Automated Speech Recognition in Medical Applications[1,2]

Michael A. Grasso[3]

Despite considerable advances in computer technology over the last 20 years, the keyboard and video display are still the principal means of entering and retrieving data. As the use of computers increases, however, the need for alternative ways of interacting with the computer grows as well. This demand was fueled by a need for an intuitive human-machine interface to accommodate increasing numbers of nontechnical users, since limitations in the interface are still obstacles to the widespread acceptance of computer automation[1]. In addition, in some situations the usual methods of interacting with the computer are impractical - for example, when the hands are otherwise occupied, as in microscopy. An approach that addresses both the need for ease of use and the need to keep the hands free is the use of automated speech recognition.

Research in automated speech recognition dates back to the 1950s, but advances in computer technology have resulted in successful commercial systems only recently. Despite these advances, true natural language processing is still several years away, and a successful speech-driven system must allow for limitations in the current technology. On the other hand, a recent article identified 70 manufacturers of products with the ability to identify spoken words[2]. These systems can be classified according to speaker dependence, continuity of speech, and vocabulary size[3,4].

Speaker Dependence

Most systems with moderate to large vocabularies require some type of user training to ensure a high accuracy rate. Speech recognition systems typically use some form of pattern matching, in which the computer compares spoken words with predefined templates to find the best match. Before this can occur, the user must create templates by reading each word in the vocabulary two or three times. Representative word phrases may also be read aloud, to identify how certain words will be spoken in context. A speech model consists of all the templates for a given vocabulary. Each operator of a speaker-dependent system must create a speech model by training the system to recognize his or her way of saying every word in the vocabulary. Depending on the vocabulary size, training can take from a few minutes to several hours.

Speaker-independent systems use generic models to recognize speech from any user. Generic models are created by combining existing templates from a variety of speakers. This approach is

1 This research was supported by grant 1R43RR07989-01 from the National Center for Research Resources.
2 This article also appeared in M.D. Computing, Vol. 12, No. 1, 1995.
3 Address correspondence to Michael A. Grasso as MikeGrasso@umbc.edu.

advantageous in that it does not require individual operators to train the system to recognize their voices. However, because the templates are not user-specific, accuracy rates are usually lower with this approach.

An alternative is the speaker-adaptive approach, which uses a generic model to eliminate initial training and then automatically generates user-specific models for each operator over time. Although initial training is eliminated, recognition accuracy is diminished until the system develops an adequate user-specific model.

Continuity of speech

Continuous speech systems can recognize words spoken in a natural rhythm. Although this approach seems more desirable at first glance, continuous speech is harder to process because of the difficulty of identifying word boundaries - as in "youth in Asia" and "euthanasia." Variability in articulation, such as the tendency to drop consonants or blur distinctions between them - as in "want it" and “wanted" - can result in further misunderstanding. To increase accuracy, speech models for continuous speech systems include information on representative word combinations and context rules.

Isolated word systems require a deliberate pause between each word. Pausing for 0.1 second after each word is unnatural and can be tiring. However, accuracy rates are usually higher with isolated word systems than with systems using continuous speech. Isolated systems work best with vocabularies that consist mainly of individual command words.

Vocabulary size

The vocabularies of various speech recognition systems range from 20 to 40,000 words. Large vocabularies cause difficulties in maintaining accuracy, but small vocabularies restrict the speaker. In addition, large vocabularies are likely to contain ambiguous words, which in speech recognition systems are words with pattern-matching templates that the computer will treat as similar -- such as the words "tree" and "three."

Grammars can impose constraints on allowable sequences of words, according to a set of rules regarding context and phrase structure. A tightly constrained grammar is one in which only a small number of words can legally follow any given word. Keeping the list of candidate words small can increase recognition accuracy and decrease latency time during pattern matching, especially with large vocabularies. However, too many grammar rules can reduce the naturalness of communication.

Building Speech Recognition Applications

Most speech recognition systems include an interface card that can be installed into the computer workstation's I/O bus. This card contains the digital signal processor and analog-to-digital converter used to recognize spoken words. More advanced cards contain an on-board processor and memory to off-load processing requirements from the computer. A noise-canceling microphone is used to capture speech input. Headphones may be included for computer-generated voice responses, if the system has text-to-speech capabilities.

A number of supplemental files are needed for a speech recognizer to work. The vocabulary file includes the words the system recognizes and the grammar rules that identify the situations and order in which words can be used. Speaker-dependent systems also have a separate voice file for each user. Finally, a program is written in a high-level language, such as C, interfaces with the speech recognition engine. An application programming interface in the form of a linkable object module library, is provided for this purpose. The task varies with different systems. The description that follows is based on the Verbex 6000 AT31 Voice Input Module[5]. Although the interfacing process can be very involved, a minimal program will contain three parts: initializing, polling, and handling speech events.

