View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CSUSB ScholarWorks

Journal of International Information Management

Volume 7 | Issue 2 Article 6

1998

Voice user interfaces (VUIs) emerge

Brian D. Lynch
Daemen College

Follow this and additional works at: http://scholarworks.lib.csusb.edu/jiim

 Part of the Management Information Systems Commons

Recommended Citation

Lynch, Brian D. (1998) "Voice user interfaces (VUIs) emerge," *Journal of International Information Management*: Vol. 7: Iss. 2, Article 6.
Available at: http://scholarworks.lib.csusb.edu/jiim/vol7/iss2/6

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Journal of International Information Management by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

Voice user interfaces (VUIs) emerge

Brian D. Lynch
Daemen College

ABSTRACT

Human beings use speech as the primary means of communication. Therefore, voice user interfaces (VUIs) represent a natural way for people to interact with computers (Fournier, 1996; Hyde, 1979; Teja and Gonnella, 1983; Witten 1982). Voice input facilitates a much higher computer input rate than keyboard or mouse driven input. Voice output permits computer generated output in devices when output screens are not available (e.g., most phones) or in situations where the user's eyes are busy elsewhere (e.g., driving a car, assembling a product, etc.). Thus, VUIs are viewed as the logical next generation in computer interfaces. Forecasts call for a rapid expansion of voice technology within our work environments over the next few years (Cone, 1997; Schwartz and Brier, 1997). This paper discusses the current state of and potential future of VUIs.

INTRODUCTION

Systems using VUIs and specialized natural language processing systems (i.e., narrow knowledge and contextual domains) are being successfully applied in a wide variety of business environments. In the next few years, experts predict that the majority of business transactions will take place between a person and an automated personality (Markoff, 1998). "Speech is not just the future of Windows, but the future of the computer itself" according to Microsoft Chairman William H. Gates III (Gross, et al., 1998). Through September 1997, U. S. voice recognition software retail sales grew nearly 3000% over the year-ago period (Wirthman, 1997). Greg Tapper, an analyst of Giga Information Group, expects that the beginning stages of speech recognition technology as the impending user interface will be seen in 1998 or shortly thereafter (Schwartz & Brier, 1997). By the year 2001, Jackie Fenn, a Gartner Group Inc. analyst, estimates that "more than 30% of general office workers will use some form of voice recognition software" (Cone, 1997). Businesses find voice technology very attractive because speech is the easiest form of communication for most people. As voice technology improves, voice input may become the most common form of computer input (Nickerson, 1998). A fully functional VUI could give speech as an output (speech synthesizer) and recognize speech as an input (speech recognition).

Published by CSUSB ScholarWorks, 1998

Early voice based applications were very simple and confined by the limitations of technology at that time. The systems were very expensive and possessed an unacceptable error rate for real life situations. Due to these constraints, the early uses of speech recognition was mainly in laboratories (Holmes, 1984; Reedy, 1979). Until recently, most speech recognition applications were discrete speech systems requiring brief pauses between words. Further, most systems required the user to spend time training the system to recognize their voice and individual tones or accents. Currently, training-free continuous speech products are beginning to emerge in the marketplace.

This paper proceeds as follows: first, VUI fundamentals are discussed; then, current VUI applications are presented; and finally, the paper draws conclusions about the potential future of VUIs.

VUI FUNDAMENTALS

Parsing Speech

All speech recognizers, biological or mechanical, have the ability to convert sound waves to internal representations. Speech recognition works by the use of phoneme recognition. Phonemes are the smallest acoustical components of a language, with roughly 80 of them making up the needs of the spoken English language. To study the patterning sounds, linguists write down speech utterances as sequences of sounds. Linguists use the *International Phonetic Alphabet* which consists of fewer than 100 sound symbols or phonemes–it can be used for any of the more than 5,000 known languages. All speech recognizers have internal models stored in memory of the acoustic patterns produced by speech then spoken speech is matched against these internal representations to facilitate the creation of a digital representation of the words. Other key design issues for voice recognition systems include: understanding speech, navigation versus dictation, discrete speech versus continuous speech, and speaker dependence versus speaker independence.

