Speech Recognition, the PC, and SAS® Software
Linda S. Fineman
SAS Institute Inc.
Cary, NC

ABSTRACT
Over the past four decades, researchers have often thought that the solution to the problem of machine recognition of speech is "just around the corner." Realization of the dream has been a long time coming, but today speech recognition is available for limited applications on both large and small computers. At SAS Institute, developers are building a speech interface to the SAS System for personal computers using the IBM Voice Communications Adapter.

AN INTRODUCTION TO SPEECH PROCESSING
Machines that talk and listen to people have long been a staple of movies, television, and science fiction writing. It may surprise the typical viewer or reader to learn that the technology being portrayed is not possible today. Fluent man-machine communication in spoken, natural English is an unrealized dream.

Background
Over the past four decades, researchers have often felt that the solution to the problem of computer speech recognition is "just around the corner." Forty years ago the advent of the sound spectrograph opened the way to a complete understanding of speech acoustics. Twenty years ago digital computers arrived on the scene and gave a new direction and motivation to the field. Ten years ago researchers thought that the tools of artificial intelligence would bring the speech recognition problem to a speedy solution.

Where are we today, nearly half a century after the first breakthroughs? A great deal is now known about how to build computer systems that can recognize and respond to a restricted set of spoken commands. A variety of practical recognition systems exists in both the laboratory and the commercial world. Speech recognition has been available for limited applications on large computers for several years. Recently, the technology has been adapted for use on the PC and a number of modestly priced recognizers have reached the market and are in use.

Realization of the dream has been long in coming for several reasons: overly high expectations, a lack of meaningful performance measures, high prices, and difficulty in developing applications. In many cases, the applications chosen were not economically sound; often, they addressed problems that were already being handled successfully by other methods.

Speech technology still has a long way to go, and both the current quality of artificially produced speech and the power of recognition devices leave much to be desired. A greater understanding of the entire speech process is needed before a recognizer whose performance will approach human ability can be built. Advances in speech processing will require contributions from a wide spectrum of disciplines: phonology, physiology, psychology, computational linguistics, and artificial intelligence, among others. Progress will not be possible if the research is conducted in isolation.

Motivations for the Work
Speech is an ancient and universal form of human communica-
Since the systems are recognizing known words by known speakers, the major source of variability is the time frame. The same word may (and will) be spoken at different speeds. Unfortunately, different speaking rates do not result in a linear speed change in all parts of a word; the vowels are much more sensitive to speed variation than are consonants. One current technique of time adjustment called dynamic programming uses a mathematical process to compare a word spoken slowly with the same word said more quickly and then produces a match. Dynamic programming is a feature of most medium- and high-performance speech recognizers.

The Future

Currently, words are generally the smallest units of speech recognition. Work is progressing in the area of phoneme recognition, but it may be some time before systems routinely incorporate this facility.

Another future development concerns the recognition of several words or phrases in succession. Connected speech systems relieve the speaker of the burden of pausing after each uttered word. The software is designed to compute the word boundaries in speech input automatically. Connected speech, however, should not be confused with the continuous speech used in everyday human conversation. The goal of fluent conversational speech cannot be achieved with the current level of linguistic knowledge.

Work is also underway to develop systems that are speaker-independent. Such systems accept voice input commands from many users, without requiring a set of training templates for each individual. Speaker independence faces serious obstacles: in the United States alone, over 200 regional dialects have been identified. Moreover, variations in pronunciation always exist among speakers, whether or not they share a common dialect.

The implementation of large vocabulary systems is another key area of research. Experience suggests that the algorithms used for small speech tasks cannot simply be scaled up to handle large vocabularies. Recognizing words from a vocabulary comparable in size to that of spoken English would be a major technological accomplishment.

Future directions for speech recognition research will be driven by the shifting needs of the user community as well as by advances in technology. It is difficult to predict whether large vocabulary, speaker-independent, and continuous speech recognition systems will be here in the foreseeable future. The features such systems should possess to be useful and the strategies that will ultimately prevail are currently matters of debate.

