is the human role as decision maker
in real-time command-control systems where his performance
is especially affected by the criticality of the consequences of
the decisions, the diversity of the decisions to be made, and
their dynamics.
The following considerations influence the design of a
man-computer interface.

• Nature, time characteristics and variability of the tasks.
• Intensity level of the task performance and the operator’s
  response requirement.
• Input and output loading of the operator.
• Operator’s and system’s physical state during task
  performance. Operator’s physical safety and other stress
  conditions.
• Operator’s level of isolation when performing the task.
• Environmental conditions.
• Training and skill level of the operator.

Based on these and the discussion of speech characteristics
in the previous section the suitability of a speech interface for
performing a given set of man-computer tasks can be
evaluated.

The speech understanding and recognition systems used to
implement a speech interface are characterized by a series of
design features which reflect the acoustic and linguistic
processing aspects of these systems. These design features
have been discussed for continuous speech understanding
systems in detail by Newell, et al. Included are the following:

• Vocabulary size and syntactical structure.
• Number of speakers, their dialects and speaking habits.
• User training and system tuning; the degree of speaker-
  independence.
• Ambient noise environment and the transducer charac-
  teristics.
• Requirements for and availability of contextual “world
  model.”
• Recognition error rate.
• Response time.

These establish the requirements for the system hardware—
the special-purpose acoustic signal processing equipment and
the general-purpose digital computer for pattern matching
and linguistic processing. Tradeoffs can be performed between
the various sets of the characteristics, especially between
those involving the interface capabilities (vocabulary, syntax,
speaker independence), performance (error rate, response
time), and equipment (processing power, storage capacity,
cost).

State of the art and potential applications

As mentioned previously, isolated-word speech recognition
systems are already being offered on the market and are being
tested. Typically, these systems can be trained to recognize
utterances employing small vocabularies and highly restrictive
syntax. The pause between words must be greater than .2
seconds.

The implementation of continuous speech recognition and
understanding systems is much more difficult. One of the
problems is the absence of word boundary indications in the
acoustic signals and the dependence of the signal representing
a word on the predecessor and successor words. This problem
makes the use of linguistic and semantic information a
necessity. The variability of individual speaking and articula-
tion habits further complicates the recognition tests. Table II
shows the recognition accuracy that has been achieved by
various experimental and prototype speech recognition
systems.

There are a number of man-computer application areas,
mainly in the “user's hands busy” category, where a speech
interface for computer input-output could provide significant
performance improvement. Among these are:

• Computer-aided fault diagnosis and isolation; computer-
  aided instruction; medical diagnosis; performance of
  scientific experiments.
• Data input in taking inventory, making observations, or
  tracking moving targets.
TABLE II—Speech Interface Performance Data

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Capability</th>
<th>Correct</th>
<th>Recognition percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vicens (1969)</td>
<td>54 isolated</td>
<td>1</td>
<td>98-100</td>
</tr>
<tr>
<td></td>
<td>54 isolated</td>
<td>10</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>580 isolated</td>
<td>1</td>
<td>91</td>
</tr>
<tr>
<td>Yilmaz (1971)</td>
<td>16 isolated</td>
<td>10</td>
<td>99</td>
</tr>
<tr>
<td>Hill (1969)</td>
<td>16 isolated</td>
<td>12 unknown</td>
<td>78</td>
</tr>
<tr>
<td>Medress (1972)</td>
<td>100 isolated</td>
<td>5</td>
<td>94</td>
</tr>
<tr>
<td>Glenn (1971)</td>
<td>10 isolated digits many</td>
<td>&gt;99</td>
<td></td>
</tr>
<tr>
<td>Doddington (1973)</td>
<td>10 continuous digits many</td>
<td>&gt;99</td>
<td></td>
</tr>
<tr>
<td>Tappert &amp; Dixon (1971)</td>
<td>296 continuous</td>
<td>several</td>
<td>75</td>
</tr>
</tbody>
</table>

- Monitoring computer-controlled processes.
- Controlling teleoperator systems and robots.

Other application areas are computer data base management, information retrieval, and computer-aided programming. The ultimate application is the perennial inventors' dream—the speech-operated typewriter for unconstrained language. It is not likely that such a device can be realized in the next decade, or even this century. Restricted versions, however, are likely to be implemented.

CONCLUDING REMARKS

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

Despite the attractive characteristics of speech described in this paper, its use in a particular man-computer task makes sense only when its use is natural for performing the task and compatible with the environment. Hence the nature of the interaction involved must be thoroughly analyzed before committing to the use of a speech interface. However, together with other modes of man-computer communication, the speech-based interfaces can help an operator to concentrate on the tasks he is performing rather than on operating the interface equipment.

ACKNOWLEDGMENTS

The author would like to thank his colleagues at The Rand Corporation, Alan S. Hoffman, Robert C. Gamnill, Gabriel F. Groner and Thomas F. Lippiatt for their helpful suggestions.

REFERENCES