Understanding Speech

True understanding of everyday speech for purposes of acting upon by the computer is still mostly science fiction. The problems with understanding speech are complex. First, the computer must recognize each word in a sentence. Then, the sentence must be parsed into its grammatical elements. Next, the computer must try to derive (understand) the meaning of what it has parsed. And finally, the computer must act upon or respond to the sentence. Currently, some prototype voice understanding systems deal with these issues by focusing on extremely narrow knowledge domains (e.g., airline reservations, locating restaurants, etc.) and thus significantly restrict the potential context of the spoken word ("The galaxy's guide . . .," 1996).

Navigation vs. Dictation

There are two primary types of human to computer interaction patterns: navigation (or command and control) and dictation. Navigation uses voice commands to access certain menu

items and accomplishes some of the command and control steps typically done with a mouse or keystrokes or touch-tone phone pads. Dictation involves entering text and numbers into applications such as word processing and spreadsheets (Rash, 1994). A dictation system might also have some navigation elements (e.g., to facilitate editing, etc.).

Discrete Speech vs. Continuous Speech

Discrete speech systems require a pause between each word. The computer software translates speech sound patterns into words which become written text or the basis for a command and control sequence. Continuous speech systems permit natural speech patterns (i.e., no pause after each spoken word).

Three properties of continuous speech make word recognition a complex task: word boundaries, co-articulatory effects and difficulty with content and function words (Fournier, 1996). The pause pattern in discrete speech systems makes word boundaries easily identifiable. In contrast, word boundaries are more difficult to identify in continuous speech. Co-articulatory effects are caused by the fact that certain letters, syllables or words are typically more emphasized than others during natural or continuous speech. As speaking rates increase, the co-articulatory effects also increase. The content-function word problem is also connected to the articulation issue. Content words are nouns, verbs, adjectives, and adverbs. Function words are pronouns, prepositions, short verbs, and articles. In natural speech, content words are more often articulated than function words (Lee, 1989).

Speaker Dependence vs. Speaker Independence

Speaker dependent systems require each user to train the system to recognize his or her voice (Reedy, 1979; Witten, 1980). Speaker independent systems require little or no system training. However, many experts agree that "for the same task, speaker-independent systems will have three to five times the error rate of speaker dependent ones" (Lee, 1989, p. 5).

VUI APPLICATIONS

Current VUI Dictation and Navigation Products

Three vendors currently dominate the PC voice dictation and navigation market: Dragon Systems, IBM, and Kurzweil. Tables 1-3 provide product information for the key product offerings for each of the three vendors.

Dragon Systems' products require one-half hour of voice training prior to use. After training, the speech recognition accuracy rate is 98%. Kurzweil's and IBM's products are usable right out of the box (i.e., no voice training required) with speech recognition accuracy rates of 90%. In addition, Kurzweil's and IBM's products adapt to (learn) users' speech and language patterns boosting recognition accuracy to 97%. Essentially, Kurzweil's and IBM's products are speaker

Published by CSUSB ScholarWorks, 1998

independent out of the box. However, to get the adaptive accuracy rates of 97%, the products require creating a user profile for each user (effectively a speaker dependent solution).

Dragon Systems' products have the largest vocabularies, but are the most expensive and least integrated with other applications. IBM's products have noticeably smaller vocabularies while Kurzweil's products have significantly slower dictation speeds. Kurzweil's products are the most integrated with other applications and the only products currently offering operating system navigation. On the negative side, their products are still primarily discrete voice systems with extremely limited dictation speeds. However, Kurzweil will soon release continuous versions of its products with dictation speeds of approximately 140 words per minute and with text to speech capabilities. Kurzweil also offers another product, VoiceCommands ($14.95), that permits formatting and editing of Microsoft Word documents with continuous speech. It is a command and control product (as opposed to a dictation product) built with natural language processing technology. The product allows you to use speech instead of the mouse and keyboard to accomplish tasks typically controlled through the menus and dialogs of Microsoft Word. By virtue of its natural language processing capabilities, the product would facilitate high level editing functions (e.g., creating and editing a table) without requiring knowledge of the Microsoft Word menus and dialog boxes.