Speech Output

One common approach to speech synthesis is the storage and recall of human-originated speech. This is the method used to make digital recordings. In the digitizing process, the system samples the voice qualities of the incoming audio signal and stores the results digitally.

It is also possible to produce intelligible synthetic speech out of computer-generated sounds or phonemes. The engineer, an applications programmer, can assemble the voice elements required to create a particular sound, customizing characteristics such as pitch, amplitude, and speaking rate. The result is readily comprehensible speech, though with an unmistakable machine accent.

Synthesis of text-to-speech introduces another major level of complexity, in order to process unrestricted text, a number of steps must be carried out in a satisfactory way. First, orthographic text must be interpreted; for example, we pronounce the acronym "USA" as a sequence of three words, but we do not expand the word "SAS." Second, the spelling must be converted to pronunciation—a nontrivial task in English. Many proper nouns will require special analysis, and the system must be able to calculate the appropriate stress assignments and phrase demarcators.

The speech output produced by modern synthesizers still only sounds natural when a human utterance has been digitized and reconstituted in whole. Concatenating words, though cheaper now, sounds much as it did twenty years ago.

MARKET SURVEY

The sales of speech technology products reached $20 million in 1985, an increase of $6 million since 1983. The industry has shown a dramatic shift in focus over the past few years: products for the IBM PC and its compatibles now dominate the field. Until very recently, most voice systems were large, expensive, stand-alone devices. Today, the biggest seller is the plug-in PC speech board.

PC Speech Boards

Not all of the available hardware for the PC can be discussed here, due to the quantity of product offerings. The following systems are representative of other speech boards on the market.

The first plug-in speech board introduced to the market was from Texas Instruments in 1982 for the TIPC. Texas Instruments' current entry is the Texas Instruments Speech Library, priced at $995. It is a speaker-dependent discrete utterance recognizer with integrated telephone functions and speech synthesis capabilities. A telephone manager is available for an additional $350.

A less expensive speech recognizer is marketed by Interstate Voice Products of Orange, California. The Vocalink SRB-LC costs $395 and is strictly a voice-activated keyboard utility. It allows users to activate keystrokes by voice and is best suited for simple applications like moving through a spreadsheet.

In late 1984, NEC Corporation introduced the SAR-10 Voice Plus, a voice input and output speech board. Priced at $1,450, the SAR-10 provides both speech recognition and audio response capabilities.

Votan, of Fremont, California, is a longtime leader in the voice technology market. The VPC 2000 is unusual among speech boards in its ability to recognize connected speech. The system costs $1,695 and includes speech synthesis capabilities.

Other board suppliers include IBM, ITT, Dragon Systems Inc., of Newton, Massachusetts, and Kurzweil Applied Intelligence, of Waltham, Massachusetts.

Large Vocabulary Speech Systems

Kurzweil Applied Intelligence developed the Kurzweil Voice System (KVS), a speech recognizer capable of understanding a vocabulary of up to 1,000 utterances. The KVS is a self-contained box that connects to an IBM PC/XT or PC/AT through an RS-232 interface. While the KVS is confined to PC compatibles at this time, the RS-232 port should allow it to become a front-end product for other machines in the future. The system costs approximately $6,500.

Voicescribe-1000, introduced this past spring by Dragon Systems Inc., also has a 1,000-word vocabulary capacity. In addition to this feature, the Voicescribe-1000 provides limited natural language recognition. It can be used for dictating notes or simple documents with a limited vocabulary, and the user is allowed to customize vocabularies.
Speech Systems Inc., of Tarzana, California, is working on a large vocabulary, connected speech recognition system based on the recognition of phonemes. The recognizer, called the Phonetic Engine, will come with a 20,000-word phonetic dictionary. Applications include data entry, dictation, and word processing. The first version of the engine will run on the IBM PC and its compatibles, but future models may be ported to other machines.

