Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10007–10010, October 1995
Colloquium Paper

This paper was presented at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Commercial applications of speech interface technology: An industry at the threshold

JOHN A. OBERTEUFFER

Voice Information Associates, Lexington, MA 02173

ABSTRACT **Speech interface technology, which includes automatic speech recognition, synthetic speech, and natural language processing, is beginning to have a significant impact on business and personal computer use. Today, powerful and inexpensive microprocessors and improved algorithms are driving commercial applications in computer command, consumer, data entry, speech-to-text, telephone, and voice verification. Robust speaker-independent recognition systems for command and navigation in personal computers are now available; telephone-based transaction and database inquiry systems using both speech synthesis and recognition are coming into use. Large-vocabulary speech interface systems for document creation and read-aloud proofing are expanding beyond niche markets. Today's applications represent a small preview of a rich future for speech interface technology that will eventually replace keyboards with microphones and loudspeakers to give easy accessibility to increasingly intelligent machines.**

Speech interface technology, which encompasses automatic speech recognition, synthesized speech, and natural language processing, comprises the areas of knowledge required for human-machine communication by voice. This paper discusses commercial applications, which are beginning to have significant impacts on business and personal use. Commercial applications of speech interface technology, which first appeared in the early 1980s, are poised now in the early 1990s at a threshold of widespread practical application. Today's applications in speech interface technology utilize speech recognition or synthesis to simply translate spoken words into commands and text or vice versa with little regard to underlying meaning. In the future as applications for human-machine communication by voice grow, the need for natural-language-processing technology to permit speech interpretation will increase. The applications and developments described below represent some very important first steps into a future that will include systems capable of understanding natural conversational speech for transcription or spoken real-time translation. Today's applications are an important bridge to that future and represent the early and productive uses of speech interface technology.

Automatic speech recognition is the ability of machines to interpret speech in order to carry out commands or generate text. An important related area is automatic speaker recognition, which is the ability of machines to identify individuals based on the characteristics of their voices. Synthetic speech, or synonymously text-to-speech, is audible speech generated by machines from standard computer-stored text. These disciplines are closely related because they both involve an analysis and understanding of human speech production and perception mechanisms. In particular, the analysis of speech into its individual components (phones) and the characterization of the acoustic waveforms of these components are common to both disciplines. Speech recognition and speech synthesis are also closely coupled at the applications level—for example, for remote database access where visual displays are not available. The use of speech recognition for input and synthetic speech for output is a powerful combination that can transform any telephone into a fully intelligent node in a computer network.

BACKGROUND

Automatic speech recognition and text-to-speech technologies have been under development since the early days of modern electronic and computer technology in the middle part of this century. A phonemic-based text-to-speech system was demonstrated at the World's Fair in 1939 by AT&T Bell Laboratories; high-speed computers in the early 1950s made the display of speech spectrograms and the application of pattern recognition techniques for automatic speech recognition practical. By the early 1980s, automatic speech recognition had progressed enough to make practical speech-driven data entry systems. Voice input computers are used in many industrial inspection applications where hands and eyes are busy with other tasks, allowing data to be input directly into a computer without keyboarding or transcription. In the late 1970s, a combination of optical character recognition and synthetic speech made possible the first reading machine for the blind. Although bulky and expensive initially, this device allowed blind people, for the first time, to access arbitrary text with no human intervention.

The development of automatic speech recognition and text-to-speech systems has been carried out by large and small companies and by universities. In the United States, the Defense Advanced Research Projects Agency (now the Advanced Research Projects Agency) has provided significant funding and encouragement to a number of important university and private developers of this technology. In addition, large companies such as IBM, AT&T, and Texas Instruments have provided major research and development funding for advanced speech technologies.

