REVIEW

Asst Technol 2001;13:116–130

User Performance With Speech Recognition: A Literature Review

Heidi Horstmann Koester, Ph.D.

Center for Ergonomics, University of Michigan, Ann Arbor, Michigan

The application of speech recognition to the computer access needs of people with disabilities continues to grow, and a greater understanding of user performance with such systems is needed. This article reviews what is known about user performance with speech recognition systems, with a focus on its application to accommodation of physical disability. Although current systems offer the potential of text entry at 150 words per minute, the literature suggests that users actually achieve somewhere between 8 and 30 words per minute. Barriers that may contribute to this gap, such as the costs associated with correcting recognition errors, are reviewed, and directions for future research are proposed. A major need is for additional research involving users who have physical disabilities.

Key Words: Speech recognition—Voice recognition—Human-computer interaction—Assistive technology devices.

Automatic speech recognition (ASR) systems have the potential to greatly improve the productivity and comfort of performing computer-based tasks for a wide variety of users. These systems allow data input into a computer by speaking into a microphone rather than by typing with a keyboard or other input device. Their use is becoming more commonplace in society in general, as physicians dictate medical reports directly to their computers and individuals control remote computer applications by speaking into their cell phones.

The use of speech recognition is also very attractive for individuals with a variety of disabilities. The many possible applications include: control of telephones, TVs, and other home appliances (Cavalier & Brown, 1998; Goette & Marchewka, 1994; Noyes, Haigh, & Starr, 1989); enhanced communication for people whose speech is consistent but not generally intelligible (Coleman & Meyers, 1991; Doyle, Leeper, Kotler, & Thomas-Stonell, 1997); enhanced telecommunications for people who are deaf (Noyes & Frankish, 1992); control of functional electrical stimulation systems (Noyes & Frankish, 1992); writing assistance for people who have difficulty composing written language (De La Paz, 1999; Higgins & Raskind, 2000; MacArthur, 1999; Raskind & Higgins, 1999); and general-purpose access to computers (Thomas, Basson, & Gardner-Bonneau, 1999).

Although all of these applications have potential benefits for individuals with disabilities, the focus of this article is on the use of ASR for general-purpose computer access, particularly text entry. The users of primary interest here are people with unimpaired speech who might use ASR to accommodate a physical disability when nonspeech methods of text entry are either too slow or too painful to meet their needs fully. This general category includes those individuals with severe physical disabilities (e.g., high-level spinal cord injuries) as well as those with localized and possibly temporary upper extremity impairments resulting from repetitive stress injuries (RSIs). For users whose physical disabilities require them to have hands-free access to a computer, ASR is an attractive option compared with potentially less efficient methods, such as mouthstick typing or on-screen keyboards. Users who are able to type manually but slowly are drawn to the use of ASR by its promise of increased speed and decreased muscle fatigue. For people whose use of "standard" manual input methods has led to an RSI or other serious bio-

Address correspondence and reprint requests to Dr. Heidi Horstmann Koester, Center for Ergonomics, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117.

mechanical stress, ASR may provide a productive alternative to continued discomfort and exacerbation of the injury, freeing users from keyboard use and its associated postural constraints.

The promise of ASR is enormous, but some basic questions regarding user performance with speech recognition have not been satisfactorily addressed. These include:

1. How well are speech recognition systems meeting the needs of users who have physical disabilities?
2. What is the range of productivity that a user of an ASR system can expect? How does this depend on the characteristics of both the user and the task?
3. What is the learning curve associated with ASR systems? How long does it take to develop a high degree of proficiency?
4. Are there human factors costs that may partially counteract the benefits of using ASR systems?
5. If so, are there methods of assessing for and delivering ASR systems that can reduce the impact of these costs and result in improved user satisfaction and productivity?

In this article, the literature on ASR systems and physical disability is reviewed, with the goal of understanding what is already known about user performance with ASR and what remains to be discovered. The article concludes with suggestions for future research in this area.

BACKGROUND

Automatic speech recognition (ASR) has been under development at least since the early 1970s. Early systems could recognize only a handful of discrete words or utterances. By the late 1980s, recognition vocabularies of several thousand words became available, with the requirement that the user speak each word consistently and discretely, with short pauses between words. Discrete ASR systems continued to improve in vocabulary size, recognition accuracy, and capabilities. By the mid-1990s, for example, systems such as IBM VoiceType[1] and Dragon Dictate[2] provided a 30,000-word vocabulary and the ability to access many Windows applications.

In 1997, a major breakthrough in ASR technology occurred with the introduction of the first consumer-affordable continuous speech recognition system. Continuous speech allows users to speak at their natural pace and rhythm, with the potential for faster and more satisfactory interaction. Major dictation systems currently available for Windows include IBM ViaVoice[3] and Dragon NaturallySpeaking.[4] The ViaVoice product also has a version for the Macintosh operating system.

In addition to the distinction between discrete and continuous recognition technology, ASR systems can be designed to be speaker dependent or speaker independent. *Speaker dependent* means that the user teaches the system about his or her particular voice by undergoing a prescribed enrollment procedure. In early systems, this procedure could last 2–3 hours, a significant up-front time cost for the user to even try the system. Current systems allow for streamlined installation and enrollment of approximately 30 minutes. The user has the option of giving the system additional training later as needed.

Speaker independent allows for walk-up-and-use applications because no enrollment procedure is necessary. This places a greater burden on the recognition algorithm because it must match the user's utterances to a generic model of spoken language rather than to one tuned to the user's particular speech characteristics. General-purpose dictation systems are rarely used in a speaker-independent fashion because of the degradation of recognition accuracy. Speaker-independent applications of speech recognition technology are found throughout daily life, however, in restricted vocabulary situations such as automated telephone operator services.

A third parameter of a speech recognition system is its vocabulary size, that is, the number of different words it can recognize. All early systems employed restricted vocabularies because of the limitations of the speech recognition technology and memory restrictions in early computers. Today, the built-in vocabulary of a system like Dragon NaturallySpeaking contains tens of thousands of words, and the user can add new words if necessary. As noted, restricted vocabulary applications do still have a role for constrained tasks that can benefit from speaker independence, such as voice operation of bank ATMs.

The literature on speech recognition is very diverse, covering all combinations of these three di-

[1] IBM Corporation, 1133 Westchester Avenue, White Plains, NY 10604; www.ibm.com.

[2] Dragon Systems, 320 Nevada Street, Newton, MA 02460; www.dragonsys.com.

[3] IBM Corporation, 1133 Westchester Avenue, White Plains, NY 10604; www.ibm.com.

[4] Dragon Systems, 320 Nevada Street, Newton, MA 02460; www.dragonsys.com.

mensions: discrete/continuous, speaker dependent/independent, and restricted/large vocabulary. This article focuses on the use of continuous speech recognition, for general-purpose dictation and computer access in a speaker-dependent fashion, by individuals whose primary need is accommodation of a physical disability. I emphasize the literature that is most relevant to that application; however, because there is no strong concentration of studies in any one combination of these dimensions, I have also reviewed some literature that relates to other aspects of speech recognition use, particularly the older discrete ASR systems, in order to offer a broader perspective.