"*Initializing*" refers to initializing the recognition engine and download the necessary vocabulary and voice files to it.

"*Polling*" refers to a loop in the main function that continually searches for speech events. With Microsoft Windows programs, this can be added to the main loop that polls for window messages and processes them. The event handler function is implicitly called whenever polling identifies a spoken word. "*Handing speech events*" means to write an event handling function that will be used to process word recognition events. This function will be implicitly called whenever the program polls for speech events. The function will most likely consist of a series of conditional branches to match the spoken word passed to the handler with possible words from the vocabulary. For example, in Microsoft Windows the event handler function is similar to a callback function that processes messages sent to a window.

Most systems provide an alternative method for interfacing with their speech recognition engines. Instead of using an application programming interface from a high-level language, they may support speech as an extension of normal keyboard input. With this approach, spoken words recognized by the system are inserted into the keyboard input stream just as though they were typed on the keyboard. The input can then be received and processed by an application program in the same way that the program would receive and process typed input. Although this is a quick method of developing speech-driven applications, the process has definite limitations.

In the first place, adding speech capabilities to a preexisting system can be awkward. It is similar to adding mouse support to normal character-based applications through a mouse driver that simply inserts text into the keyboard input stream. There is a lack of coupling between the input device and the application. Information from speech input is sent to the application, but the application cannot communicate with the input device, or even know it is there. Simply adding speech to an existing user interface can decrease the integrity of the system[6]. It is best to design speech-driven interfaces from scratch.

Current Uses of Speech Recognition Technology

The main uses of speech recognition system that have been described in the literature are for template-based reporting, natural language processing, integration of speech with other methods of input, and data entry in environments where the hands are busy.

Template-Based Reporting Systems

These system are popular in radiology, pathology, endoscopy, and emergency medicine. They have large vocabularies, recognize discrete speech (meaning that the speaker pauses between words), and are speaker-adaptive systems designed to generate template-based reports using fill-in forms, trigger phrases, and free-form speech. Turnaround time is decreased and accuracy is increased by eliminating the need for dictation and transcription by clerical personnel.

Reactions to this approach has been mixed. For autopsy pathology, it has been noted that a greater degree of computer literacy is required and that the need for typed input is not eliminated[7]. When applied to endoscopy, the process took longer than standard dictation and nevertheless collected less information[8]. These problems were attributed to the fact that therapeutic endoscopic procedures are complex and not suited to a template-based reporting format. The free-form speech method, in which single words are printed as they are spoken, was found to be too slow to be useful[9]. This was probably due to increased computational requirements associated with larger vocabularies (up to 40,000 words). On the positive side, the formality of the process seemed to provide other benefits. One researcher noted that 80% of emergency room reports were adequately completed with a speech recognition system, as compared with 30% when reports were dictated or handwritten records were used[10].

Natural Language Processing

A group at Stanford University is studying the use of speech as an improved interface for medical systems. Initial work focused on the development of three prototype speech-driven interfaces[11] along with research on how clinicians would like to speak to a medical decision-support system[12]. It was noted that the use of template-based dictation with fill-in forms worked will only when the documentation task was limited to a few standardized reports. Template-based reporting may be inadequate in clinical domains, because the required documentation is less standardized. At the same time, current speech recognition technology does not permit the processing of free-form natural language. Methods that circumvent shortcomings in the current technology while maintaining the flexibility and naturalness of speech are being explored.

Three prototype systems were developed that were more complex linguistically than template-based reporting, and the typical entries could not easily be selected from a simple presentation of menus. The systems had a speaker-independent vocabulary of more than 38,000 words using continuous speech. In addition, Windows-based graphics were used as control and feedback mechanisms for the various grammatical rules in the system. This use of graphics to display the visual context in which the various grammatical rules applied was shown to improve the speed and accuracy of recognition except when the grammar was complex. Overall the evidence suggests that graphical guidance can be used effectively when the vocabulary is sufficiently constrained.

Speech with Other Input Devices

A different approach for speech recognition is to develop systems that use speech in combination with other input devices. The goal in this case is not to replace the keyboard or mouse but to simplify or accelerate the input process.