Connected Speech Recognition Systems
As mentioned earlier, both Votan and Speech Systems Inc. are involved in the design of connected speech recognition systems. A third company, Interstate Voice Products, is also active in this area of development.

The Interstate Voice Products Series 4000 connected speech recognizer is a VLSI-based peripheral that can provide a speech recognition capability for virtually any machine. The Series 4000 was originally developed by Verixx, a former division of Exxon Enterprises. In mid-1985, Interstate acquired the rights to manufacture, service, and sell the system. The recognizer is speaker-dependent and is designed to perform well in both noisy and quiet environments. It comes in a compact unit that can be mounted on a wall or shelf.

A PC VOICE SYSTEM
SAS Institute began its speech recognition effort a year and a half ago, with the purchase of a speech board from Texas Instruments. Since then, an IBM speech board has been acquired, and this is the device being used in the present implementation.

Basic Components
The IBM Personal Computer Voice Communications Option (VCO) is an adapter card that fits into an expansion slot on any of the IBM PC, PC/XT, and PC/AT systems. Memory requirements vary depending on the application, the DOS level, and the functions being used. The adapter consists of a specialized microprocessor, memory, and support hardware.

The VCO and its applications software support the following functions: discrete utterance speech recognition, text-to-speech synthesis, modem communications, telephone monitoring, and audio record and playback. The software interface consists of several routines that are executable on the main PC processor. An application program can access these routines by issuing a DOS interrupt call.

The Recognizer
The discrete utterance function of the VCO provides the capability to recognize previously trained utterances consisting of a word or short phrase. Each input utterance must be followed by a brief pause in order to be recognized. Connected or continuous speech cannot be processed.

The recognizer is speaker-dependent, so each user of the system must create and store a personal template file. This file is a list of voice patterns corresponding to the words in the vocabulary and is produced during training. An application may require more than one vocabulary; conversely, one vocabulary may be used by several different applications.

Defining a Language
The VCO requires a language definition for each speech-driven application. As with English, a language consists of a vocabulary and a grammar. The vocabulary is the set of utterances that can be recognized, and the grammar is the set of rules that impose order on the utterances of the language. The grammar limits what utterances can be spoken at any point in the input stream, and this improves the accuracy and speed of recognition. In this system, the grammar rules are in the form of regular expressions.

Each grammar rule can specify an output string representing a list of keystrokes that an application can retrieve at run-time. This feature can be used to define what actions should be taken when a particular utterance is recognized. Processing the output string is one means by which an application can act on an utterance.

Provided with the VCO is a stand-alone program that compiles a language definition in text form into a binary object file for use by the speech recognizer.

Managing Vocabularies
Once a language has been defined, the users must train the recognizer to recognize the way they speak the words in the vocabulary. Training involves saying the same utterance one or more times, while the system listens and constructs templates for the utterance. A model, or prototype voice print, is then derived from the collection of templates. The model becomes a member of the vocabulary and is assigned an integer ID.

The size of a model varies somewhat depending on the duration of the trained utterance. Typically, a model will require between 200 and 400 bytes of storage. Models may be built that are adaptable so that templates can be added over time to enhance recognition performance. Models that are adaptable require about 50% more memory.

Recognition response time increases as a function of the size of the vocabulary. To ensure a healthy performance level for the recognizer, a restriction must be placed on the number of utterances valid at any point in time. The maximum number of active utterances on this system is 64. Although any single vocabulary is limited to 64 elements, applications are permitted to switch active languages as processing proceeds. Loading or activating a new language always requires extra interrupt calls to the speech board and often involves additional disk I/O, so execution may slow down.

Listening and Recognition
Speech recognition really comprises two independent parallel processes: listening for utterances and recognizing utterances. The two activities can operate in parallel, by allowing the system to accept speech input at any time, or in serial form, to regulate the timing of utterances. In the latter mode, users can speak only when a command is required.

