Voice Recognition Technology: Has It Come of Age?

Joseph R. Zumalt

Voice recognition software allows computer users to bypass their keyboards and use their voices to enter text. While the library literature is somewhat silent about voice recognition technology, the medical and legal communities have reported some success using it. Voice recognition software was tested for dictation accuracy and usability within an agriculture library at the University of Illinois. Dragon NaturallySpeaking 8.0 was found to be more accurate than speech recognition within Microsoft Office 2003. Helpful Web sites and a short history regarding this breakthrough technology are included.

Typing, or keyboarding, as it is referred to today, is perhaps a more important activity than it was one hundred years ago, as Penn State University Library was one of the first to catalog its entire collection on the typewriter in 1902–03.[1] Along with many other white-collar professions, librarianship is dependent upon interaction with a computer. Metadata creation, online chat, blogs, discussion lists, and e-mail have multiplied the need for keyboarding. However, other data input methods are becoming increasingly available. Handwriting recognition and voice recognition are two additional ways for users to interact with their computer.

It seems inevitable that nearly everyone may use their voice to interact with computers in the future; it is just a question of how soon. A glimpse of the future can be found in the television series *Star Trek*, in which the characters used speech almost exclusively to interact with their computers. In a humorous vignette in the movie *Star Trek IV: The Voyage Home*, the twenty-fourth-century character, Lt. Commander Montgomery Scott (Scotty), has to interact with a twentieth-century computer. After attempting to use his mouse as a microphone, he is instructed by another character to use the keyboard. Without any experience in using a keyboard, Scotty then proceeds to rattle off a complex series of equations for "transparent aluminum" at approximately two hundred words per minute.

Voice recognition technology (VRT) has been touted as the next "killer application" several times. However, many have tried this technology and put it aside, vowing to return to it when the technology improves. Thus, it is easy to understand why this technology has not gained more traction. While VRT has been widely accepted by individuals with disabilities that prohibit or hinder their ability to type, the vast majority of people who are able to type continue to use their keyboarding skills. Keyboarding skills developed over years or decades have been seen as sufficient or superior to VRT. Because VRT applications require hardware not demanded of other software, such as a good sound card, a microphone, additional software, and some initial training, most users are not going to try it without some very compelling reasons. Other problems include financial constraints amongst the principal software developers and suppliers, likely leaving many customers waiting to see what Microsoft will do. Even though there has been hesitance to embrace this technology, the steady progress of computing power has certainly helped move VRT toward greater acceptability.

In late 1997, Sara Hedberg wrote a provocative article titled "Dictating This Article to My Computer: Automatic Speech Recognition Is Coming of Age." Hedberg's observations certainly resonated with the author and many other writers reviewing this technology. Her bottom line: she did not think the application was good enough for her because she was a fast typist, and correcting errors produced by the software when it did not recognize her words slowed her down. Nonetheless, she believed a slower typist could make better use of the program. Hedberg said she looked forward to the next ten years of development in VRT, when she thought it could become more prevalent, especially with Microsoft becoming more involved.[2]

What developments have occurred in voice recognition in the last eight years or so since Hedberg's article? Has VRT improved enough to really come of age? How can we use it to benefit librarians? This piece will help to fill in some of the gaps in the library literature and bring it up to date, and report on this technology's use in one library setting.

A short history of VRT software

IBM has been active in the field of VRT for many years. James K. Baker, a researcher at IBM who wrote several articles in the late 1970s about this technology, was one of the pioneers. He and others decided to create their own privately held company, Dragon Systems. The company produced a software package called Dragon Dictate in the early 1990s, which was a discrete-speech package (a distinct pause was required between every word spoken). While this required pause was annoying to the majority of users, the software gained a loyal following, some of whom are still using versions of it to this day. In 1997, Dragon Systems stopped development of

Joseph R. Zumalt (jzumalt@uiuc.edu) is Assistant Professor and Assistant ACES Librarian, Isaac Funk Family Library, University of Illinois at Urbana-Champaign.

Dragon Dictate and introduced a new product, Dragon NaturallySpeaking (also referred to as DNS).