Beginning in about 1990 the combination of powerful inexpensive microprocessors and improved algorithms for decoding speech patterns made possible voice command systems for personal computers and telephone-based systems. More expensive, but very powerful, large-vocabulary systems for the creation of text entirely by voice also have become available. Inexpensive chip-based text-to-speech systems allow talking dictionaries and word translators to be sold as consumer products, while more expensive systems provide concept to speech from powerful databases for telephone access by remote users. With these new platforms and analytical tools

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

available, the number of commercial PC-based applications increased significantly in the early 1990s. The success of these recent applications highlights the need for increased sensitivity to human factors in the implementation of speech technologies. Many early commercial systems were offered without much understanding of the difficulty of implementing these semihuman technologies. Recently, speech interface technology systems have sought to overcome the natural limitations of speech understanding with more user-friendly interfaces.

TECHNOLOGY

Although automatic speech recognition and synthetic speech technology often require the use of significant computing hardware resources, the technology is essentially software based. Digital signal processors are used by many systems, but some speech systems utilize only analog/digital converters and general-purpose computing hardware. Low-end systems may run on single chips; mid-range systems are generally PC based with digital-to-analog and analog-to-digital conversion of speech signals carried out on a separate plug-in card. In some systems this card also includes a digital signal processor for speech analysis or synthesis. In most automatic speech recognition systems the pattern search and matching algorithms run on the main microprocessor.

THE ADVANCED SPEECH TECHNOLOGY MARKET

The automatic speech recognition market can be organized into six major segments as shown in Table 1. In the early 1990s significant growth in new applications has occurred in three of these segments: computer control, speech-to-text, and telephone. In the computer control segment, a number of small and large companies introduced speech input/output products for a few hundred dollars. These products, with both automatic speech recognition and text-to-speech capability, are bundled with popular sound boards for PCs. They provide automatic speech recognition of 1000 to 2000 words for controlling the Windows interface. Their synthetic speech capability can be used for reading text from the screen in word-processing or spreadsheet programs. The initial paucity of software applications, which could take full advantage of this speech capability, limits the popularity of these systems. This is not unlike the early days of the computer mouse when only a few DOS applications allowed point-and-click functions and none included sophisticated operations like drag-and-drop, which the current Windows interface for PC computers allows with mouse-pointing devices.

In the speech-to-text segment of the market, advances in large-vocabulary systems from several companies have been introduced. Systems for voice generation of text with vocabularies of 7000 words for memo and letter generation are available for less than $3000. Medical report generation systems with vocabularies of 50,000 words, selling for over $20,000 each, are finding acceptance in many hospital emergency rooms. In the telephone market there are a growing number of important applications, including operator services and interactive voice response, that use automatic speech recognition. Other applications in information services use both synthetic speech and automatic speech capability for database access. Through these telephone applications, the technology is now being introduced to a very broad market and is being recognized as essential for automating human telecommunications interfaces.

Table 1. Automatic speech recognition market segments

Segments	Applications
Computer control	Disabled, CAD
Consumer	Appliances, toys
Data entry	QA inspection, sorting
Speech-to-text	Text generation
Telephone	Operator services, IVR
Voice verification	Physical entry, network access

Table 2. Synthetic speech market segments

Segments	Applications
Assistive technology	Reading machines, voice output aids
Consumer	Pocket translator, games
Education	Talking word processors
Telephone	Database access
Voice interface	Data entry confirmation, task instructions

The synthetic speech market is segmented into five areas, as shown in Table 2. The telephone segment will see the greatest growth and the most new applications in the near term. In addition, a number of profitable applications exist for consumer products. As more advanced telephone information services are offered, the need for text-to-speech to replace or complement digitally recorded speech will grow. Talking translators and talking dictionaries with very large vocabularies are now available for a few hundred dollars.

