EFFECT OF ERROR CORRECTION STRATEGY ON SPEECH DICTATION THROUGHPUT

James R. Lewis
International Business Machines Corp.
West Palm Beach, Florida

The eight participants in this experiment used two different commercially available speech recognition dictation systems to complete a variety of reading transcription tasks. Participants enrolled fully in both systems. They received training in two correction strategies for both systems: multimodal correction (voice plus mouse plus keyboard) and hands-free correction (voice-only), and used both strategies during the experiment. The key findings were:

- Both dictation systems were equally accurate.
- Throughput (corrected words per minute) was significantly (63%) faster using multimodal correction.
- Speaking rates were the same for both systems and correction strategies, averaging around 105-110 utterances (words and commands) per minute.
- Correction speeds for the multimodal correction strategy (13.2 seconds per correction) were significantly faster than (a little more than twice as fast as) those for hands-free correction (29.1 seconds per correction).
- At the end of the experiment, participants indicated they significantly preferred the multimodal correction strategy.

THE MODEL

Background

Within the last few years, there have been tremendous breakthroughs in the availability of affordable computerized speech dictation applications. Of particular importance has been the introduction of systems that accept continuous speech rather than discrete speech as input. The accuracy of these systems has also been improving, although achieving 100% accurate transcription remains a goal rather than a reality.

Given accuracy less than 100%, it is necessary for users of dictation systems to correct errors. Although it's obvious that faster error correction necessarily yields faster overall throughput of correct text, modeling the effect of speed of error correction on throughput yields some interesting observations.

Performance Model of Computerized Dictation Throughput

There are three key variables that drive throughput (where throughput is defined as the number of correct words produced per minute, or cwpm): (1) the accuracy of the speech recognition system, (2) the speaking rate of the user, and (3) the time required to correct an error. Figure 1 shows projected throughputs for a hypothetical 1000-word document as a function of recognition accuracy ranging from 50% to 100% (at intervals of 5%), speaking rates of 100 or 150 wpm (S100 and S150 respectively), and correction speeds of 5, 10, 15, or 30 seconds per correction (spc) (E05, E10, E15 and E30 respectively). Figure 2 focuses on the projected throughputs as a function of accuracy ranging from 90 to 100% (at intervals of 1%).

The figures show that at lower levels of recognition accuracy, it makes almost no difference whether the speaking rate is 100 or 150 wpm. The key determinant of throughput given poor accuracy is correction speed. This pattern holds until the

higher levels of accuracy. When a recognizer has perfect accuracy, correction speed is irrelevant, but even when recognition is as high as 95%, correction speed still plays a key role in determining throughput, although speaking rate is beginning to play an important role as well. It appears that with anything other than a perfect (or very near perfect) recognition system, the speed with which users can make corrections has a dramatic effect on the throughput of the system.

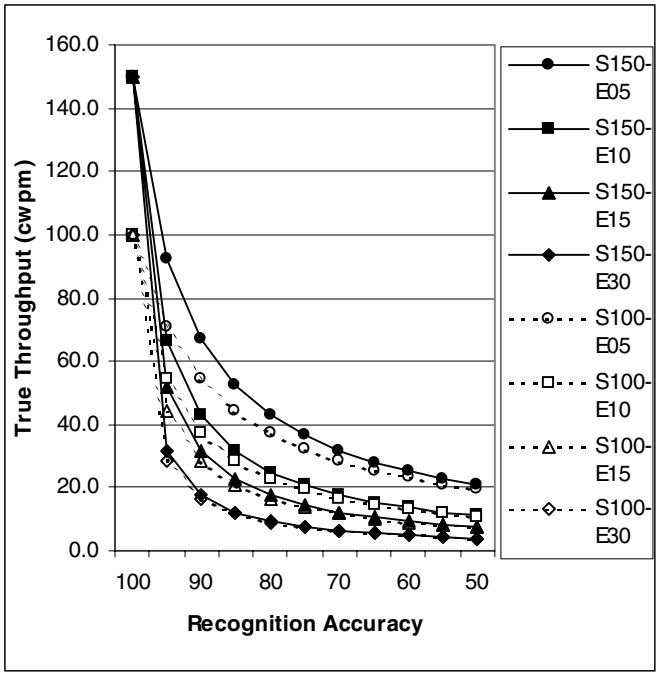

Figure 1. A performance model of computerized dictation throughput for recognition accuracy ranging from 50 to 100%. (Points are throughput measured in correct words per minute.)