Articles for review were located by keyword searches of all the science, health, and general journal databases available through the University of Michigan library, including CINAHL, Medline, InfoTrac, OCLC FirstSearch, and NARIC. The keywords used were "voice recognition" and "speech recognition." Additional articles were obtained by searching the proceedings from assistive technology conferences such as CSUN and RESNA, as well as from colleagues in the field.

USER PERFORMANCE WITH ASR

User-System Performance With Discrete Speech Recognition Systems

Discrete speech recognition (DSR) systems require the user to say each word individually, separated by a short pause from the preceding and following words. This was the first type of speech recognition system available, and discrete systems are generally no longer developed, supported, or maintained by commercial manufacturers for general-purpose computer input. However, reports of user performance with these systems are reviewed here for two reasons. First, the discrete technology is still used in some applications for people with disabilities. For example, the requirement of speaking slowly, word by word, can yield better recognition accuracy for users who have dysarthria and users who are ventilator-dependent or have other respiratory impairments. Some users with cognitive impairments also can benefit from the imposed word-by-word structure of a DSR system (Higgins & Raskind, 2000). Second, much of the scientific literature on user performance with speech recognition deals only with discrete systems, and its review provides a perspective from which to understand the literature on continuous speech recognition systems, which will be discussed in the next section.

Recognition Accuracy

One metric of user-system performance is the recognition rate, measured as the percent of words or utterances accurately recognized. All data in this section are from users without disabilities. Reports from early systems, using limited vocabularies of less than 70 words, letters, and digits, tend to range from 90% (Dabbagh & Damper, 1985; Shurick, Williges, & Maynard, 1985) to better than 95% accuracy (Frankish & Noyes, 1990; Noyes & Frankish, 1994). However, there is one report of very poor recognition, averaging 50% for 12 subjects after 10 days of practice, with a restricted-vocabulary system for control of home appliances (Noyes et al., 1989). More advanced systems, with vocabularies of several thousand words, have reported recognition rates of 94% to 98% for well-trained subjects (Karl, Pettey, & Shneiderman, 1993).

Recognition accuracy has been observed to be sensitive to a variety of factors, including time on task, task type, individual variation, user training and experience, vocabulary domain, background noise, and microphone position. Frankish, Jones, and Hapeshi (1992) noted that recognition accuracy on a digit entry task decreased by several percentage points within minutes of starting the task because of changes in the subjects' vocal style as the task proceeded. In a test on one of the earliest large-vocabulary systems (a precursor to IBM's VoiceType product), recognition accuracy for 12 subjects across two test sessions averaged 92% overall for a reading task (Brown & Vosburgh, 1989). In a composition task, however, these same subjects achieved only 85% accuracy. In both tasks, the users with the worst recognition experienced roughly three times as many errors as the best users. A second study using this same system showed that average recognition accuracy for 12 subjects improved from 91.6% to 94.4% after 4 weeks of coaching and retraining (Danis, 1989). In addition, the variation between best and worst subjects decreased over this same period. In a later study, in which seven readers dictated radiology reports to the IBM VoiceType system, recognition accuracy of 94% was achieved "under optimal conditions" for radiology content alone (Zemmel, Park, Maurer, Leslie, & Edlich, 1997). However, for general English words within the reports, the accuracy dropped to 77%. An increase in background noise caused a similar drop in accuracy, as did changes in microphone position. On this basis, the authors concluded that discrete SR was inadequate for medical emergency room and radiology

dictation based on observed performance in those environments.

Productivity

A second performance metric is overall user productivity. The results for the productivity of discrete SR systems relative to standard input methods are mixed, depending on the task and subject population. Many empirical studies that report task time for discrete speech recognition have focused on tasks that require relatively small vocabularies and were therefore better suited to the available ASR technology, such as alphanumeric data entry, computer programming, or entering text editing or spreadsheet commands.

Two studies report positive productivity results on constrained tasks for speech recognition relative to other input methods. Karl et al. (1993) observed that when subjects without disabilities used voice instead of a mouse to enter word processing commands, time for four specific tasks was reduced by 19%. In a study of four subjects using a speech-enabled computer-aided design system, users were able to complete 62% of the tasks within a fixed time when speech input was available, and only 38% of the tasks when speech was not available (Martin, 1989).

Four other studies on constrained tasks reported neutral or negative results for speech input. Leggett and Williams (1984) asked 24 subjects to enter and edit segments of specified computer programs, using speech input and standard keyboard input. During a 10-minute period, subjects could complete more tasks on average using the keyboard (70%) than they could using speech input (53%). However, given that subjects were vastly more experienced with keyboard use than with speech input, the authors interpreted these results as encouraging for the use of speech as a computer programming input modality. Indeed, approximately one third of the subjects had equivalent performance with voice and keyboard. A study with an early speech-controlled text editor showed no significant difference in speed relative to keyboard, even for inexperienced typists (Morrison, Green, Shaw, & Payne, 1984). In a similar experiment with spreadsheet tasks, however, subjects who used speech in combination with keyboard/mouse to enter commands took almost 50% longer to complete the experimental task than those who used only the keyboard and mouse (Molnar & Kletke, 1996). Valk (1990) found that a touchscreen was faster, and preferred, relative to speech input for 30 subjects who used both methods to input commands for an industrial control task.

For general dictation and text entry, performance with discrete SR systems has steadily improved over the years. For one early system, in which the user spelled out each word using the military alphabet, text entry rates were approximately 8 words per minute (wpm) (Dabbagh & Damper, 1985). By 1997, anecdotal reports of transcription rates for highly skilled users without disabilities approached 25–30 wpm (Mello, 1997), not generally competitive with skilled touch typists but perhaps sufficiently fast for many tasks. One empirical study suggests that extensive practice may not even be necessary to achieve these relatively high speeds (Dirks & Dirks, 1997). Forty-four undergraduates without disabilities were introduced to the Dragon Dictate system, and after initial enrollment and a short practice session, totaling roughly 1½ hours, their text entry rates on a 5-minute typing test averaged 28 wpm. These authors did not report the recognition accuracy their subjects achieved, nor did they describe how or whether subjects corrected misrecognition errors during the dictation task. They did note that errors overall were lower in the dictated text compared with similar text that subjects typed with the keyboard.

Studies Involving Users With Disabilities

I have found four studies relating to performance with discrete ASR systems by individuals whose primary need for ASR is the accommodation of physical disabilities. The first is a descriptive study in which 29 people with a variety of disabilities evaluated a voice-operated system for control of household appliances (Noyes et al., 1989). Specific results for recognition accuracy or success at operating the system were not reported. Overall, the researchers suggest that evaluators liked the idea of using voice control, but that this particular implementation was not very successful because of poor recognition.

The second study is of a single case that directly compared text entry rate with a discrete speech recognition system to mouthstick typing (Dalton & Peterson, 1997). The subject was a well-trained user with a high-level spinal cord injury. He achieved 20 wpm with the speech system, compared to 13 wpm using his mouthstick on a standard keyboard. This individual was also more accurate with speech input (98%) than with his mouthstick (95%).

The third report was a larger scale study involving a custom-built text-editing application, called

StoryWriter, which was designed for journalists who had incurred RSIs in their hands and upper extremities (Danis et al., 1994; Karat, Lai, Danis, & Wolf, 1999). Anecdotal reports of the accuracy achieved by these subjects ranged from percentages in the low 80s to the mid 90s. The authors attribute the relatively low accuracy to the challenging noise conditions experienced within the open newsroom. They do not report on productivity measures with the system, but note that some users credited the system with success in allowing them to return to work following their RSI.