Dragon NaturallySpeaking was actually part of a line of several products produced by Dragon Systems, including products tailored to the legal and medical communities. It was the first continuous-speech package (you could speak at normal speed, reportedly at up to 160 words per minute). Dragon NaturallySpeaking was updated approximately five times before the owners of Dragon Systems sold their company to a European company called Lernout and Hauspie, which had a competing product called VoiceXpress. Lernout and Hauspie promptly proceeded to have financial troubles in Asia and went bankrupt. In 2001, a technology company spun off from Xerox called ScanSoft acquired the remnants of VoiceXpress and Dragon NaturallySpeaking. IBM also was active in speech recognition research and marketed a product line called ViaVoice. At present, ScanSoft also distributes and supports IBM's ViaVoice. Microsoft has also been working on VRT for years; Bill Gates has emphasized the bright future of VRT in many speeches and books. However, they did not bring out this functionality until the Office XP suite, which marked the entrance of Word 2002.

Several niche companies have moved into the market. A company called Speaking Solutions sells training manuals and tip sheets for all the different limited-edition packages, and it has partnered with several companies in this field such as Plantronics, the maker of microphones. They also offer a significant amount of training scattered at sites all around the country. Partly because of this, large numbers of business schools are now offering courses to train students on how to use VRT software. A small company has created a CD-ROM–based video guide series for the NaturallySpeaking line of products. Another company, KnowBrainer, hosts the Unofficial NaturallySpeaking Public Forum, an excellent source of information about technical issues. They also have recently begun building both laptop and desktop computers with parts meant to maximize performance for VRT.

Literature review

To obtain a complete picture of the activity in this area, one needs to include the phrases "voice recognition" and "speech recognition" to retrieve relevant searches. The Library Literature database prefers the broad subject heading "speech processing systems." Some articles were written about VRT in its infancy. In addition, with the rapidly changing nature of this technology, it is important to remember that studies that are several years old may now be obsolete.

The promise of alternatives to keyboarding has been very alluring to those in organizations trying to serve those with disabilities. Tanya Goette did a field study on the use of VRT in the late 1990s by people with disabilities, some of who were successful using VRT and others who were not.[3] Some individuals, such as Janine Ladato, who has been battling multiple sclerosis for years, have found VRT ". . . magical, wonderful, and definitely worth the effort needed to learn to adjust to it."[4] In addition, it has been reported that voice recognition software has proven helpful to a man with aphasia, an inability to articulate ideas in written language.[5]

While it is encouraging to read success stories from the rehabilitation literature, it is difficult to infer how average users will react to the technology. However, it seems VRT software has been gaining some momentum. Increasing numbers of business schools, universities, and community colleges utilize VRT software. For example, Seton Hall University gave IBM's ViaVoice to all incoming freshmen, which would seem to indicate a trend toward greater acceptance of this technology.

Several technology writers have tackled this subject over the last few years. Many, like Hedberg, have been hesitant to switch to voice recognition, partly because of the pragmatic uses of the technology and partly due to their advanced typing skills. However, Jon Udell in a recent issue of *InfoWorld* states that ". . . dictation technology may finally have crossed the threshold of practicality for me."[7] David Pogue, a technology writer for the *New York Times,* has embraced this technology for several years now.[8] Both have produced short videos available on the Internet that give the viewer a look into how VRT software functions. In addition, several interesting videos using Microsoft Office's speech recognition engine are available on the Web showing applications for construction engineers.[9]

Unfortunately, not a great deal has been written about voice recognition in the last ten years in the library literature. One of the few is an informative piece concerning the use of Dragon Dictate in cataloging. David Bertuca does a nice job of framing the issues regarding the pros and cons of this software.[10] Most of the literature naturally comes from the legal and medical professions, both of which generate a large volume of documents, and which also use a large specialized vocabulary.

Method

Because technology and software changes rapidly, testing VRT had to be done on several different computers. These tests were performed on more than one machine as opposed to a single machine, so absolute comparisons between versions was not possible. However, each of these machines is loosely comparable, allowing for a close comparison. These machines are basic, office-type machines, certainly not expensive, state-of-the-art models.

- (Running DNS 6.0) Intel 1.9 GHz, 512 MB RAM, purchased in spring 2002
- (Running DNS 7.3) AMD 2400+, 2.0 GHz, 768 MB RAM, purchased in spring 2004
- (Running DNS 8.0) Intel 2.53 GHz, 512 MB RAM, purchased in spring 2003
- (Running DNS 8.0, using existing speech files converted from DNS 6.0) Intel 1.9 GHz, 512 MB RAM, purchased in spring 2002
- (Running Speech Recognition for Microsoft Office 2002), 2.6 GHz, 512 MB RAM, purchased in summer 2003
- (Running Speech Recognition for Microsoft Office 2003) Intel 2.53 GHz, 512 MB RAM, purchased in spring 2003