The flow of the speech recognition process is as follows:

1. The listening process is initiated.
2. The user inputs a sequence of utterances.
3. The listening process translates each utterance into a template (voice print) and places it on a queue.
4. The recognizer retrieves the oldest template from the queue and tries to match it against the current set of active vocabulary models. If a pattern match occurs, the utterance is translated into an integer ID representing an element in the active vocabulary.

An apt analogy for the processes of listening and recognizing is the operation of a keyboard. Starting and stopping the listening process is similar to locking and unlocking the keyboard. Unlocking the keyboard allows the user to type ahead, just as leaving the listening process running permits the speaker to talk ahead. Locking the keyboard to allow data entry only at appropriate times is similar to stopping the listening process. An utterance can be compared to a keystroke, and the talkahead queue is like the keyboard buffer. Recognizing an utterance is like translating a keystroke to a character.
A SPEECH INTERFACE TO SAS SOFTWARE

Work is in progress to implement a speech input facility for the SAS System for PCs. The goal is to enhance the environment of SAS by providing the capability to issue verbal commands and dictate program text.

Talking to the Display Manager

Nearly every PC user has experienced the frustration of forgetting which functions have been assigned to which function keys. In the SAS Display Manager System, each function key can be set to execute one or more display manager commands or SAS statements. Yet, there is nothing mnemonic about the labels on function keys and no way to guess probable key assignments (short of using the KEYS command to display the KEYS window and to view the current function keys settings).

Using voice commands is a natural solution to this problem. It is much easier to remember to say "output" to display the OUTPUT window than to remember to hit the correct function key. One feature of the SAS speech interface is the ability to speak commands such as "help," "log," "zoom," and so on. It should be noted that in all cases speech input is an alternative, not a replacement, for typed commands. When to speak and when to type are matters of personal preference.

Voice commands can be used to call and manage any primary or special window. Speaking "tabname," "notepad," or "titles" calls up those special windows, and utterances such as "top" and "bottom" can be used to scroll the contents. In theory, any display manager command that can be executed from the command line or with a function key can also be spoken.

Inserting Text

The SAS speech interface provides a facility for dictating portions of program text. Common SAS keywords, such as DATA and PROC, can be spoken rather than typed. To use this facility, you first move the cursor to the appropriate position in the PROGRAM EDITOR window. The text is printed starting at the designated location.

Due to the 64-word size limitation, one vocabulary is not sufficient to handle the complete set of SAS keywords. Specialized vocabularies have been created for several common procedures. For example, the list of vocabulary words for the CHART procedure includes the utterances "hbar," "vbar," "block," "pie," and "star." During a SAS session, vocabularies are activated as needed. The PROC CHART vocabulary is loaded and made active following recognition of the utterance "chart."

Training Facilities

Special screens have been defined to facilitate the process of training and refining vocabulary models. Since the recognizer is speaker-dependent, it is necessary to create a personal template file for each vocabulary that you plan to use. The procedure has been designed so that a vocabulary can be trained all at once or over as many sessions as desired. The training routine can be halted or restarted at any point.

Once a vocabulary has been trained, you can test its recognition performance. A special window displays the words in the vocabulary and prompts you to speak. The system displays a confidence score (between 1 and 100) each time an utterance is recognized. This number is a measure of how well your current pronunciation matches that of the stored model. A running average of the recognition scores is also displayed.

If an utterance is not being recognized with an acceptable degree of accuracy, it may be beneficial to refine the vocabulary model. This is accomplished by retraining the utterance and using the old model and the new template to derive a more representative voice print for the word. A window is provided for this process.

In some cases, refining a model is unsuccessful, or an utterance may require nontrivial changes. The best approach under these circumstances is to destroy the old model and create a new one from scratch. This capability is accessed through another special window.

CONCLUSION

Speech recognition systems have finally attained a level of performance that makes it possible to develop meaningful applications. The current generation of recognizers is more reliable and less expensive than ever before. Now is the perfect time for software developers to consider the appropriateness of speech input and output for their applications.

SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.

REFERENCES