Finally, Schwartz and Johnson (1999) surveyed 28 users of DSR systems regarding their perceptions of its effectiveness. All these users received one initial training session followed by one follow-up session approximately 1 month later. Although 75% of the subjects said they needed more training, 75% rated their overall experience with DSR as good or better. The remaining 25% were no longer using their systems. No measurements of accuracy or productivity were taken, but 51% of the subjects believed that they could enter text at 20 wpm or greater using speech.

Summary

Overall, reported accuracy for general dictation with DSR systems has ranged from 77% (Zemmel et al., 1997) to about 95% (Danis, 1989). Productivity for experienced users on text transcription tasks ranges from 20 to 30 wpm. The only empirical performance measurements reported for a user with a physical disability are for a single individual who achieved 98% accuracy and a 20-wpm text entry rate (Dalton & Peterson, 1997).

User-System Performance With Continuous Speech Recognition Systems

The heyday of discrete speech recognition has passed. Continuous speech recognition (CSR) systems that will recognize tens of thousands of words are now available for less than a few hundred dollars. Popular reviews of such systems suggest that users can employ natural speech at their natural pace, with resulting dictation speed of up to 150 wpm and more than 95% recognition accuracy (Mello, 1997; O'Malley, 1997; Zafar, Overhage, & McDonald, 1999). However, only a few scientific studies provide data with which to interpret the validity of these claims.

With continuous speech recognition, published reports are much more focused on general- or specific-purpose text entry, rather than on the restricted-vocabulary applications that mark so much of the discrete speech recognition literature. Overall, however, the results with continuous speech recognition are not markedly different from those reported for discrete speech recognition.

Recognition Accuracy

Devine, Gaehde, and Curtis (2000) compared the accuracy achieved with three different CSR systems. Accuracy was measured for 12 physicians performing a medical dictation task immediately after initial training and enrollment with each system. The "out-of-the-box" accuracy ranged from an average of 85% to 93% for the different systems. The study did not examine, however, whether accuracy changed with additional time and practice. A team at IBM, however, has looked at this issue, at least for a handful of subjects (Karat, Horn, Halverson, & Karat, 2000). Across three different CSR systems, novices who used speech for a text entry task achieved a recognition accuracy of 89%, whereas long-term (several years of use) and "extended-use" (20 hours of use) subjects achieved accuracy of 92% to 94%.

The IBM MedSpeak system for dictation of radiology reports is somewhat of an anomaly (Lai & Vergo, 1997). It is the only system of those reviewed that was employed in a speaker-independent fashion, meaning that the users did not have to undergo an enrollment procedure to teach their voice patterns to the system. Although speaker independence generally implies lower recognition accuracy, three MedSpeak users achieved an average accuracy of 97%. This is surprisingly good, yet, as I will discuss, the system was not judged to be good enough to replace the radiologists' current methods of report creation.

Productivity

With the MedSpeak system, the time required for a report to go from beginning of report creation to final signature decreased by 76.8%, compared with the traditional system of physician dictation followed by secretarial transcription (Lai & Vergo, 1997). In one sense, this represents a huge productivity gain in terms of how quickly a report is available to the referring physician. However, the use of Medspeak took more of the radiologists' time, about 35–50% more time than the simpler acts of dictating into a recorder and signing the transcribed report. So, despite the impressively high accuracy and overall productivity gains, the test group of physicians did not embrace the MedSpeak system as an acceptable alternative to the status quo.

In perhaps the most detailed study of text entry with speech recognition, 12 subjects were asked to perform four text entry tasks using a CSR system and four similar tasks using the standard keyboard and mouse (Halverson, Horn, Karat, & Karat, 1999; Karat, Halverson, Horn, & Karat, 1999). Subjects had approximately 2 hours of training and practice with the speech system before performing the tasks. Results showed that, using speech, subjects could transcribe text at an average of 13.6 wpm (including time required to correct recognition errors). With the keyboard and mouse, their transcription speed averaged 32.5 wpm. A separate group of four subjects, who were part of the research team, was followed across 20 hours of speech recognition use. These extended-use subjects achieved transcription speeds averaging 25.1 wpm. One expert, with several years of experience, tested at 31.0 wpm (Karat et al., 2000). These results suggest that initial performance with CSR is relatively slow but can be expected to improve significantly with practice. It is not clear from this study, however, whether, or at what point, text entry speed with speech would exceed that with keyboard and mouse.

This same research group also examined performance during simple composition tasks, in which users were asked to write their own text rather than transcribe it from hard copy. The absolute rates achieved on initial use, at 7.8 wpm for speech and 19.0 wpm for keyboard/mouse, were much slower than transcription speeds, but the relative difference between the two input methods remained roughly the same (Karat, Halverson, et al., 1999).

Gould, Conti, and Hovanyecz (1983) also focused on composition tasks. Using a human typist to enter the subjects' utterances, they simulated a variety of different types of "listening typewriters," including one that allowed continuous speech with unlimited vocabulary. This they called the "ultimate, but unachievable" speech recognizer. For eight novice subjects, the composition time for short letters was significantly faster with this system than with handwriting, with a composition rate of 11.5 wpm compared to around 8 wpm for writing, and subjects preferred this system to writing. In a second experiment, users with dictation experience were employed. They achieved slightly faster, but similar, performance as the novices, but they did not express the same degree of preference for the listening typewriter.

Studies Involving Users With Disabilities

I have not been able to find any empirical studies that report on the productivity of CSR use by people who are using it to accommodate physical disabilities. In the area of learning disabilities, one study suggests that CSR use can enhance reading and spelling ability, but no data on the performance of the students with the CSR system itself (i.e., recognition accuracy or text entry rate) were presented (Higgins & Raskind, 2000). Use of the technology to meet the needs of people with physical disabilities is often mentioned during general discussions of ASR, with the assumption that this user group may be more tolerant of the "immature" technology because they are limited in their ability to use the standard keyboard and mouse input methods (e.g., Danis & Karat, 1995; Seelbach, 1995; Shneiderman, 2000). It is unclear at this time whether this assumption is warranted. I simply have not been able to find any performance data on CSR use by people with physical disabilities.

TABLE 1. User performance with continuous speech recognition systems

User experience	Recognition accuracy	Text entry speed (wpm): Transcription	Text entry speed (wpm): Composition
Initial use	85–93%	14	7.8
Extended use	94%	25–30	Not available

Note. All data were collected from subjects without disabilities. Text entry speeds include time required to correct recognition errors. The transcription and composition speeds for subjects using standard keyboard and mouse are 32.5 and 19.0 wpm, respectively (wpm = words per minute). Sources: Devine et al., 2000; Karat, Halverson, Horn, & Karat, 1999; Karat, Horn, Halverson, & Karat, 2000.

Summary

Table 1 summarizes the published performance data on performance with CSR systems. For users without disabilities, speeds on text entry tasks using speech input are generally slower than speeds on the same tasks using keyboard and mouse. In interpreting these results, note that it can be difficult to make a fair comparison with more frequently used methods such as the keyboard because many subjects have already developed a high degree of skill with these methods. A more significant issue is that only a handful of empirical studies even involve users without disabilities, and we have found no relevant studies employing users who have physical disabilities.