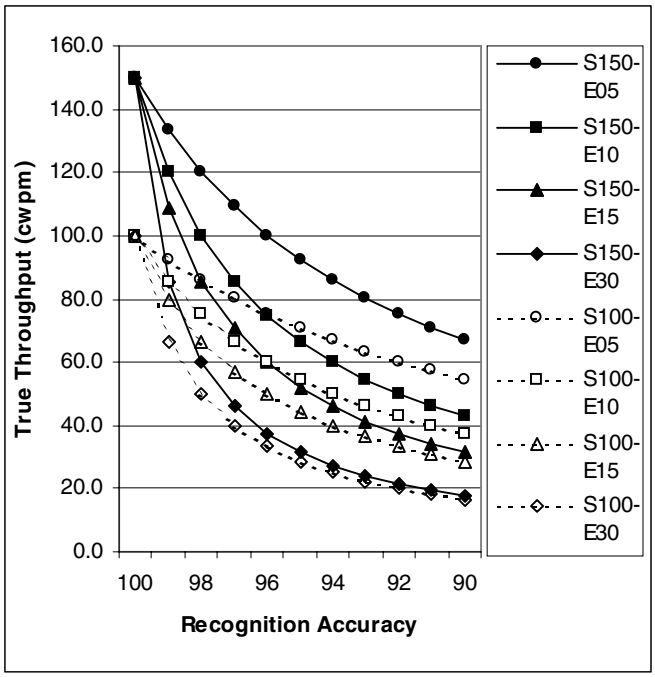

Figure 2. The performance model of computerized dictation throughput for recognition accuracy ranging from 90 to 100%. (Points are throughput measured in correct words per minute.)

Correction Strategy and Correction Speed

Most current computerized dictation products offer several ways for users to accomplish the correction of misrecognition errors. Two different methods of error correction are hands-free (voice-only) and multimodal (using voice, mouse, and keyboard). One way to reduce the time spent correcting errors is to choose the most efficient correction method.

Many users of speech products have the initial belief that correction by voice-only will be faster than using the mouse and keyboard along with voice commands as appropriate. Human factors performance analysis of the correction task would suggest otherwise because a multimodal approach allows for more overlapping of actions. For example, as the user prepares to issue a voice command, he or she can simultaneously be positioning the cursor on the misrecognized word. Also, in many (though not all) situations, the use of the mouse or keyboard can be inherently faster than

a voice command. It takes about 150 ms to press a key (Card, Moran, and Newell, 1983), but about 200 ms to issue a one-syllable speech command (Massaro, 1975). Speech commands longer than one syllable will take, on average, about 200 ms per syllable, and most speech commands will be longer than one syllable.

To empirically evaluate the magnitude of the difference in correction speed and throughput as a function of correction strategy, the following experiment was designed and conducted.

THE EXPERIMENT

Purpose

The purpose of this study was to compare the effect of two speech dictation correction strategies on the speed of true throughput (corrected words per minute, or cwpm). The two strategies were hands-free (voice only) and multimodal (voice, mouse and keyboard) correction. Because multimodal correction allows users to overlap voice commands with other manual activities, it seems likely that multimodal correction should be faster than hands-free correction, with corresponding improvements in the true throughput (corrected words per minute, or cwpm) of dictation. If true, this information would be useful to people who use computerized speech dictation products to help them be as efficient as possible when producing documents.

Method

Participants. Eight people (four males and four females) participated. All participants were familiar with Windows® 95. Each participant worked for about four hours to complete the study. The participants had mixed experience with prior speech dictation products (from expert to novice). Unfortunately, gender and experience were confounded variables (with males having more experience than females in this study). For this reason, neither gender nor experience receive treatment in the statistical analyses.

Materials and Procedure. Participants used the released versions of two different commercially available speech dictation products installed on a Micron™ Millenia™ LXE system (a Pentium (TM) 200 MHz MMX processor running Windows 95 with 64 Mb extended memory, on-board SoundBlaster™ compatible sound system). The microphone attached to the system was the one shipped with the product in use at the time.

Participants enrolled fully in both systems. They received training in two correction strategies for both systems: multimodal correction and hands-free correction, and used both strategies during the experiment.

Experimental Design. The study was a within-subjects design providing sufficient counterbalancing for the independent variables in this study. Each participant completed four experimental conditions (two correction strategies by two dictation systems) reading four test texts, one per condition. Across the experiment, texts were paired with experimental conditions an equal number of times.

Dependent measures

Accuracy. This measure is a percentage. It is the number of words and commands correctly recognized and acted upon during a dictation session (times 100) divided by the total number of utterances (words and commands) issued.

True Throughput (correct words per minute, or cwpm). This measure is a rate. It is the number of correctly recognized words issued per minute in a dictation session (where the time for a dictation session includes the time for word entry and correction of misrecognitions). This measure is important because it is the truest measure of throughput from a user's point of view. It does not take into account the number of correctly recognized commands because commands exist in a dictation system to control text characteristics and to allow recovery from recognition error.