Data collection can be a bottleneck in many computer applications. One system, design to assist in the collection of stereological data, combines speech input with a digitizing pad[13]. Each data set consists of an object name, recorded by voice, followed by X and Y coordinates, entered with a digitizing pad. The system is used for boundary analysis and histomorphometry of bone and skin. It has a small speaker-dependent vocabulary (less than 50 words) for object names and voice commands, and recognizes discrete speech. This simple interface has low computational requirements and therefore a high chance of success. It allows the user to choose between eight control words and between 20 object names. The combination of speech and a digitizing pad was shown to accelerate the data collection process.

A system developed at the Massachusetts Institute of Technology uses speech as an auxiliary channel to support window navigation. With the product, Xspeak, window navigational tasks usually performed with a mouse are controlled by speech instead[14]. Xspeak was developed with the assumption that most successful speech recognition systems have small vocabularies, are speaker-dependent, and use discrete speech, and that speech input is more valuable when it is combined with other input devices.

The X Windows system uses a spatial metaphor to organize applications on a monitor in three dimensions. However, it uses a two-dimensional device for window navigation: the mouse. When there are many overlapping windows, it can be difficult to reach some applications directly with the mouse. Xspeak was therefore designed to improve navigation in this type of environment. Each window is associated with a voice template. When the word represented by a template is spoken, the window is moved to the foreground and the mouse pointer is moved to the middle of the window. Initial testing revealed that while speech was not faster than the mouse for simple change-of-focus tasks, the advantage shifted toward speech if the window needed was partly or completely hidden. Another observation was that the users most inclined to choose speech input increased the number of overlapping windows or the degree of overlap.

Data Collection When the Hands Are Busy

Several applications use a speech-driven approach to facilitate the collection of data in a hands-busy environment. One group studied the feasibility of using speech recognition to record clinical data during dental examination[15]. Systems of this type would eliminate the need for a dental assistant to record results. The speech method was shown to be slower. However, when the time needed to transfer results recorded by the dental assistant into the computer was considered, the speech method was faster. Speech input also had more errors, although the difference was not statistically significant. Overall the study suggested that speech recognition may be a viable alternative to traditional charting methods.

Another group designed a speech interface for an anesthetist's record keeping system[16]. Anesthetists are responsible for recording information on drugs administered during medical procedures. A long interval between an event and its recording can compromise the completeness and accuracy of the manual record. However, voice entry allows collection of the data during the medical procedure, while the anesthetist's hands are busy. The system used a vocabulary of around 300 words. Preliminary testing showed an accuracy rate of 96%, even in a noisy operating room.

Hands-busy data collection has also been applied to the analysis of bone scintigraphic data[17]. Such diagrams are analyzed to study metastases of malignant tumors. A speech system was developed to allow doctors to enter the results of image readings into the computer while looking at the images instead of the terminal. In 580 voice-entered reports, response time was shortened in comparison with dictation or writing by hand.

To test the feasibility of using speech recognition to permit hands-free and eyes-free collection of data during laboratory research[18], my colleagues and I developed a speech-driven system for recording histopathological data during microscopy. Only speech input and computer-generated voice responses were used with a continuous-speech, speaker-dependent system with a vocabulary of 900 words, based on the Pathology Code Table[19]. The overall accuracy rate was 97%. However, additional work is needed to reduce the training requirements and improve audible feedback.

A Comparative Study

Successful development of a speech recognition application requires an adequate understanding of the limitations in the current technology. True natural language processing by computers is still several years away and there is little conclusive evidence that speech recognition is superior to the keyboard or other input devices. In a study comparing data entry with the keyboard, light pen and speech recognition, 20 intensive care nurses used each approach to enter data at the computer and were asked to evaluate each method[20]. Keyboard input was selected as the quickest, easiest, most accurate, and most desirable method of data entry. In other words, under normal circumstances, keyboard input is preferred to speech recognition.

This finding highlights an important caveat: speech recognition is best used when its advantages offset its limitations. Consider the main reasons for choosing on of the four types of speech-driven applications I have described: template-based reporting eliminates the need for dictation and transcription by clerical personnel and is easier for the physician than direct keyboard entry, natural language processing offers an intuitive and spontaneous form of computer interaction that is also easy for the user, speech recognition integrated with other devices offers increased functionality and thus can simplify or accelerate data entry, and "hands-busy" speech recognition provides an alternative method of input when the user's hands are otherwise occupied. Notice that even though template-based reporting can theoretically reduce personnel requirements, the same end can been achieved with keyboard-based templates instead of speech. Natural-language processing and speech recognition plus other devices expedite data entry by overcoming certain limitations of standard input devices or combining speech with other devices for increased functionality. Another caveat is that speech is a new approach with unique requirements that need to be addressed at the beginning of the design process; the

problems associated with speech recognition are unlikely to be resolved if they are not considered from the start.