BARRIERS TO SUCCESS WITH ASR

Although today's CSR systems offer the potential for input at 150 wpm, the published reports suggest a reality somewhere between 8 and 30 wpm. What accounts for this gap? This section reviews a range of possible barriers to successful use of ASR technology.

1. Is Speech Well-Suited to the Task?
2. Learning and Training Requirements
3. "Cognitive Costs" During ASR Use
4. Physical Interference With Other Tasks
5. Vocal Fatigue
6. Social and Work Environment Considerations
7. Technical Barriers

Is Speech Well-Suited to the Task?

It has been suggested that although speech is a natural medium for human-to-human communication, it just may not be that natural for human-to-computer communication. Newell, Arnott, Carter, and Cruickshank (1990) explored this idea in a partial replication of Gould et al.'s (1983) "listening typewriter" study, in which a human operator simulated a CSR system with an almost unlimited vocabulary. The 20 subjects achieved an average composition speed of 7.9 wpm using the simulated speech recognizer for a series of letter composition tasks. One reason for this relatively slow speed is that only 39% of the words uttered by subjects, on average, appeared in the final text. The majority of their words involved thinking out loud about what they wanted to say and negotiating with the "typewriter" to edit their text. The authors concluded that the use of speech for text composition may be inefficient and that designing an appropriate command structure for a speech interface is not a trivial endeavor.

ASR may be less beneficial in tasks like composing text, where "thinking" time is often the limiting factor rather than speed of getting the thoughts into the computer (Martin, 1989). In Gould et al.'s (1983) simulation, subjects spent only 10–18% of their time in a composition task actually dictating the words. So even if the speech technology enables faster word entry time, it may not affect the overall task time in a significant way. Some believe that ASR may be most beneficial for tasks involving frequent and short interactions between the user and the system, such as command entry or command and control of a computer operating system (Caton & Bethoney, 1998; Martin, 1989).

Learning and Training Requirements

Learning and training are among the most frequently mentioned issues influencing success with ASR technology (e.g., Biermann, Fineman, & Heidlage, 1992; Henry, 1998; Horner, Feyen, Ashlock, & Levine, 1993). For successful use, the system must learn how the user speaks, which typically involves a standard enrollment process in which the user says specific words or paragraphs in response to system prompts. The user must learn how to speak in such a way as to maximize recognition accuracy, by using clear and consistent tone, volume, and pronunciation. The user must also learn the most effective technique for identifying errors and correcting the system when the inevitable misrecognition occurs. One corporate reseller of ASR suggests that a minimum of 20 hours of training is necessary (Henry, 1998).

Finally, a memory burden is involved in learning the commands necessary for effective interaction with an ASR system (MacArthur, 1999). As with any sophisticated computer application, there are dozens of possible commands. Even a novice may need to learn a fairly large subset of these commands, particularly if he or she is using ASR in a "hands-free" or almost hands-free mode. The commands fall into four major categories:

1. Special words needed for dictation. This includes the words necessary to generate punctuation marks, numbers, or dates in the desired format and may also include knowledge of the military alphabet for spelling out words.
2. Special words needed for editing. This includes actions such as capitalization, tabs, blank lines, cursor movements, insertions, deletions, cutting, and pasting.
3. Spoken equivalents to application software or operating system commands. This includes commands for launching applications, opening and closing windows, copying and deleting files, and menu or button commands within a given application.
4. Commands needed to operate the ASR system itself. This includes basic operational commands for turning the microphone on and off, correcting recognition errors, and updating voice training, as well as more advanced commands for playing back dictation, using text-to-speech capabilities, or creating voice macros.

The design of the command vocabulary can play a significant role in the overall usability of the system. Garberg's (1995) study involving 26 subjects exposed to a set of 42 commands showed that the

particular choice of a word for a given command can make a big difference in a user's subsequent ability to recall it. In an attempt to enhance memorability, a voice command within an application may be chosen to have the same name as the menu or button equivalent (Jones, Frankish, & Hapeshi, 1992). However, this strategy generally results in short voice commands, which may be harder for the speech recognizer to discriminate. If special words are chosen for the voice commands to enhance recognizability, this results in essentially a second set of application commands for the user to learn. Furthermore, the user's age is a major factor. In Garberg's (1995) study, subjects over the age of 60 recalled less than half the commands recalled by younger subjects.

"Cognitive Costs" During ASR Use

In addition to the cognitive effort required up front to learn to use an ASR system, there are also cognitive costs involved once the system has been learned. The major source of these costs lies in the necessity of identifying and correcting recognition errors. These issues are discussed first, followed by an outline of other cognitive costs incurred during ASR use.

Performance Consequences of Recognition Errors

For skilled users, accuracy with ASR may approach that achieved with standard keyboard input. However, the consequences of errors with a keyboard compared to speech recognition are very different. The detection of keyboard errors generally requires relatively little attention, whereas most errors with ASR cannot be "felt," but must be specifically identified by looking at the display (Karat, Halverson, et al., 1999). Similarly, correction of keyboard errors is more straightforward, requiring an average of 3 seconds, compared to 25 or more seconds to repair a speech recognition error (Karat et al., 2000). This time adds up over the course of a several-page document. Even at 95% accuracy, an average of 5 errors will occur per 100 words dictated. At an error rate of only 5%, a five-page document with a total of 2,500 words contains 125 errors. The text itself could be dictated in 25 minutes or so, at a rate of 100 wpm, but the error correction could require an additional hour of time (Halverson et al., 1999). In other words, when transcribing text with ASR, about one third of the time may be spent reading the text, and the remaining two thirds of the time may be spent correcting recognition errors, even for users with high accuracy. For inexperienced users with lower recognition accuracy, the situation is even worse.

Anecdotally, 95% accuracy has been reported to be the threshold of user acceptance and successful use (Caton and Bethoney, 1998; Karat, Lai, et al., 1999). This criterion is based on extensive user feedback from trials of IBM speech recognition technology in applications ranging from radiology to journalism, as well as via observations during product testing at PC Week Labs.[5] A study by Casali, Williges, and Dryden (1990) showed how seemingly small differences in recognition accuracy can lead to fairly large differences in performance and to user acceptability. Eighteen subjects used a simulated speech recognizer to perform a data entry task involving digits, short words, and a few commands. Among the conditions simulated were three different levels of recognition accuracy: 91%, 95%, and 99%. As expected, task completion time significantly improved with each increment in recognition accuracy. However, the improvement was not linear. Jumping from 91% to 95% accuracy yielded a 22% improvement in task time, while jumping from 95% to 99% yielded a smaller improvement of 14%. The accuracy level also had a significant effect on subjects' acceptability ratings, with, again, the biggest jump in acceptability occurring on the jump from 91% to 95% accuracy. These results highlight the importance of being precise when talking about recognition accuracy (i.e., anecdotally reporting that it is "above 90%" gives too broad a range) and suggest that 95% accuracy may indeed provide the best combination of attainable accuracy with good user performance and acceptability.