Table 1. Dragon Systems' Voice Product Offerings

Features	Naturally-Speaking Standard	Naturally Speaking Preferred	Naturally Speaking Professional
Discrete or Continuous Speech	Continuous	Continuous	Continuous
Vocabulary Size	230,000 total vocabulary 35,000 active	230,000 total vocabulary 42,000 active	230,000 total vocabulary 55,000 active
Dictation Speed	160 words per minute	160 words per minute	160 words per minute
Dictation to other applications	Yes	Yes	Yes
Navigation of other applications	Microsoft Word Corel WordPerfect	Microsoft Word Corel WordPerfect	Microsoft Word Corel WordPerfect
Multiple users	Yes	Yes	Yes
Text-to-Speech	No	Yes	Yes
Suggested Retail Price	$109	$229	$695

(Source: Dragon Systems, 1998)

Table 2. IBM's Voice Product Offerings

Features	Simply Speaking Gold	Via Voice	Via Voice Gold
Discrete or Continuous Speech	Continuous	Continuous	Continuous
Vocabulary Size	22,000 64,000	22,000 64,000	22,000 64,000
Dictation Speed	100 words per minute	125 words per minute	125 words per minute
Dictation to other applications	Microsoft Word and Other Applications	Microsoft Word	Microsoft Word and Other Applications
Navigation of other applications	Yes	No	Yes
Multiple users	Yes	Yes	Yes
Text-to-Speech	Yes	Yes	Yes
Suggested Retail Price	$34.95	$74.00	$124.00

(Source: IBM, 1998)

Table 3. Kurzweil's Voice Product Offerings

Features	VoicePad	VoicePlus	VoicePro
Discrete or Continuous Speech	Discrete for words Continuous for digits	Discrete for words Continuous for digits	Discrete for words Continuous for digits
Vocabulary Size	200,000 total vocabulary 20,000 active	200,000 total vocabulary 30,000 active	200,000 total vocabulary 60,000 active
Dictation Speed	60 words per minute	60 words per minute	60 words per minute
Dictation to other applications	No	Yes	Yes
Navigation of other applications	No	Most applications and Windows 3.1 x and 95	Most applications and Windows 3.1x and 95
Multiple users	Yes	Yes	Yes
Text-to-Speech	No	No	Nos
Suggested Retail Price	$29.95	$49.95	$69.95

(Source: Kurzweil, 1998)

Published by CSUSB ScholarWorks, 1998

Other Business Oriented VUI Applications

Small vocabulary voice systems now permit hands-free voice dialing from business, home, and mobile phones (Bayle, 1993). A simple voice recognition application which determines whether people say "collect," "operator," "third party," "credit card," or "person to person" saves AT&T several hundred million dollars a year (Markoff, 1998). During the 1997 Christmas season, the United Parcel Service (UPS) deployed a VUI system that gives package tracking information in response to a caller's spoken commands. On Christmas Eve, the system handled 193,000 calls. The system paid for itself in three months and operating costs are about one-third of the cost of using workers to handle the package tracking calls (Gross et al., 1998; Markoff, 1998).

VUIs are also being used to fight phone card fraud via speaker dependent passwords. Analysts estimate that fraud accounts for $1.3 billion of the $6 billion in revenue collected by the calling card industry (Thyfault, 1995). Using the unique fingerprint-like characteristics of individual voices, voice recognition systems can also be used as security systems to only allow authorized personnel into restricted physical areas of "digital" areas (Eng, 1995; Stair, 1998). Voice prints could also be used to catch criminals (e.g., obscene or harassing phone callers, bomb threats, ransom demands, etc.). A voice print database could be built by police to facilitate forensic identification of criminals by their voices ("Voice recognition to catch crooks," 1996).

Small to moderately large vocabulary voice systems (2000-5000 words) facilitate interactive 24-hour banking, credit card querying, and remote tax information queries (Barney, 1995; "Minnesota's rotary . . .," 1995; Sharman, 1995; Wildstrom, 1995). American Express is developing a voice recognition system, code named PARIS, to take airline reservations. American Express estimates that the system will slash transaction costs in half and cut average transaction time to two minutes from seven minutes (Thyfault, 1997).

Voice based systems are used in direct marketing and marketing research (Blyth, 1994; Fenn & Hodgdon, 1995; Syedain, 1994). Burlington Industries and General Motors have reported significant production gains due to quality and control related voice systems in use on production lines (Andrea, 1995; "Lumber mill scales . . .," 1995). Voice systems can be used by equipment operators on the factory floor to give basic commands to machines while they use their hands to perform other operations (Stair, 1998).