On the other hand, the comparative study used the same data-entry format (single-character entry) for input with the keyboard, light pen, and speech. Light pens are more practical than the keyboard when the user is selecting items from a list. Speech input is better suited for entering command words or entire phrases. It is not surprising, therefore, that the keyboard was preferred.

Conclusion

Automated speech recognition can address two issues in the human-computer interaction: the demand for ease of use and constraints on the user's ability to work with the keyboard or mouse. The technology is still limited, however. Success involves using a small to medium-sized vocabulary (less than 1000 words) and well-defined grammatical rules, which limit the number of words that are valid at any one time. Even dictation systems with large vocabularies work best when the operator uses only a subset of key words. A small vocabulary and tightly constrained grammar are especially important when a system is intended to support continuous speech. Finally, a successful system must also provide some training mechanism (speaker-dependent or speaker-adaptive) to increase recognition accuracy.

References

1 Landau JA, Norwich KH, Evans SJ. Automatic Speech Recognition - Can it Improve the Man-Machine Interface in Medical Expert Systems? Int J Biomed Comput 1989; 24:111-117.

2 1994 Buyer's Guide. Voice Processing Magazine 1993; 5(12):35.

3 Bergeron B, Locke S. Speech Recognition as a User Interface. M.D. Comput 1990; 7(5):329-334.

4 Peacocke RD, Graf DH. An Introduction to Speech and Speaker Recognition. Computer 1990; 23(8):26-33.

5 Verbex voice application software support library, Verbex Voice Systems, Inc., Edison, New Jersey, 1990.

6 Wulfman CE, Isaacs EA, Webber BL, Fagan LM. Integration Discontinuity: Interface Users and Systems. Tech. Report KSL-88-12, Knowledge Systems Laboratory, Stanford University, Palo Alto, California, 1988.

7 Klatt EC. Voice-Activated Dictation for Autopsy Pathology. Comp Biol Med 1991; 21(6):429-433.

8 Massey BT, Geenen JE, Hogan WJ. Evaluation of a Voice Recognition System for Generation of Therapeutic ERCP Reports. Gastrointest Endosc 1991; 37(6):617-620.

9 Dershaw DD. Voice-Activated Radiology Reports. Radiology 1988; 187:284.

10 Hollbrook JA. Generating Medical Documentation Through Voice Input: The Emergency Room. Top Health Rec Manage 1992; 12(3):58-63.

[11] Issacs E, Wulfman CE, Rohn JA, Lane CD, Fagan LM. Graphical Access to Medical Expert System: IV. Experiments to Determine the Role of Spoken Input. Methods Inform Med 1993; 32(1):18-32.

[12] Wulfman CE, Rua M, Lane CD, Shortliffe EH, Fagan LM. Graphical Access to Medical Expert System: V. Integration with Continuous-Speech Recognition. Methods Inform Med 1993; 32(1):33-46.

[13] McMillan PJ, Harris JG. Datavoice: A Microcomputer-Based General Purpose Voice-Controlled Data-Collection System. Comput Biol Med 1990; 20(6):415-419.

[14] Schmandt C, Ackerman MS, and Hindus D. Augmenting a Window System with Speech Input. Computer 1990; 23(8):50-56.

[15] Feldman CA, Stevens D. Pilot Study on the Feasibility of a Computerized Speech Recognition Charting System. Community Dent Oral Epidemiol 1990; 18:213-215.

[16] Smith NT, Brian RA, Pettus DC, Jones BR, Quinn ML, Sarnat L. Recognition Accuracy with a Voice-Recognition System Designed for Anesthesia Record Keeping. J Clin Monit 1990; 6(4):299-306.

[17] Ikerira H, Matsumoto T, Iinuma TA, et al. Analysis of Bone Scintigram Data Using Speech Recognition Reporting System. Radiation Medicine 1990; 8(1):8-12.

[18] Grasso MA, Grasso CT. Feasibility Study of Voice-Driven Data Collection in Animal Drug Toxicology Studies. Comput Biol Med 1994; 24:4:289-294.

[19] National Center for Toxicological Research, Post Experiment Information System Pathology Code Table Reference Manual, TDMS Document #1118-PCT-4.0, Jefferson, Ark: 1985.

[20] Murchie CJ and Kenny GNC. Comparison of Keyboard, Light Pen and voice recognition as methods of Data Input. Int J Clin Monit Comput 1988; 5:243-246.