Net Words Per Minute (nwpm). This measure is a rate. It is the number of correctly recognized words issued per minute in a dictation session (where the time for a dictation session includes only the time for word entry).

Speaking Rate. This is the number of words and commands uttered per minute while reading.

Rate of Correction. This is the average number of seconds per correction.

After-scenario questionnaire (ASQ). This measure is a rating taken with a 3-item standardized questionnaire (Lewis, 1995). Administration of this questionnaire immediately followed completion of each individual dictation task.

Results

Accuracy. Collapsed across correction strategy, the accuracy of the two dictation systems was almost identical (differing by a statistically and practically insignificant 0.8%).

Throughput. ANOVA on the throughput data indicated a significant main effect for correction strategy ($F(1,6)=21.0$, $p=.004$). The throughput difference as a function of correction strategy was of surprisingly large magnitude, with 31.0 wpm for efficient multimodal correction, 63% faster than the hands-free throughput of 19.0 wpm (values for both strategies collapsed over system).

Speaking Rate. Participants' rate of speaking was independent of system, correction strategy, or gender, and averaged about 105-110 words and commands per minute.

Correction Rate. ANOVA indicated a significant main effect of correction strategy ($F(1,6)=18.7$, $p=.005$). The difference in correction speeds (collapsed over system) as a function of strategy were large, with the correction speed for the multimodal strategy (13.2 seconds per correction) over twice as fast as hands-free correction (29.1 seconds per correction).

After-Scenario Questionnaire. The only significant effect from the ANOVA on ASQ scores was a main effect of correction strategy ($F(1,6)=9.1$, $p=.02$). Participants gave a better rating for the multimodal strategy (2.5) than the hands-free strategy (3.4). (The ASQ scale runs from a best possible score of 1.0 to a worst possible score of 7.0.)

Net Words Per Minute. The ANOVA on nwpm indicated no significant differences. Correlation data (Table 1) showed that nwpm bears no significant relationship to cwpm (corrected words per minute, true throughput as defined in this paper), but does correlate highly with measures of speaking rate. This indicates that its use as a measure of true throughput (with corresponding implications for productivity) would be misleading.

Table 1. Correlations Between nwpm, Throughput (cwpm) and Speaking Rate (wpm) for each experimental condition

Correlation	X-MM	X-HF	Y-MM	Y-HF
nwpm-Throughput	0.03	0.39	0.33	-0.03
nwpm-Speaking Rate	0.96	0.98	0.96	0.94

Table Notes

With a sample size of 8, only correlations that exceed .62 are significant ($p<.05$). Correlations that exceed .83 are highly significant ($p<.005$).

X=one product, Y=the other product, MM=multimodal, HF=hands-free

Final Correction Strategy Preferences. After completing all the tasks in this experiment, participants chose their preferred correction strategy. 7 of 8 participants indicated that they preferred multimodal correction over hands-free (observed percentage: 87.5%; 90% binomial confidence interval: 53.0 - 99.5%).

Discussion

The key findings from the experiment were:

- Both dictation systems were equally accurate.
- Throughput (corrected words per minute) was significantly (63%) faster using multimodal correction.
- Speaking rates were the same for both systems and correction strategies, averaging around 105-110 utterances (words and commands) per minute.
- Correction speeds for the multimodal correction strategy (13.2 seconds per correction) were significantly faster than (a little more than twice as fast as) those for hands-free correction (29.1 seconds per correction).

- At the end of the experiment, participants indicated they significantly preferred the multimodal correction strategy.

These experimental results demonstrate the greater efficiency of multimodal correction compared to hands-free correction and are consistent with the model of throughput presented in the introduction. This effect held across two different recognition systems, indicating its generalizability. Granted, some users of computerized dictation products might want to use a hands-free strategy (either as a result of difficulties using mice and keyboards or simply as a result of a desire to control their computers by speech only), but designers of such systems should be aware of the relative efficiencies of the two correction strategies, and provide appropriate guidance to their users.

TRADEMARKS

Micron and Millenia are trademarks of Micron Corp.

Pentium is a trademark of Intel Corp.

SoundBlaster is a trademark of Creative Labs Corp.

Windows is a registered trademark of Microsoft Corp.

REFERENCES

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

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57-78.

Massaro, D. W. (1975). Preperceptual images, processing time, and perceptual units in speech perception. In D. W. Massaro (ed.), *Understanding language: An information-processing analysis of speech perception, reading, and psycholinguistics* (pp. 125-150). New York, NY: Academic Press.