Identifying Recognition Errors

Because any ASR system makes mistakes, users must first decide how they are going to identify errors. The two basic strategies are (1) a "proofreading" strategy, in which the user focuses first on dictating the entire text and then corrects errors as a group in a second step; and (2) an "in-line" strategy, in which the user identifies and corrects errors as they occur (Karat, Halverson, et al., 1999). When people use nonspeech input methods to enter text, they typically use an in-line strategy. Indeed, many keyboard users often detect and correct their errors in-line without even looking at the display. The typist can "feel" when an error has occurred. Novice users of ASR, perhaps transferring

[5] PC Week Labs, a subsidiary of ZDNet, 650 Townsend Street, San Francisco, CA 94103; www.zdnet.com.

their experience with nonspeech methods, typically use an in-line strategy more often than a proofreading strategy to identify errors (Karat, Halverson, et al., 1999). In-line identification has the advantage of taking care of the error while the correct intent is still fresh in the user's mind. However, it has the disadvantage of requiring considerable attention and possible distraction from the primary task. When an error is detected in-line, performance of the user's primary task stops while the error is corrected, potentially disrupting the rhythm of primary task activities.

The proofreading strategy has the advantage of reducing the amount of the user's attention that is diverted from the primary task. Because the user, in effect, does not worry about errors until after the entire text, or at least a major portion of it, is completed, he or she can focus more completely on dictating or composing the text. This may be why more frequent use of the proofreading strategy is correlated with greater expertise with speech recognition (Karat, Halverson, et al., 1999). One drawback, especially when recognition accuracy is not above 95%, is the risk of forgetting the original intent when there is a significant time delay between when the error occurred and when it is identified. In a typical example taken from one of our research subjects, the user actually said, "Suddenly, Tom opened his eyes and sat up." The system's attempt at recognizing this sentence was "Suddenly, open his license that up." Seeing that garbled sentence even a few minutes after the utterance, one could easily forget the original intent. This also makes it harder for someone other than the speaker to do the correction process. Systems such as Dragon NaturallySpeaking have a feature that allows for audio playback of what a user said as a way to refresh the user's memory in these situations. But use of this feature can bring its own costs, such as the knowledge of when and how to use it, the time required for its use, and the cognitive effort required to map what one hears in the playback to the errors one sees on the screen. It is preferable to avoid this situation altogether.

Correcting Recognition Errors

Once an error has been detected, the next step is generally to correct it. Although there are numerous specific methods, both within and between different ASR systems, the correction methods generally fall into one of three categories:

1. Use a "Scratch That" or "Undo" command to erase the immediately previous utterance. Then redictate the utterance. This general method has been named SCRUNDO by the research group at IBM (Halverson et al., 1999).
2. Select the erroneous word or phrase by a voice command (e.g., "Select 'license that'"). Then redictate the correct word or phrase in its place (e.g., "eyes and sat").
3. Select the erroneous word or phrase by a voice command and open the error correction dialogue (e.g., "Correct 'license that'"). Then choose from one of the following correction methods:
   a. select the correction from a pick list of alternate recognitions;
   b. spell the word by voice; or
   c. spell the word by another input method such as the keyboard.

   For either of the spelling correction methods, the pick list of alternates will change as the user types in letters for the word. If the desired word or phrase appears in the numbered pick list, it can be selected at any time by giving the appropriate voice command (e.g., "Choose 3").

Although all of these methods *can* be used to correct recognition errors, only the last method, to open up the correction dialogue, is actually *intended* to be used for this purpose (Halverson et al., 1999). The SCRUNDO method is primarily designed for correcting utterances that are misspoken or inadvertent vocalizations, such as coughs. The select-then-redictate method is actually intended for making editing changes to the text. Only the correction dialogue method teaches the system more about how to correctly interpret the user's voice. By consistently using the correction dialogue to fix recognition errors, the user helps the recognizer improve subsequent accuracy. Conversely, when the user relies on SCRUNDO or select-redictate, the system does not learn from its recognition mistakes, and recognition accuracy is much less likely to improve.

The way people actually correct errors has been shown to be quite different than the way the system designers originally intended. In examining the correction behavior of 12 novice ASR users, a research group from IBM found that the most common strategy among people who are not specifically coached to use a particular method is to select then redictate the incorrect word (Halverson et al., 1999; Karat, Halverson, et al., 1999). Subjects used this method 38% of the time. They deleted then re-entered (SCRUNDO) 23% of the time. They used the correction dialogue only 8% of the time. The remaining 32% of the correction time was spent fixing recognition errors that occurred

during the correction process itself (Karat, Halverson, et al., 1999).

Several problems are associated with this situation, in which three methods can be used for fixing recognition errors but only one is really intended for this purpose. First, the mere fact that there are three different methods creates a cognitive load on the user. There is a load associated with learning the purpose of each method, the appropriate conditions for its use, and the exact steps for proper use. Even the choice between two simple methods can cost a user 1–2 seconds each time a choice is made (Olson & Nilsen, 1988). With ASR systems, making the right choice is made more difficult by the fact that all three methods only *appear* to be equivalent. That is, externally, the results of each method are the same, but internally their effects on the system are quite different. ASR systems as currently designed do nothing to help users realize that difference or to reinforce the proper choice of correction methods.

Given the challenge in learning to make the "correct" choice, people often do what seems easiest or what they remember most readily. Hence, they often use the more straightforward SCRUNDO or select-redictate methods rather than the correction dialogue. In doing so, they inadvertently restrict the recognition accuracy they can achieve. In addition, these seemingly simpler methods may not work that well when used for fixing recognition errors. For example, when using the select-redictate method, the recognition accuracy for the redictated text averages only 47% (Halverson et al., 1999). This means that the majority of the time, users must repeat this method multiple times to get the system to recognize the intended word or phrase correctly.

Appropriate use of the correction dialogue provides a powerful opportunity to teach the system more about the user's speech patterns. However, its use is associated with cognitive loads, which may be part of the reason people fail to use it as often as they should. The user must first tell the system what text he or she wants to correct. On the surface, this appears to be a straightforward process in which the user issues a command such as "Correct *carefully.*" However, if the word to be corrected occurs more than once in the document, the system may not select the intended word, even if it is the one closest to the current cursor position. This can be very frustrating with a common word such as *the,* leading to situations where a user repeats a "Correct *the*" command numerous times before the system will select the intended one. In addition, because the correction dialogue window covers up part of the text, the user may not even notice that the word being corrected is not the one intended. To avoid these problems, the user can select a short phrase, which is more likely to be unambiguous. This is effective but, of course, it must be learned or taught.

Once the correction dialogue box is open, there are several additional sources of cognitive load. One is the presence of the "pick list," a list of candidate words or phrases that can be selected to replace the word currently being corrected. This can be a powerful feature for reducing the letter-by-letter spelling that is required, as long as the desired word actually appears in the pick list, but it must be visually scanned and commands for its use must be remembered. A second source of cognitive load is that, typically, at least a portion of the desired word or phrase must be spelled out letter by letter. For hands-free users, who do the spelling by voice, successful recognition of spoken spelling requires using the vocal pattern suggested in the system manual (e.g., "say the letters continuously and quickly, not one at a time" [Dragon Systems, 2000, p. 19]), or knowledge of the military alphabet to reduce the ambiguity between letter sounds. To avoid any problems with spoken spelling, the user may type on the keyboard if he or she is able, but in either case the knowledge of *how* to spell the desired words may also contribute to cognitive load. Finally, the user must remember which methods are appropriate for use within the correction dialogue. For example, in Dragon NaturallySpeaking, the replacement word can be entered only by spelling or by selection from the pick list. Users may commonly, and erroneously, attempt to redictate the word within the correction dialogue, then wonder why it is not recognized.