In order to evaluate the software, the testers utilized the last three versions of Dragon NaturallySpeaking Preferred, versions 6.0, 7.3, and 8.0. Current VRT software requires the user to spend time dictating specifically selected passages to a computer, which then analyzes the speech patterns of the user's individual voice for future use. Accurate, instant recognition of any voice, seen on many science fiction programs, is not currently possible. Each tester trained all versions with the minimum period allowed, about five minutes worth of dictation per program. Two other speech recognition programs were also evaluated, Microsoft Office 2002 and Microsoft Office Professional Edition 2003, each of which also required about five minutes of training. After training was completed, two passages from the *Voice and Articulation Drillbook* by Grant Fairbanks were chosen: the "Amplifier Passage" and the "Rainbow Passage," widely used because they contain all the different sounds in the English language.[11] To provide some continuity with an earlier review, the author also chose Lincoln's Gettysburg Address.

Additionally a basic Toshiba Pocket PC, the e750 running Microsoft Windows Mobile 2003, was used to test the ability of Dragon NaturallySpeaking 8.0 to transcribe text. Included with the DNS 8.0 Preferred package is the ScanSoft Voice Recorder, which captures dictation on a Pocket PC. To test the efficacy of performing additional training, testers were retrained for an additional thirty minutes on the two newest versions, DNS 8.0 and the Speech Engine for Microsoft Office 2003, and retested.

Test results

As the test results demonstrate, Dragon NaturallySpeaking has become more accurate with each new version (see table 1). As is to be expected, transcription with the Toshiba Pocket PC is not quite as accurate because of the need to record the speech on the handheld device and then transcribe it. Unfortunately for those relying on the Microsoft Office Speech Engine, accuracy is not nearly as strong as with the NaturallySpeaking line, even compared against transcription using the Pocket PC. The tests also reveal the usefulness of additional training. The results are substantially improved with as little as thirty minutes of additional training, and use of DNS 8.0 in particular improved after additional training was integrated into the new speech files.

Microsoft has a bad habit of throwing in the words "but" and "long" into dictation, requiring a great deal of care in proofreading. One of the benefits that Lernout and Hauspie's legacy voice recognition product, VoiceXpress, brought to speech recognition was the "disfluency filter," a feature that eliminated almost completely the stray "ohs and ahs," which are very annoying to most of us not accustomed to dictating so carefully. Without this filter, the user was required to make corrections when these words were inadvertently placed in the dictated document.

One of the criticisms of this particular type of software has often been that people can type faster than these voice recognition systems can produce text. Is this really true? How fast can the average person type? Several sources have stated that the average typist types about thirty to forty words per minute. The typical test takes one to three minutes to check typing speed. The results of a personnel company reviewing 3,475 job applicants in the middle 1990s, taking a five-minute timed test, revealed a mean (average score) of forty words per minute, a median (middle score) of thirty-eight words per minute, and a mode (most numerous score) of thirty-one words per minute. Only the top 30 percent of typists could type more than fifty words per minute. Only about 10 percent can type approximately sixty words per minute or faster. Only two-tenths of 1 percent can type one hundred words per minute or faster. The stated intent of the Web page was to try to educate people that average typing really was not sixty words per minute.[12] Note that the author of the current piece is about an average typist by the above definition, as tested by Mavis Beacon 10 and the online typing test at www.typingtest.com.

Though it is doubtful a tested typing speed can be maintained for any length of time, most everyone has anecdotal evidence of people being able to speak for quite extended periods of time. While many company brochures for voice recognition programs tout the ability of the programs to recognize dictation speeds of up to 165 words per minute, the requirement to dictate punctuation in addition to words definitely slows down the real dictation speed. Ninety to one hundred words per minute seems a more realistic average dictation speed with voice recognition programs such as Dragon NaturallySpeaking or Microsoft Office Speech Recognition.

Case study in an academic library

Another way to test this software in a library setting is to use it to do actual data entry.

The Agricultural Communications Documentation Center (ACDC) in the Isaac Funk Family Library at the University of Illinois searches for and indexes citations pertaining to agricultural communications, an area not very well-covered by such agricultural databases as AGRICOLA and CAB Abstracts. The citations, some now abstracted, are entered into a Microsoft Access database that is viewable and searchable through the center's Web site. Patrons search the database, using the center's customized thesaurus, and can request documents that fit their research interests.

The center's Microsoft Access database holds over 27,000 index citations. All of the citations pertain to agricultural communications, but their scope ranges from advertising circulars to research journal articles. The center's database averages approximately 230 new citations every month.