Students at St. John's University use an interactive voice response system (IVRS) to register for classes and access their final grades ("St. John's University . . .," 1997). Pine Forest High School (Pensacola, Florida) uses an IVRS to permit students and parents to obtain 24-hour real-time access to grades and attendance records. The system also lets the school automatically call student homes to deliver recorded information and messages as well as to conduct automated surveys. A survey found that 86% of parents felt that the new system helped them become more involved with their child's school activities ("Fla. High School . . .," 1996).

Hospital and law firms use large vocabulary voice systems (30,000-60,000 words) to dictate correspondence, other documents, and reports without involving intermediary typists (Cohen & Beshers, 1995; "Software opens doors . . .," 1995; Zinn, 1991). Goldstein, Martin, and Sin-del

(1996) found that significant time savings and practice pattern enhancements may be derived from using a voice system to record patient records. Westlaw has a VUI that permits subscribers to enter search commands and queries by voice ("A keyboard . . .," 1994). Natural language processing could also facilitate the data mining of corporate databases. If the VUI was intelligent enough to parse spoken words into queries, this would permit data mining by people who do not have the necessary skill set to program their queries using conventional keyboard and mouse driven database query interfaces (Schwartz & Brier, 1997).

VUIs could be used to help organizations comply with the Americans with Disabilities Act (ADA) Fournier, 1996; "Software opens doors . . .," 1995). VUIs would facilitate computer interaction by individuals who are unable to use their hands due to physical disabilities. In particular, VUIs could be used to assist workers with repetitive strain injuries (RSI) and carpal tunnel syndrome (a specific RSI). VUIs could also assist people with fluent oral communication skills, but disabilities in written communications (Wetzel, 1996). VUIs could also benefit people who lack keyboarding skills. VUIs could even increase the efficiency of keyboard fluent individuals. For example, Dirks and Dirks (1997) found that 20% of a sample of undergraduate business communication students could dictate straight copy faster than entering it via keyboard after a 1½ hour tutorial with a discrete voice system.

VUIs could also prove extremely beneficial for written languages not derived from the Phoenician alphabet. For example, the complicated writing schemes of Japanese and Chinese make use of the standard keyboard problematic (Cone, 1997).

CONCLUSIONS

For most people, continuous speech input facilitates a much higher computer input rate than keyboard or mouse driven input. Voice output permits computer generated output in cases when output screens are not available (e.g., most phones) or the user's eyes are busy elsewhere (e.g., driving a car, assembling a product, etc.). VUIs may also provide a cost effective means for dealing with the ADA and repetitive motion injuries. Given the potential benefits, the forecasts of rapidly increasing penetration rates of VUI's into our work environments (Cone, 1997; Schwartz & Brier, 1997) make intuitive sense. Nonetheless, skeptics argue that the current structure of many work environments (e.g., cubicles) will hinder the wide deployment of VUIs. People working in close proximity to each other can hear each other. Even assuming speaker dependent systems, you may not want anyone to hear what you are inputting to or outputting from your computer (Johnston & Levin, 1997). Yet, realistically VUIs are still in their emergent phase and their future looks very bright. As evidenced by the discussion and examples cited throughout this paper, simple voice recognition systems and specialized natural language processing systems (i.e., narrow knowledge and contextual domains) will continue to facilitate the expansion of computer technology into our business environments and our everyday lives.

Published by CSUSB ScholarWorks, 1998

REFERENCES

A keyboard. How quaint? (1994). *American Libraries, 25*(4), 352.

Americans with Disabilities Act (ADA). (1992). Washington, DC: Department of Justice. U. S. Equal Employment Opportunity Commission.

Andera, F. (1989). Voice input to computers: How will it affect the teaching of business communications? *The Bulletin, 52,* 18-20.

Barney, D. (1995, March). Telephony gets an improved interface. *Infoworld, 17,* 45.

Bayle, S. (1993, March). The freedom to dial hands-free. *Communications,* 36-40.

Blyth, B. (1995). ASR: The last word in self-completion. *Marketing Research, 7*(1), 42-44.

Cohen, E. E. & Beshers, C. (1995). Making automated speech recognition work. *The CPA Journal, 65*(2), 72-74.

Cone, E. (1997, November). Voice comes to the desktop--New speech products let users just say the word (or words) to control their office applications. *Information Week,* 218-220.

Dirks, R. & Dirks, M. (1997). Introducing business communication students to automated speech recognition. *Journal of Business Education, 72*(1), 153-156.

Eng, P. M. (1995, February). A set of prints may foil thievery. *Business Week,* 118.

