# USER PERFORMANCE WITH CONTINUOUS SPEECH RECOGNITION SYSTEMS

Heidi Horstmann Koester and Simon P. Levine, University of Michigan

## ABSTRACT

The University of Michigan Rehabilitation Engineering and Research Center (UM RERC) on Ergonomics is just beginning a three-year study on user performance with continuous speech recognition systems. 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 paper outlines what is known about user performance with speech recognition systems and presents the plan of a project designed to enhance understanding in this area.

## STATEMENT OF THE PROBLEM

Continuous speech recognition (CSR) 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 simply by speaking into a microphone, without requiring the speaker to pause between each word. For users whose severe physical disabilities require them to have some sort of hands-free access to a computer, CSR is an attractive option compared to potentially less efficient methods such as mouthstick typing or two-switch Morse code. For users whose use of "standard" manual input methods has led to a repetitive stress injury or other serious biomechanical stress, CSR may provide a productive alternative to continued discomfort and exacerbation of the injury, freeing users from keyboard use and its associated postural constraints.

While the promise of CSR is enormous and sales of voice recognition systems continue to grow, some basic questions regarding user performance with voice recognition have not been satisfactorily addressed. These include:

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

This study will provide a fuller understanding of the role of CSR systems in meeting the needs of people with disabilities by addressing these questions. The project will apply this new understanding to devise and evaluate methods of improving user performance with CSR systems.

## BACKGROUND

### *Voice Recognition (VR) Systems*

Automatic voice recognition (VR) has been under development since the early 1970's. Early systems could recognize only a handful of discrete words or utterances. By the late 1980's, 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 word VR systems have continued to improve in vocabulary size and recognition accuracy. In 1997 a major breakthrough in VR technology occurred with the first consumer-affordable continuous speech recognition system. Continuous speech allows users to speak at their natural pace and rhythm, resulting in faster and potentially more satisfactory interaction.

### *User-System Performance for Discrete VR Systems*

The vast majority of existing literature on user performance with VR deals with discrete systems only. One metric of user-system performance is the recognition rate, measured as percent of words accurately recognized. One early system, with a limited vocabulary of 70 utterances, was able to recognize up to 90% accurately for a well-trained able-bodied subject (Dabbagh and Damper, 1985). More advanced systems, with several thousand word vocabularies, have reported recognition rates of 94% to 98% for well-trained subjects with and without severe upper extremity disabilities (Karl et al., 1993; Dalton et al., 1997). This is comparable to the accuracy of a skilled typist or mouthstick user (Dalton et al., 1997).

A second performance metric is overall user productivity. Discrete VR systems may or may not provide improved performance relative to standard input methods, depending on the task and subject population. For example, Karl et al.

(1993) observed that when able-bodied subjects used voice instead of a mouse to enter word processing commands, time for four specific tasks was reduced by 19%. In a similar experiment for spreadsheet tasks, however, subjects performed more slowly with the voice interface (Molnar, 1996). Zemmel (1996a, 1996b) concluded that discrete VR was inadequate for medical emergency room and radiology dictation based on observed performance in those environments.

For general dictation and text entry, which is an important VR application for users with disabilities, performance with discrete VR 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 of approximately 8 words per minute (wpm) were achieved (Dabbagh and Damper, 1985). By 1997, rates for highly skilled able-bodied users approached 25 - 30 wpm (Mello, 1997), not generally competitive with skilled touch typists but perhaps sufficiently fast for certain workplace dictation tasks. There are very few reports directly comparing text entry rate with VR to other input methods for users with physical disabilities. In one single case study, a well-trained user with a high level spinal cord injury achieved 20 wpm with a discrete VR system, as compared to 13 wpm using his mouthstick on a standard keyboard (Dalton et al., 1997).

*Human Factors Issues in Discrete VR Systems*

Human factors issues with discrete VR systems are an important influence on user performance. While several such issues have been mentioned in the literature, including learning/training, other cognitive and perceptual aspects of interacting with a VR system, the capacity of the human vocal system, and the task environment, there are very few specific reports examining their quantitative impact on user-system performance.

Learning and training is one of the most frequently mentioned issues (e.g., Horner et al., 1993; Biermann et al., 1992). For successful use of a discrete VR system, the system must learn how the user speaks, which typically involves a standard enrollment process where the user says specific words in response to system prompts. The user must learn how to speak in such a way as to maximize recognition accuracy, by using a consistent tone of voice and the proper pause between each word. The user must also learn the most effective technique for correcting the system when the inevitable misrecognition occurs.

