Overcoming Unusability: Developing efficient strategies in speech recognition systems

John Karat, Daniel B. Horn*, Christine A. Halverson+, and Clare-Marie Karat

IBM T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532 USA
+1 914 784 7832
ckarat, jkarat@us.ibm.com

*University of Michigan
Collaboratory for Research on Electronic Work
701 Tappan Street, C2420
Ann Arbor, MI 48109-1234
danhorn@umich.edu

+SRI International
Computer Human Interaction Center
333 Ravenswood Av.
Menlo Park, CA 94025
krys@speech.sri.com

ABSTRACT

This paper describes changes in user error correction strategies over time in the use of large vocabulary desktop automatic speech recognition (ASR) systems. Users with minimal practice with such systems were found to have considerable difficulty with error correction [1,2]. Users with more extensive use were found to have improved overall performance compared to initial use subjects. This is attributed to development of multimodal strategies for error correction rather than to significantly improved speech recognition rates or use of speech-based error correction techniques. These results point to the importance of multimodal interaction in the acceptance of speech recognition technology.

Keywords

Speech recognition, multimodal interaction, input techniques, speech user interfaces

INTRODUCTION

In the past few years, large vocabulary desktop speech recognition systems from a number of vendors have become widely available (the studies reported here used systems from Dragon Systems, IBM, and L&H). Misleading claims – such as the ability to enter text at the rate of 140 words per minute – have encouraged many to try the technology. However, it is a not-too-well-kept secret that within a few months of acquisition most of the systems purchased end up not being regularly used. Recent work [1,2] sheds some light on the problems of early system use. These problems seem to lie more in the difficulty of error correction than in the initial entry of text into the systems. In this paper we briefly present some findings related to changes in user behavior with extended use of the systems.

Our earlier work painted a rather unflattering picture of the initial user experience with current automatic speech recognition (ASR) systems. Users were less productive with ASR than with keyboard and mouse on a variety of tasks (averaging about 14 corrected words per minute [cwpm] for voice input for short transcription tasks versus about 32 cwpm with typing and pointing). The problem, at least for novices, was not text entry speed or error rate as much as it was in the time and effort required making corrections. While the number of errors in text entry was similar for the two techniques, it took much longer to make corrections in the ASR systems (an average of 26 seconds for the ASR systems versus 3 seconds for keyboard-mouse). Comparing text entry with speech recognition and typing, novice users can generally speak faster than they can type and have similar numbers of speech and typing errors, but take much longer to correct dictation errors than typing errors.

In this paper we examine what happens with extended use of ASR systems to see if and how users overcome these difficulties. For users who gain some experience with ASR systems, we wanted to know whether they become more efficient with initial techniques or develop new error correction strategies.

DEVELOPING STRATEGIES FOR EFFICIENT ERROR CORRECTION

There can be several ways to account for the difficulty that novice users have dictating with ASR systems. First, it may simply take time to master the interaction techniques. If this is the case, we might expect to see improvements in performance without changing the user's error correction strategies. Second, novice users might be selecting inefficient speech correction strategies. In general, there are many ways in which an error might be recognized and corrected, and some methods might be more efficient in different contexts. Third, while speech seems to offer promise for entering text, speech techniques might not be efficient for error correction. Novices in our study persisted in attempting to use speech to make corrections [1,2], however, this might not be the optimal strategy.

With experience a number of things might happen. We might expect some performance improvement due to adaptation – both by the user to the system and by the system to the user – which results in better recognition accuracy. Second, we might expect users to learn more about the different correction mechanisms within speech and to better select the most appropriate speech commands in different correction contexts. Third, we might expect users to better integrate speech with other input techniques and select methods across modalities more efficiently.

	Speech	Correction Steps	Error Cascades	Multimodal Corrections	Keyboard Corrections	Speech Corrections
Novices (N=12)	13.6 cwpm	7.3	32%	8%	2%	90%
Extended Use (N=4)	25.8 cwpm	3.8	3.3%	42%	11%	47%
Expert A (N=1)	31.0 cwpm	3.3	0.6%	60%	38%	2%

Table 1. Summary of error correction in transcribed text for various subjects.

Table 1 summarizes the data from several studies. Novice user data reported previously [1,2], and the associated extended use data from that study are presented with data from an evaluation of several regular users of ASR systems. (Data from a single user with several years experience using the technology who we believe to be typical of this population is presented here).

Overall measures of performance improve with experience. Error rates (errors divided by total words) were about 11% for Novices, 8% for Extended Use subjects and 6% for our Expert. In general, throughput improved with experience, though it remained no better than our measure of keyboard-mouse productivity from the novices (32.5 cwpm). While improved overall recognition rate accounts for some of the performance improvement, a more substantial improvement is accounted for by a reduction in time taken per correction. Extended Use and Experts subjects take about half of the steps to make a correction that novices do (3.8 and 3.3 steps per correction compared to 7.3). Most of this reduction is accounted for by a decrease in the number of times making an error correction results in a new error (an error cascade).

Experience seems to provide improved error correction, generally through avoidance of compound errors. Why is this? Associated with these changes is a movement toward the use of keyboard and mouse (and away from using speech) in making error corrections. Novices use speech alone to make corrections in almost all cases, and in general only used other devices when speech failed repeatedly. For the Extended Use subjects, we can see an increase in the use of multimodal corrections (corrections in which keyboard or mouse are used to either select the text to be corrected, or to type in the correction), and an increase in keyboard-mouse only corrections. For the expert subject, there are very few instances of speech only corrections (only 2%). Here, almost all text selection is done with the mouse - and while speech is still used for reentering corrected text, keyboard only corrections account for 38% of the episodes. These differences are extremely striking, as the chief weakness of speech correction is the potential for misrecognitions which can lead to cascading errors – in fact, for our novice users, single word speech corrections were misrecognized nearly half the time! We do not have sufficient data to draw any conclusions about when users elect to completely abandon speech for corrections and when they elect to use it for text reentry (although we can suggest that experience provides the user with clues about when redictation is likely to succeed or fail).

CONCLUSIONS

From these observations we conclude that there is a tendency with experience to employ keyboard-mouse techniques over speech-based techniques for making error corrections in ASR systems. A mouse is preferred by over speech for selecting text. Additionally, the use of these techniques contributes to improved overall user performance and satisfaction with the technology. However, for our limited sample, the performance did not reach levels that are remarkably above keyboard-mouse performance. Significantly, all of the experienced users in this study (except for one expert who was physically incapable of using keyboard-mouse) continue to use keyboard-mouse as their standard means of text entry in their day to day work.

This does not mean that we think the technology has been shown to be unusable. There are certainly contexts in which it can been seen as an acceptable input technique (e.g., for devices in which keyboard-mouse are not practical or for contexts in which the users' hands are busy). However, for the existing human-computer interaction paradigm of workstation interaction, we do not expect to see a mass user declaration of keyboard obsolescence in the near future.

REFERENCES

1. Halverson, C., Horn, D., Karat, C., & Karat, J. (1999). The beauty of errors: patterns of error correction in desktop speech systems, in *Proceedings of INTERACT '99* (Edinburgh, September 1999), IOS Press, 133-140..
2. Karat, C., Halverson, C., Horn, D., & Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems, in *Proceedings of CHI '99* (Pittsburgh PA, May 1999), ACM Press, 568-575.