Table 1. Voice recognition package comparison results, mistakes per passage

	Amplifier passage	Rainbow passage	Gettysburg Address
NaturallySpeaking 6.0	16	20	26
NaturallySpeaking 7.3	12	19	13
NaturallySpeaking 8.0	12	16	5
Toshiba e750 Pocket PC	19	17	11
Microsoft Office 2002	24	33	28
Microsoft Office 2003	25	20	33
Additional Training Trials			
NaturallySpeaking 8.0	8	10	2
Microsoft Office 2003	15	28	15

Table 2. Documents created per hour

	Novice voice user	Experienced voice user
Typing alone	32 documents/hour	17 documents/hour
Typing and voice	12 documents/hour	21 documents/hour
Voice alone	4.33 documents/hour	9 documents/hour

In order to provide a common frame of reference for the user who may not be familiar with VRT, two authors created documents using both keyboarding and voice recognition methods. The novice VRT user with better typing skills typed three selected paragraphs from Abraham Lincoln's "We cannot escape history" speech in seven minutes and voiced it in eight minutes. The more experienced VRT user with average typing skills took fourteen minutes and thirty seconds to type it in, but was able to accomplish the same task in five minutes using voice input.

A random set of new ACDC documents were collected for use in this study. DNS 7.3 Preferred voice recognition software was used on a 2.0 GHz Pentium 4 machine with 768 MB of RAM. Both users, one without any prior voice-recognition experience and another with about seven years of experience using DNS Preferred software, entered as many documents as possible in one hour using three methods: typing only; typing and voice entry; and voice entry only. Three trials of each were averaged to provide a score for each method (see table 2).

The test results confirm that a person with faster typing skills will be able to enter in more documents per hour. However, the slower typist was more experienced with the voice recognition software and was able to better gauge when to type and when voice would be more advantageous, actually entering in more documents with a combination of typing and voicing than by typing alone, twenty-one documents per hour versus seventeen documents per hour.

How might these tests have compared with even earlier versions of the software, which are still running on obsolescent equipment? While this might be difficult to ascertain, it is possible to get a glimpse of how computer equipment has evolved over the years. Benchmark programs have been used for many years to test the performance of computers and to chart the progress of computers. An interesting Web site that has done this for a number of years is www.tomshardware.com. This site has come out with an extensive list of benchmark tests evaluating microprocessors introduced since the mid-1990s. For example, on the microprocessor benchmark PCMark04, which tests CPUs, the Pentium 233 computer from 1997 scored 230; whereas one of the test machines, a Pentium 4, 2.53 GHz, scored 3317; and a latest generation Pentium 4, 3.8 GHz machine scored 5922.

Directions for future research

One of the significant challenges of VRT is the problem of noise in a shared environment. This shared environment is very common in such large workspaces as most cataloging and acquisition departments. A possible solution to this problem could be the use of a device called a Sylencer or Stenomask, a small, handheld mask with a microphone inside. This accessory is used extensively in courtrooms and law offices to take transcripts and depositions and could be a possible solution to the noise and privacy concerns raised by speaking out loud in an essentially public space.

Dragon NaturallySpeaking Professional, a more advanced version of the software, allows for use of a network. It would be a good pilot project to deploy this software in a large department with multiple workstations, perhaps a cataloging department. This would allow for the use of the same voice profile across multiple workstations. One of the difficulties of prior versions of the software was the need to voice train for each separate machine, but this is not as big of a concern now.

More could have been done to test the effects of different hardware and peripherals. While the tested machines were just basic, office-oriented machines with simple peripherals, better computing equipment makes a difference in dictation accuracy. A better processor with the recently released 64-bit AMD chip would improve performance, as would a faster hard drive, a better sound card, and a better microphone.

Conclusion

Why should someone go to the trouble of learning this software, which requires some additional hardware and software and some training time, when one is already satisfied with the speed of typing? While dictation technology has not advanced to the state of reliability seen in popular science fiction, it can be useful in a large number of contexts. For example, persons with disabilities may find voice recognition allows them to reach out to the world in ways not possible before. A slow, "hunt and peck" typist may be able to use this technology effectively to do most of his or her document creation faster than by typing.

Finally, some who are very fast typists may choose to use it only for such simple documents as e-mail or while they are chatting with friends over the Internet or creating their blogs. At the 2005 American Library Association Midwinter Meeting in Boston, the Public Library Association sponsored a blog (i.e., Web log, essentially an online journal). With the increasing influence during the 2004 election cycle of blogs, increased typing on a computer will inevitably see an increase of such repetitive stress injuries as carpal tunnel syndrome. This technology could help alleviate this persistent problem in our white-collar profession.