The time and effort involved in learning effective use of a discrete VR system, as well as to repeatedly decide on the optimal correction strategy for each misrecognition, are examples of the "cognitive cost" of using the system. Other examples are the conscious effort required to speak each word discretely and then attend to the system' s recognition response. This response typically takes the form of a "pick list" of candidate words that potentially match the user' s utterance; the user has the option of visually searching this list for the correct word and choosing it verbally. The presence of these cognitive activities is what primarily distinguishes use of discrete VR from speaking naturally in a conversation. The need to frequently engage in them during human-computer interaction can be both tiring and time-consuming to the user (Card, Moran, and Newell, 1983; Koester and Levine, 1996). For example, in a clinical case study the time involved in correcting misrecognized words accounted for more than 50% of the task time (Koester and Hilker, 1995, unpublished).

There has also been some suggestion in the literature that use of voice recognition can have unanticipated physical consequences. While decreasing the biomechanical load on upper extremities and postural systems, discrete VR can exact a greater load on the vocal system. This may cause only minor discomfort for some, but Kambeyanda et al. (1997) report on four individuals who developed chronic vocal stress requiring treatment after one year of using a discrete VR system.

Finally, the conditions of the work environment in which VR is used can have a significant impact on user performance. Key characteristics include placement and stability of the microphone, workplace background noise, and the extent to which VR use disturbs others in the environment. Zemmel et al. (1996a, 1996b) found VR not suitable for hospital emergency room or radiology environments due to background noise and other environmental issues.

Subjective comments of discrete VR users corroborate the presence of some significant human factors issues. Even users who enjoy using VR overall have commented on short term memory challenges and consequent interference with task domain, "tedious" nature of talking all the time, voice fatigue, and the frustration of attending to and correcting errors (Biermann et al., 1992).

*User-System Performance for Continuous VR Systems*

Continuous speech recognition (CSR) systems which 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 100 wpm and 95% recognition accuracy (Mello, 1997; O' Malley, 1997). However, we have found no empirical validation of these claims in the literature, either for "mainstream" or physically disabled users.

While the ability to use natural speech at a natural pace would be expected to reduce the impact of human factors issues on performance with CSR, many of the cognitive and perceptual activities required for interaction with a discrete VR system are still present with continuous speech recognition. In particular, misrecognition errors still occur, and need to be identified and corrected. This process is in fact somewhat more complicated than with discrete VR, since there are more choices for when to check for errors and how to correct for them. Effective interaction still requires development of a mental model of how the system works, an understanding of which error correction strategy is best suited to a particular situation, shared attention between the task domain and output of the CSR system, and memorization of specific commands for executing the chosen error correction strategy. To date there have been no reports of how these activities impact user performance and satisfaction or how to design effective training interventions to reduce any negative impact they may have.

*Implications*

The above review reveals the following major gaps in understanding user performance with voice recognition systems:

1. There is only a small amount of user performance data with discrete VR systems and almost none on continuous speech recognition systems. Very little of the existing data focuses on users who are physically disabled.
2. Human factors issues involved in the use of VR are briefly discussed in most studies, suggesting that cognitive/perceptual overhead and the potential for vocal stress may in some cases combine to significantly reduce user performance and comfort, but the magnitude and prevalence of these effects have not been reported.
3. An ergonomics perspective has not been a primary focus. Interventions which may enhance user performance, such as redesigning the work task, sharing input between voice and other channels, or customizing user training, have not been explicitly discussed or studied.

## EXPERIMENTAL DESIGN AND METHODS

*Overview*

The UM RERC project plans to address these gaps through a series of experiments, divided into a baseline phase, an intervention phase, and an evaluation phase. The baseline phase consists of three experiments, each examining user performance with CSR systems from a different perspective. In the intervention phase, new intervention methods will be developed based on findings of the baseline phase. The evaluation phase will evaluate the success of these new methods at enhancing user performance by repeating some of the protocols employed in the baseline phase.

*Baseline Phase*

Clinical Case Review. Clinical files of clients who are using CSR systems will be examined in order to get a clear understanding of current practices and level of success. For each client, the case review will assess client and clinician expectations of performance with the CSR system at time of recommendation and answer such questions as: Why was a CSR system recommended? What other possibilities were rejected and why? Were specific productivity goals established for this client? What training methods were used? Clients will then be contacted directly by telephone to assess their current usage patterns and satisfaction with the system.