In summary, the correction of recognition errors *can* be a smooth process, but in reality, users often choose an inappropriate method that may be attractively simple in the short term but reduces the chances of long-term success. The challenge of correction is compounded by recognition errors that occur within the correction process itself. Assistive technology clinicians applying ASR to the needs of their clients have long been aware of this issue and have often had to devise specialized training methods in an attempt to reduce the impact of misrecognitions (Horner et al., 1993).

Preventing Recognition Errors

Because the cost of recognition errors is relatively high, users have a strong incentive to prevent as many of them as possible. If successful, er-

ror prevention can reduce the time and irritation associated with recognition errors, but prevention techniques themselves may have some costs. The typical advice given to reduce recognition errors is to maintain a consistent, natural volume and pace while speaking; to speak naturally but clearly, enunciating each word, as a newscaster might speak; to avoid saying "um" or thinking out loud; and to speak in phrases rather than one word at a time (Dragon Systems, 2000). Although the word "naturally" is sprinkled throughout this advice, in reality it is not natural for most people to speak like a newscaster, much less to maintain that speech consistently over time. Learning and maintaining this vocal pattern takes time and cognitive energy.

Even with some conscious effort to continue to speak in a consistent and appropriate manner, a consistent voice pattern is not always easy to achieve. Frankish et al. (1992) observed that recognition accuracy can drift with time on task. In a numeric entry task involving 16 subjects, accuracy declined by about 4% over the first half hour of the task. Follow-up studies revealed that this was not because of fatigue, but was primarily the result of changes in speech patterns during performance of the task as compared to the initial enrollment process. This problem is exacerbated when the user has a cold or other temporary condition that affects vocal quality.

A final error prevention method involves anticipating errors before they occur. For example, if a user is about to enter an acronym, proper noun, or other word that might be expected to result in a recognition error, he or she can spell the word out rather than speaking it. This requires the foresight to anticipate a problem and the correct recall of how to do this (it typically requires a special command, such as "Spell," followed by the letter-by-letter spelling of the word). Both of these actions require cognitive energy and have an associated cost.

Other Cognitive Costs

Some subtler forms of cognitive cost may also be involved in the use of ASR systems. First, the "choice-of-methods" issue often exists on a gross level. Because most ASR users also have a nonspeech method that they can use for computer input, they are continually faced with the decision of which method to use in which circumstance (Jones et al., 1992).

Second, the need to correct recognition errors means that use of the ASR system is far from transparent; the user's attention must be shared between the primary task domain and the output of the ASR system. Frankish and Noyes (1990) supply empirical evidence for the view that error detection and correction activities can interfere with performance of the primary task. During a numeric data entry task, subjects were asked to enter four-digit numbers using speech input. Their memory for the digit strings was nearly flawless in cases where no recognition errors occurred during entry. However, of entries in which a recognition error did occur, almost 12% also contained a memory recall error.

Third, Shneiderman (2000) has recently raised an interesting hypothesis about whether the use of the speech channel itself, even in conditions of perfect recognition, can interfere with short-term memory and problem-solving ability. He points out that more cognitive resources are required for speaking than for physical activity. In relation to computer use, this means that it is easier to type and think simultaneously than it is to speak and think simultaneously. The evaluation of the StoryWriter ASR system for journalists provides some anecdotal evidence for this viewpoint (Danis et al., 1994).

Physical Interference With Other Tasks

A typical user of ASR is connected to the system via a headset microphone, which is plugged in to the computer's sound card or USB port. For some users, this tethering represents a nuisance, interfering with the performance of other tasks that may be necessary during computer use. For example, performing pressure relief may require backing away from the work desk, making it necessary to remove, then replace, the headset each time pressure relief must occur. Similarly, accessing paper materials or other items that are not directly within reach of the computer may also require removal of the headset. If the phone rings, the user may also need to remove the headset and remember to turn off the ASR microphone before answering. Individuals who use head-controlled pointers such as the HeadMaster Plus[6] must either attempt to wear two headsets simultaneously or find a different solution.

Some of these issues can be solved by a more thorough integration of technologies or improved workstation design. In addition, wireless alternatives to the default microphone do exist, but they

[6] Prentke Romich Company, 1022 Heyl Road, Wooster, OH 44691; www.prentrom.com.

represent an added cost and may not always work as well as the hard-wired headset microphone.

Vocal Fatigue

There has also been some suggestion in the literature that use of speech recognition can have unanticipated physical consequences. Although decreasing the biomechanical load on upper extremities and postural systems, ASR can exact a greater load on the vocal system. This may cause only minor discomfort for some, but Kambeyanda, Singer, and Cronk (1997) report on four individuals who developed chronic vocal stress requiring treatment after 1 year of using a DSR system. Because use of CSR requires fewer starts and stops and allows for more natural vocal patterns, its stress on the voice should be less than older discrete systems (Grubbs, 2000). But there are anecdotal reports of vocal fatigue and injury among CSR users, although the prevalence of the problem is unclear (Grubbs, 2000).

Social and Work Environment Considerations

The work environment in which ASR is used can have a significant impact on user performance. Key considerations include placement and stability of the microphone, workplace background noise, and the extent to which ASR use disturbs others in the environment (Gardner-Bonneau, 1999). People who work in a noisy environment or simply like to listen to moderately loud music may find that they need to retrain the system frequently or accept lower accuracy levels. The need to speak aloud also reduces the user's privacy in some situations. Zemmel et al. (1997) found ASR unsuitable for hospital emergency room or radiology environments because of background noise and other environmental issues.

Technical Barriers

Finally, although speech recognition technology has advanced significantly over the past decade, there are still some technical barriers to successful, hassle-free use. First, even though today's personal computers have amazing hardware capabilities at a fairly low cost, use of ASR still requires a major share of computing resources (Blotzer, 2000; Lenker, 1998). The "minimum system requirements" listed on the packages of most ASR systems significantly underestimate what is required for satisfactory performance. During the course of normal computer use, multiple applications may be open simultaneously, which can significantly degrade the recognition response time (Gardner-Bonneau, 1999). Because ASR tends to "hog" system resources, some users may not open it for use unless they are certain that they are going to use it (e.g., for typing a large amount of text). Second, there remain various problems in compatibility between the ASR software and other applications. In the past, a user's only option was to dictate into a specialized ASR application, then cut and paste the results into the desired application, such as a word processor. Today's ASR systems are much more closely integrated with user applications, and users can generally dictate directly into the most popular application programs. However, the ASR system may not work in the same way with all software. With Dragon NaturallySpeaking, for example, some applications support "select-and-say" commands issued by voice, but others do not. Still others may not work with speech input at all (Blotzer, 2000).