Useful Web sites

www.scansoft.com. This is the Web site for the company which now owns two of the best sellers in the industry, the NaturallySpeaking line of products originated by Dragon Systems and later acquired by Lernout and Hauspie, and also the ViaVoice line of products originated by IBM.

www.microsoft.com. Microsoft Corporation has included a speech recognition engine with their Office products since Office XP. They include extensive files on how to maximize the use of their speech engine.

www.speakingsolutions.com. This company provides training opportunities, produces teaching manuals, and offers access to technical solutions for voice recognition problems.

www.sayican.com. This company's founder has written a couple of books on voice recognition and has produced a six-hour, three-CD-ROM video guide to the last couple of versions of Dragon NaturallySpeaking.

www.knowbrainer.com. This helpful Web site provides an Unofficial Naturally Speaking Public Forum for all Dragon NaturallySpeaking products. It actually seems to work in tandem with the company's official site. There seems to be more traffic on it, which acts as a great place to ask for and receive help about problems with using the software.

http://weblog.infoworld.com/udell/2004/11/04.html. This technical writer has produced a short video available on the Internet documenting his initial experience with Dragon NaturallySpeaking 8.

www.davidpogue.com. This journalist working with the *New York Times* has been using voice recognition software for many years very successfully. He has a link to an instructive video regarding his use of Dragon NaturallySpeaking 8.0.

www.bsci.auburn.edu/faculty/willi14/src/demo.htm. A professor in the construction field with a Web site devoted to the exploration of VRT, with some videos demonstrating the Microsoft Office Speech Engine.

www.fivestarstaff.com/publication_typing.htm. This interesting Web site from a human resources firm reports on the real-world typing test results of more than 4,000 job applicants.

www.typingtest.com. This Web site provides an easy-to-use online typing test to establish a baseline tying speed.

http://web.aces.uiuc.edu/agcomdb/docctr.html. The Agricultural Communications Documentation Center at the University of Illinois provides a searchable database of over 26,000 citations to materials in such areas as agricultural communications, agricultural broadcasting, and agricultural education.

www.tomshardware.com. This Web site performs benchmark test on computers and reports the latest developments in the computing industry.

www.talk-tech.net/pages/sylencer.html. This company produces the Sylencer, a handheld device that offers privacy to those using a microphone in open areas.

References

1. Pennsylvania State University, "A Brief History of Cataloging at Penn State," Penn State Libraries Web site. Accessed May 9, 2005, www.libraries.psu.edu/tas/cataloging/dept/history.htm.

2. Sara Reese Hedberg, "Dictating This Article to My Computer: Automatic Speech Recognition Is Coming of Age," *IEEE Expert* 12, no. 6 (Nov./Dec. 1997): 9–11.

3. Tanya Goette, "Keys to the Adoption and Use of Voice Recognition Technology in Organizations," *Library Computing* 19, no. 3/4 (2000): 235–44.

4. Janine Lodato, "Advances in Voice Recognition: A First Look at the Magic of Voice Recognition Technology," *Futurist* 38, no. 7 (Jan./Feb. 2005): 7–8.

5. Carolyn Bruce, Anne Edmundson, and Michael Coleman, "Writing with Voice: An Investigation of the Use of a Voice Recognition System as a Writing Aid for a Man with Aphasia," *International Journal of Language and Communications Disorders* 38, no. 2 (2003): 131–48.

6. Janet Rae-Dupree, "Let's Talk," *U.S. News and World Report* 134, no. 16 (May 12, 2003): 58–59.

7. Jon Udell, "Its Master's Voice," *InfoWorld* 26, no. 46 (Nov. 15, 2004): 38.

8. David Pogue, "Speaking Naturally, Anew," *New York Times* Online, Dec. 2, 2004. Accessed Feb. 2, 2005. www.nytimes.com/ref/membercenter/nytarchive.html.

9. Steve Williams, "Speech Recognition in Construction," Department of Building Science, College of Architecture, Design, and Construction, Auburn University Web site. Accessed Feb. 2, 2005, www.bsci.auburn.edu/faculty/willi14/src/introduction.htm.

10. David Bertuca, "Voice Recognition Software and OCLC: Technology That Works," *OCLC Systems and Services* 16, no. 2 (2000): 69–75.

11. Grant Fairbanks, *Voice and Articulation Drillbook*, 2nd ed., (New York: Harper, 1960), 114, 127.

12. Teresia R. Ostrach, "Typing Speed: How Fast Is Average?" Five Star Staffing Web site. Accessed Feb. 2, 2005, www.fivestarstaff.com/publication_typing.htm.