RECENT MARKET TRENDS

Some significant trends are apparent in the speech interface technology market. Overall, the growth of this market is being driven by the availability of new products. But other factors, such as cost, integration and acceptance by users, vendors, and applications developers, are important determinants in the penetration of a potential market. Consolidation for the companies involved in this industry is a future trend. A number of still small companies have been in the business of speech recognition and synthetic speech since the early 1980s. Many are currently seeking major partners for business arrangements ranging from mergers to joint marketing agreements. Market expansion also is occurring: a number of large computer and telecommunications companies that have been actively involved in advanced speech technology research for a number of years are just now beginning to offer products and technology licenses commercially. Most speech interface technology vendors are working with users to understand the important human factors issues that affect product use and application. This market demands an understanding of the complexity of speech technology as well as a willingness to provide significant customer support in designing user-friendly interfaces. Overcoming attitudinal barriers is common to any new technology, but it is especially important in this humanlike speech technology whose quality, while impressive in mid-1993, is still far below the level of speech understanding and pronunciation ability of even modestly educated people.

MARKET SIZE

The growth of the advanced speech technology commercial market is represented in Fig. 1, which shows the estimated market value from 1988 through 1997. The automatic speech recognition market will grow to $345 million by 1994 from $92 million in 1990, showing an average compound growth rate of about 40%. The synthetic speech market is smaller, growing to $148 million in 1994 from $51 million in 1988, an approximately 30% compound annual growth rate. A total market for 1994 of about a half billion dollars is projected with the market by the end of the decade for speech interface technologies estimated to be above $2 billion. These numbers represent sales of commercial products but do not include research and development dollars invested in advanced speech technologies

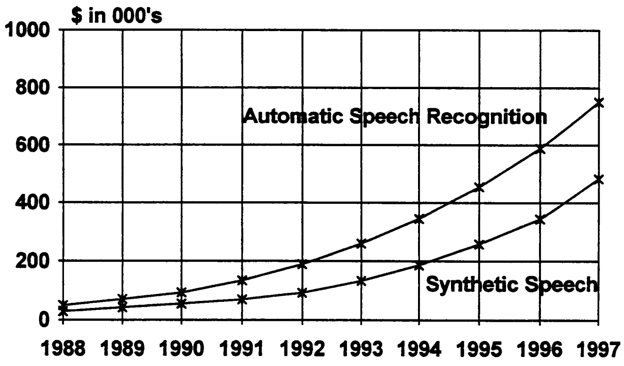

FIG. 1. End-user market revenue for advanced speech technologies. (From VIA Information Associates.)

each year by small and large companies or government programs.

RECENT SIGNIFICANT COMMERCIAL DEVELOPMENTS

Developments in several segments of the automatic speech recognition market reflect trends in speech interface product growth. In the consumer market, a voice-controlled VCR programmer/TV remote has been introduced. This well-thought-out product prompts for spoken commands like time and date to set various VCR and TV functions. It provides an easy alternative to punching in a series of instructions on a small keyboard in a darkened room. This speaker-dependent device is based on a low-cost chip. Inexpensive digital signal-processing chips for speech and other signal-computing applications have been introduced by several developers and will have a significant impact on the application of speech interface technologies in price-sensitive mass market applications. At the other end of the technology spectrum, several vendors are offering systems with vocabularies up to 20,000 to 50,000 words that recognize discrete words and permit the creation of text entirely by voice.

Major players in the personal computer market are introducing new speech technology products. Several major vendors are offering plug-in sound boards with speech recognition and synthetic speech capabilities for a few hundred dollars. These products with vocabularies of a few hundred to a few thousand words provide the ability to navigate around various Windows menus and applications by voice.

In the telecommunications segment, speech interface technologies are automating many long-distance toll calls. The driving force for automation and the use of automatic speech recognition in telecommunications is productivity. Live operator time is estimated at $17 million per second nationwide, so even partial automation of operator-related tasks can yield major benefits to telephone companies. Most of the regional Bell operating companies have pilot tested or introduced voice dialing and/or voice access for various services.