Summary

The literature suggests that a variety of barriers to successful use of ASR systems may exist, although specific information on their impact and prevalence is not always available. The goal in reviewing these barriers is not to suggest that they are insurmountable or that successful use of ASR is not achievable. The purpose is to show that there may be reasons why ASR is not a panacea for all individuals who need alternatives to standard methods of computer input, as well as to highlight why we might need to study these issues in more depth. We have a long way to go to determine how large or small these barriers truly are for people who have disabilities.

As ASR technology advances—as it certainly will—will these barriers still exist? Fundamental issues such as the suitability of speech as a medium for computer interaction, the possibility of interference between speech and other cognitive activities, and the impact of ASR use within social and work environments are relevant regardless of the sophistication of the speech recognition technology. Some potential barriers, particularly those associated with recognition errors, may be mitigated as expected recognition accuracy improves beyond 95%. However, in the foreseeable future, perfect accuracy is not realistic, so the need for identifying and correcting recognition errors is expected to remain for quite some time. Effective interaction will still require cognitive tasks such as development of a mental model of how the system works and how to speak to it, memorization of spe-

cific commands related to ASR use, an understanding of which error correction strategy is best suited to a particular situation and the particular steps required to execute that strategy, and shared attention between the task domain and operation of the ASR system.

The presence of these cognitive activities is what distinguishes use of an ASR system from the natural speech of a conversation. The need to frequently engage in even short-duration cognitive actions during human-computer interaction can be both tiring and time-consuming to the user (Card, Moran, & Newell, 1983; Koester & Levine, 1996).

SUMMARY AND IMPLICATIONS

This review sheds some light on the research questions listed at the beginning of the article, and it also reveals major gaps in our current understanding. The questions are repeated here, with a brief summary of the extent of our knowledge.

1. How well are speech recognition systems meeting the needs of users who have disabilities, particularly for the accommodation of physical disabilities? Schwartz and Johnson's (1999) survey of discrete speech users suggests that those who continue using their systems (about 75%) are satisfied with them. We have found no data on the satisfaction of continuous speech users.
2. What is the range of productivity that a user of an ASR system can expect? How does this depend on the characteristics of both the user and the task? Relatively little empirical information is available on performance of general computer access tasks using CSR systems (see Table 1). For text entry rate, reports on users without disabilities range from 8 to 30 wpm. We have seen no data on the use of continuous speech systems for common tasks such as command and control of the operating system desktop or surfing the Web. And none of the literature we found focuses on use of CSR to accommodate physical disabilities. The lack of data relative to users who have physical disabilities is particularly striking, especially in light of the common assumption that people with disabilities will be enthusiastic and early adopters of ASR technology (Danis & Karat, 1995; Seelbach, 1995; Shneiderman, 2000).
3. What is the learning curve associated with ASR systems? How long does it take to develop a high degree of proficiency? Only a fraction of the existing data describes skilled use, or how skill develops over time. Studies by Karat and colleagues (Halverson et al., 1999; Karat, Halverson, et al., 1999) suggest that today's continuous speech systems can be used with some success after only 2 hours of training, but that skill is still developing even after 20 hours of use over a period of weeks.
4. Are there human factors costs that may partially counteract the benefits of using ASR systems? Many studies suggest that human factors issues and other barriers are involved in the use of ASR. Cognitive and perceptual overhead, the potential for vocal stress, and related challenges may, in some cases, combine to significantly reduce user performance and comfort. However, the magnitude and prevalence of these effects are only partially understood.
5. If so, are there ASR system assessment and implementation methods that can reduce the impact of these costs and result in improved user satisfaction and productivity? Several clinicians who work with speech recognition users have shared some of their hard-earned wisdom about issues such as appropriate training methods and technical procedures for successfully installing and maintaining ASR systems (Cantor, 2001; Grott & Schwartz, 2001; Horner et al., 1993; Lenker, 1998). However, we have found no research studies specifically focused on developing and evaluating clinical interventions that involve ASR, and these would be a valuable complement to the qualitative information that already exists.

Opportunities abound for research in this area. We need to better understand user satisfaction, learning and training requirements, and user productivity. We also need greater insight about the extent to which the potential barriers discussed here affect user performance and how to reduce their effects. A key component is the extensive involvement of people with disabilities in any study. At the University of Michigan Rehabilitation Engineering Research Center, we are in the midst of a 3-year project that is addressing these issues in the particular context of ASR use for text entry by users with physical disabilities. Carefully designed research and the integration of that research into clinical practice are important parts of effectively applying today's speech recognition systems and helping users to reap the full benefits of the technology.

Acknowledgments: This research is supported by a Rehabilitation Engineering Research Center grant #H133E980007 from the National Institute

on Disability and Rehabilitation Research (NIDRR), U.S. Department of Education.

REFERENCES

Biermann, A., Fineman, L., & Heidlage, J.F. (1992). A voice- and touch-driven natural language editor and its performance. *International Journal of Man-Machine Studies, 37,* 1–21.

Blotzer, M. J. (2000). The master's voice. *Occupational Hazards, 62*(12), 55.

Brown, N. R., & Vosburgh, A. M. (1989). Evaluating the accuracy of a large-vocabulary speech recognition system. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 296–300). Santa Monica, CA: Human Factors and Ergonomics Society.

Cantor, A. (2001). Speech recognition: An accommodation planning perspective. In *CSUN 2001 Conference Proceedings.* Retrieved June 13, 2001, from http://www.csun.edu/cod/conf2001/proceedings/0190cantor.html

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Lawrenceville, NJ: Erlbaum.

Casali, S. P., Williges, B. H., & Dryden, R. D. (1990). Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. *Human Factors, 32,* 183–196.

Caton, M., & Bethoney, H. (1998). It isn't easy to wreck a nice beach. *PC Week, 15*(9), 157.

Cavalier, A. R., & Brown, C. C. (1998). From passivity to participation: The transformational possibilities of speech-recognition technology. *Teaching Exceptional Children, 30*(6), 60–65.

Coleman, C., & Meyers, L. (1991). Computer recognition of the speech of adults with cerebral palsy and dysarthria. *Augmentative and Alternative Communication, 7*(1), 34–42.

Dabbagh, H. H., & Damper, R. I. (1985). Text composition by voice: Design issues and implementations. *AAC: Augmentative and Alternative Communication, 1,* 84–93.

Dalton, J. R., & Peterson, C. Q. (1997). The use of voice recognition as a control interface for word processing. *Occupational Therapy in Health Care, 11,* 75–81.

Danis, C. M. (1989). Developing successful speakers for an automatic speech recognition system. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 300–304). Santa Monica, CA: Human Factors and Ergonomics Society.

Danis, C. M., Comerford, L., Janke, E., Davies, K., DeVries, J., & Bertrand, A. (1994). StoryWriter: A speech-oriented editor. In *CHI '94 Conference Companion* (pp. 277–278). Boston, MA: Association for Computing Machinery.

Danis, C., & Karat, J. (1995). Technology-driven design of speech recognition systems. In *Proceedings of DIS '95* (pp. 17–24). Boston, MA: Association for Computing Machinery.

De La Paz, S. (1999). Composing via dictation and speech recognition systems: Compensatory technology for students with learning disabilities. *Learning Disability Quarterly, 22,* 173–182.

Devine, E. G., Gaehde, S. A., & Curtis, A. C. (2000). Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *Journal of the American Medical Informatics Association, 7,* 462–468.