Longitudinal Study of Novice Users. This will track the user performance of twelve CSR novices from first introduction of the system through the development of at least "intermediate" proficiency. User performance on paragraph dictation tasks will be measured at four intervals from system introduction to final recommendation, then once each at system delivery, 6 months post-delivery, and 12 months post-delivery. Sessions will be videotaped, allowing for measures such as recognition rate and overall words per minute, as well as detailed analyses of time costs associated with recognition errors and other component activities. Users will also complete a survey of user satisfaction and other subjective measures at the end of each test session. Both the quantitative performance and survey data will be analyzed. Subjective information from clients and their clinicians will also be used to better understanding the factors that influence user performance (e.g., human factors issues, service delivery issues).

Study of Expert Performance. The performance and satisfaction of at least six individuals who are frequent, highly practiced users of CSR systems will be measured using methods similar to the longitudinal study described above. Data will be compared to that from the novices at 12 months post-delivery to determine whether novice performance after one year approaches that of experts. Subjective information gathered from the expert users will build understanding of how expertise is achieved.

*Intervention and Evaluation Phases*

The purpose of the intervention phase is to design new or revised intervention methods in order to enhance user performance with CSR systems. These will be based on the findings of the baseline phase. For example, baseline data from novices and experts will provide a much clearer picture of the range of performance that can be expected. This can then be compared to the demands of a task environment as well as the expected performance with other input methods to predict and test the potential of CSR either alone or combined with other methods in meeting a client' s needs. As a second example, detailed analysis of videotaped sessions will yield a better understanding of any performance "bottlenecks," e.g., due to difficulty in choosing the best error correction method, difficulty in remembering system commands, or other cognitive challenges in using the system. Methods to reduce the effect of these can then be implemented and tested, possibly through revised training methods or simple memory aids. If viable interventions are identified, longitudinal studies with a new novice group will be repeated to more fully assess their effectiveness. Interventions will also be employed with the original novice group and a comparison made with the new novice group exclusively trained with the interventions.

We expect this three-year study to provide insight into the human factors issues involved in the use of continuous speech recognition systems as well as information regarding service delivery methods to enhance user performance with CSR systems.

## REFERENCES


1. Biermann, A., Fineman, L., and Heidlage, J.F. (1992). A voice- and touch-driven natural language editor and its performance. *International Journal of Man-machine Studies, 37*, 1-21.
2. Card, S.K., Moran, T.P., and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Lawrenceville NJ: Erlbaum Associates.
3. Dabbagh, H.H. and Damper, R.I. (1985). Text composition by voice: design issues and implementations. *AAC Augmentative and Alternative Communication, 1*, 84-93.
4. Dalton, J.R. and Peterson, C.Q. (1997). The use of voice recognition as a control interface for word processing. *Occupational Therapy in Health Care, 11*, 75-81.
5. Koester, H.H. and Hilker D. (unpublished, 1995). Clinical case study: performance of a client with C5-6 quadriplegia on the VoiceType voice recognition system.
6. Horner, J.E., Feyen, R.G., Ashlock, G., and Levine, S.P. (1993). Specialized approach to teaching voice recogntion computer interfaces. *Proceedings of RESNA '93*, 449-451.
7. Kambeyanda, D., Singer, L., and Cronk, S. (1997). Potential problems associated with use of speech recognition products. *Assistive Technology, 9*, 95-101.
8. Karl, L.R., Pettey, M., and Shneiderman, B. (1993). Speech versus mouse commands for word processing: an empirical evaluation. *International Journal of Man-machine Studies, 39*, 667-687.
9. Koester, H.H. and Levine, S.P. (1996). The effect of a word prediction feature on user performance. *AAC Augmentative and Alternative Communication, 12*, 155-168.
10. Mello, J.P. (1997). NaturallySpeaking: Voice recognition breakthrough. *PC World, 15*, 80-81.
11. Molnar, K.K. and 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.
12. O' Malley, C. (1997). Dragon slays the voice robot.*Popular Science, 251: 61.*
13. Ottenbacher, K.J. (1986). *Evaluating Clinical Change: Strategies for Occupational and Physical Therapists*. Baltimore: Williams and Wilkins.
14. Zemmel, N.J., Park, S.M., Schweitzer, J. et al. (1996a). Status of VoiceType dictation for Windows for the emergency physician. *Journal of Emergency Medicine, 14*, 511-515.
15. Zemmel, N.J., Park, S.M., Maurer, E.J. et al. (1996b). Evaluation of VoiceType dictation for Windows for the radiologist. *Medical Progress and Technology, 21*, 177-180.


## ACKNOWLEDGMENTS

This research is supported by a Rehabilitation Engineering Research Center grant from the National Institute on

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

**ADDRESS**
Heidi Horstmann Koester, Ph.D.
2408 Antietam
Ann Arbor MI 48105
hhk@umich.edu