Dragon Systems (1998, August). www.dragonsystems.com.

Fenn, J. & Hodgdon, P. (1995, August). How emerging artificial intelligence technologies will affect direct marketing. *Direct Marketing, 58,* 28-30.

Fla. high school links gradebook & more to IVR phone system. (1996). *Technological Horizons in Education, 23*(9), 59-61.

Fourier, R. S. (1996, March/April). Developments in voice-technology. *Journal of Education for Business, 71*(4), 241-245.

Goldstein, M., Martin, G., & Sin-del, B. The use of a computerized voice to text dictation system in the neonatal intensive care unit: The ergonomic model. *Pediatrics, 98*(3), 584.

Gross, N., Judge, P. C., Port, O., & Wildstrom, S. H. (1998, February 23). Let's talk! *Business Week,* 60-76.

Holmes, J. N. (Ed.). (1984, October). *Proceedings of the 1st international conference on speech technology,* 23-25. Brighton, UK: IFS (Publications) Lt.

Hyde, S. R. (1979). Automatic speech recognition: A critical survey and discussion of the literature. In N. R. Dixon & T. B. Martin (Eds.), *Automatic speech and speaker recognition,* 16-55. New York: IEEE Press.

IBM (1998, August). http://www.software.ibm.com/is/voicetype/

Johnston, S. J. & Levin, R. (1997, April). Computers conversant with speech. *Information Week, 28,* 44.

Kurzweil (August, 1998). http://www.lhs.com/dictation.

Lee, K. F. (1989). *Automatic speech recognition: The development of the SPHINX system.* Norwell, MA: Kluwer Academic Publishers.

Lumber mill scales voice-based data entry. (1995, September). *Communications News, 32,* 26-29.

Markoff, J. (1998, August 21). It isn't human, but the voice on the line is ready to help. *New York Times,* 1, 17.

Minnesota's rotary dialers can get automated tax info. (1995, April). *Communications News,* 53.

Nickerson, R. C. (1996). *Business and information systems.* Addison-Wesley Educational Publishers, Inc.

Rash, W. (1994, December). Signing off on handwriting and sounding off on voice. *PC Week,* 203-207.

Reedy, D. R. (1979). Speech recognition by machine: A review. In N. R. Dixon & T. B. Martin (Eds.), *Automatic Speech & Speaker Recognition,* 56-86, USA: IEEE Press.

Schwartz, E. & Brier, S. (1997, November 17). Voice recognition may become the next UI. *Infoworld,* 3.

Sharman, M. (1995, April). Speech recognition technology helps banks enhance customer service. *Speech Recognition, 178,* 60-61.

Software opens doors for disabled. (1995, April). *Communications News, 32,* 21.

Stair, R. M. (1998). *Principles of information systems: A managerial approach.* Course Technology, An International Thomson Publishing Company.

St. John's University enhances student services with interactive voice response system. (1997). *Technological Horizons in Education Journal, 25*(2), 87-88.

Syedain, H. (1994, November 17). Technology finds a new voice. *Marketing,* 14.

Teja, E. R. & Gonnella, G. (1983). *Voice technology,* Reston, VA: Reston Publishing Company.

The galaxy's guide to the hitch-hiker. (1996). *Economist, 339*(7965). 77.

Thyfault, M. E. (1996, December 16). Just say it and pay it. *Information Week,* 96.

Thyfault, M. E. (1997, July 14). Voice recognition enters the mainstream. *Information Week,* 20.

Voice recognition to catch crooks. (1996). *Futurist, 30*(1), 46.

Wetzel, K. (1996). Speech-recognizing computers: A written-communication tool for students with learning disabilities? *Journal of Learning Disabilities, 29*(4), 371-372.

Wildstrom, S. H. (1995, April 24). This secretary really listens. *Business Week,* 19.

Wirthman, L. (1997, November). Computer voice recognition is past the whispering stage. *Investor's Business Daily, 17,* A10.

Witten, I. H. (1980). Communicating with microcomputers: An introduction to the technology of man-computer communication. *Computer and People Series.* B. R. Gaines (Ed.). London: Academic Press.

Witten, I. H. (1982). Principles of computer speech. *Computer and People Series.* London: Academic Press.

Zinn, T. K. (1991, December). Voice recognition: The key to hospital dominance. *Computers in Healthcare,* 14-15.