Dirks, R., & Dirks, M. J. (1997). Introducing business communication students to automated speech recognition: Comparing performance using voice input and keyboarding. *Journal of Education for Business, 72*(3), 153–156.

Doyle, P. C., Leeper, H. A., Kotler, A., & Thomas-Stonell, N. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development, 34,* 309–316.

Dragon Systems. (2000). *User's guide to Dragon NaturallySpeaking 5.* Newton, MA: Author.

Frankish, C., Jones, D., & Hapeshi, K. (1992). Decline in accuracy of automatic speech recognition as a function of time on task: Fatigue or voice drift? *International Journal of Man-Machine Studies, 36,* 797–816.

Frankish, C., & Noyes, J. (1990). Sources of human error in data entry tasks using speech input. *Human Factors, 32,* 697–716.

Garberg, R. B. (1995). Automatic speech recognition applications: A study of methods for defining command vocabularies. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 203–207). Santa Monica, CA: Human Factors and Ergonomics Society.

Gardner-Bonneau, D. (1999). The future of voice-interactive applications. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 295–299). Boston: Kluwer Academic.

Goette, T., & Marchewka, J. T. (1994). Voice recognition technology for persons who have motoric disabilities. *Journal of Rehabilitation, 60,* 38–41.

Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM, 26,* 295–308.

Grott, R., & Schwartz, P. (2001, June). *Speech recognition from alpha to zulu.* Instructional course and handout presented at the RESNA conference, Reno, NV.

Grubbs, L. (2000). Watch what you say. *PC World.com* [Online]. Available: http://www.pcworld.com/features/article/0,aid,16766,pg,1,00.asp

Halverson, C. A., Horn, D. B., Karat, C., & Karat, J. (1999). The beauty of errors: Patterns of error correction in desktop speech systems. In *Proceedings of Human-Computer Interaction—INTERACT '99* (pp. 133–140). Edinburgh, Scotland: IOS Press.

Henry, J. (1998). Combatting the high cost of doing business—Personal odyssey leads to creation of speech-recognition training VAR. *Computer Reseller News, 810,* 138.

Higgins, E. L., & Raskind, M. H. (2000). Speaking to read: The effects of continuous vs. discrete speech recognition systems on the reading and spelling of children with learning disabilities. *Journal of Special Education Technology, 15*(1), 19–30.

Horner, J. E., Feyen, R. G., Ashlock, G., & Levine, S. P. (1993). Specialized approach to teaching voice recognition computer interfaces. In *Proceedings of the RESNA '93 Conference* (pp. 449–451). Washington, DC: RESNA.

Jones, D., Frankish, C. R., & Hapeshi, K. (1992). Automatic speech recognition in practice. *Behavior and Information Technology, 11,* 109–122.

Kambeyanda, D., Singer, L., & Cronk, S. (1997). Potential problems associated with use of speech recognition products. *Assistive Technology, 9,* 95–101.

Karat, C., Halverson, C. A., Horn, D. B., & Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the CHI '99 Conference* (pp. 568–574). Boston, MA: Association for Computing Machinery.

Karat, J., Horn, D. B., Halverson, C. A., & Karat, C. (2000, April). *Overcoming unusability: Developing efficient strat-*

egies in speech recognition systems. Poster session presented at CHI 2000, the ACM Conference on Human Factors in Computer Systems, The Hague, Netherlands.

Karat, J., Lai, J., Danis, C., & Wolf, C. (1999). Speech user interface evolution. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 1–35). Boston: Kluwer Academic.

Karl, L. R., Pettey, M., & Shneiderman, B. (1993). Speech versus mouse commands for word processing: An empirical evaluation. *International Journal of Man-Machine Studies, 39,* 667–687.

Koester, H. H., & Levine, S. P. (1996). The effect of a word prediction feature on user performance. *AAC: Augmentative and Alternative Communication, 12,* 155–168.

Lai, J., & Vergo, J. (1997). MedSpeak: Report creation with continuous speech recognition. In *Proceedings of the CHI '97 Conference* (pp. 431–438). Boston, MA: Association for Computing Machinery.

Leggett, J., & Williams, G. (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies, 21,* 493–520.

Lenker, J. (1998). Naturally speaking. *Paraplegia News, 52*(6), 37.

MacArthur, C. (1999). Overcoming barriers to writing: Computer support for basic writing skills. *Reading and Writing Quarterly, 15,* 169–192.

Martin, G. (1989). The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies, 30,* 355–375.

Mello, J. P. (1997). NaturallySpeaking: Voice recognition breakthrough. *PC World, 15,* 80–81.

Molnar, K. K., & Kletke, M. G. (1996). The impacts on user performance and satisfaction of a voice-based front-end interface for a standard software tool. *International Journal of Human-Computer Studies, 45,* 287–303.

Morrison, D. L., Green, T. R. G., Shaw, A. C., & Payne, S. J. (1984). Speech-controlled text-editing: Effects of input modality and of command structure. *International Journal of Human-Computer Studies, 21,* 49–63.

Newell, A. F., Arnott, J. L., Carter, K., & Cruickshank, G. (1990). Listening typewriter simulation studies. *International Journal of Human-Computer Studies, 33,* 1–19.

Noyes, J. M., & Frankish, C. R. (1992). Speech recognition technology for individuals with disabilities. *AAC: Augmentative and Alternative Communication, 8,* 297–303.

Noyes, J. M., & Frankish, C. R. (1994). Errors and error correction in automatic speech recognition systems. *Ergonomics, 37,* 1943–1957.

Noyes, J. M., Haigh, R., & Starr, A. F. (1989). Automatic speech recognition for disabled people. *Applied Ergonomics, 20,* 293–298.

Olson, J. S., & Nilsen, E. (1988). Analysis of the cognition involved in spreadsheet software interaction. *Human Computer Interaction, 3,* 309–349.

O'Malley, C. (1997). Dragon slays the voice robot. *Popular Science, 251*(5), 61.

Raskind, M. H., & Higgins, E. L. (1999). Speaking to read: The effects of speech recognition technology on the reading and spelling performance of children with learning disabilities. *Annals of Dyslexia, 49,* 251–281.

Schwartz, P., & Johnson, J. (1999). The effectiveness of speech recognition technology. In *Proceedings of RESNA '99 Conference* (pp. 77–79). Washington, DC: RESNA.

Seelbach, C. (1995). A perspective on early commercial applications of voice-processing technology for telecommunications and aids for the handicapped. *Proceedings of the National Academy of Sciences of the United States of America, 92,* 9989–9990.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM, 43,* 63–65.

Thomas, J. C., Basson, S., & Gardner-Bonneau, D. (1999). Universal access and assistive technology. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 135–145). Boston: Kluwer Academic.

Valk, M. A. (1990). Comparing touchscreen to speech input in the control of a simple batch process. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 419–423). Santa Monica, CA: Human Factors and Ergonomics Society.

Zafar, A., Overhage, M., & McDonald, C. J. (1999). Continuous speech recognition for clinicians. *Journal of the Medical Informatics Association, 6,* 195–204.

Zemmel, N. J., Park, S. M., Maurer, E. J., Leslie, L. F., & Edlich, R. F. (1997). Evaluation of VoiceType dictation for Windows for the radiologist. *Medical Progress Through Technology, 21,* 177–180.