Advanced information services have been demonstrated by several companies. One of these services provides automated stock quotations using speaker-independent speech recognition for inquiry input and synthetic speech for output information. Callers to the service ask for one of three stock exchanges, New York, Toronto, or NASDAQ, and then ask for any one of 2000 companies on each stock exchange by name. Within a few seconds, up-to-date stock quotation information is provided by a synthetic speech system. This system represents a powerful marriage of automatic speech recognition in text-to-speech for totally automated remote database access.

An equally exciting application using text-to-speech is a system developed for the American Automobile Association in Orlando, FL. This system allows a caller to receive driving instructions over the telephone via synthetic speech. Users can call a special number and then Touch-Tone in telephone numbers representing a starting point and a destination point in Orlando. The system uses the two telephone numbers to determine physical locations that are correlated with map information about the city of Orlando. The system calculates the best route from the starting point to the end point, taking into account both the shortest distance and the use of major highways. Having determined the route, the system then generates the text for the driving instructions, which can be faxed to the inquirer or provided by synthetic speech. The information can be accessed using a car telephone while actually driving the route. The driving directions provide real-time information to the driver. The feasibility of this service and cost would be unimaginable if live operators were involved. Even an automated system using prerecorded speech would be impossible since the combination of possible driving instructions within the city of Orlando is astronomical. It should be noted that the driving instructions that are provided do not even exist before the call is made but are created in response to a request. The instructions, which are generated by an artificial intelligence system, provide not just text-to-speech but also concept-to-speech in a system that is a powerful demonstration of computer telephone integration.

FUTURE APPLICATIONS

In the near term, consumer products, voice input/output-capable hardware for PCs, telephone applications, and large-vocabulary text generation systems will dominate developments in speech interface technology. The decreasing cost of hardware will impact both low-end and high-end applications. At the low end, this hardware will make possible consumer applications with speech input and output for appliance control and instructions for use. At the high end, the decrease in cost of ever more powerful platforms for personal computers will make very large vocabulary systems both less expensive and more capable. Intel's Pentium chip will provide high-end, large-vocabulary, automatic speech recognition systems with sufficient power to provide real-time, continuous speech, speaker-independent systems for text generation. As more speech recognition in text-to-speech systems become widely available for personal computers, more applications software will be written to take particular advantage of these input/output modalities. As the software becomes available and its utility is discovered by personal and business users, the demand for speech systems will increase, allowing further reductions in cost.

Before the end of the century, speech recognition and text-to-speech systems will be applied to hand-held computers. The speech interface is ideally suited to these devices because of its small space requirements and low cost. Speech input provides far more capability than pen input; speech output is competitive with small screen displays in many applications. New lap-top computers and personal digital assistants, with voice input for recording and voice annotation, will incorporate voice-powered navigation systems.

Voice dialing of telephones, whose introduction has begun on a modest scale, will become widely used in the United States, particularly with car phones and in many other applications. Limited network-based speech recognition systems have been introduced recently both in trial and actual pilot applications. In some areas it is possible to voice dial from any cellular phone using either numbers that are recognized on a speaker-independent basis or speed dialing from a personal speaker-dependent directory. This system allows voice dialing and speed dialing by voice for easy hands-free/eyes-free telephone use in a variety of situations. Telephone services using speech input and output will continue to increase. Automated directory assistance has begun in Canada using speech recognition technology. This system allows users to complete many directory assistance calls without live operator assistance by recognizing city names and the names of common businesses.

An application that exemplifies speech interface technologies and that may find significant application in the future is the voice-controlled automated attendant. One advanced speech recognition developer is currently demonstrating such a system in its offices, which have approximately 2000 employees. The system makes it possible to ask for any one of these employees by name. The automatic speech recognition system will recognize the person's name, speak it back with the appropriate extension number using synthetic speech, and route the call. This system is fast and accurate and may be used as a telephone directory. This implementation represents another powerful integration of automatic speech recognition and synthetic speech in a system representative of those that will become commonplace by the turn of the century.