# Model-based and Empirical Evaluation of Multimodal Interactive Error Correction

**Bernhard Suhm**
*Interactive Systems Laboratories*
Carnegie Mellon University/
Universität Karlsruhe
bsuhm@ira.uka.de

**Brad Myers**
*Human Computer Interaction Institute*
Carnegie Mellon University
bam@cs.cmu.edu

**Alex Waibel**
*Interactive Systems Laboratories*
Carnegie Mellon University /
Universität Karlsruhe
ahw@cs.cmu.edu

## ABSTRACT

Our research addresses the problem of error correction in speech user interfaces. Previous work hypothesized that switching modality could speed up interactive correction of recognition errors (so-called *multimodal* error correction). We present a user study that compares, on a dictation task, multimodal error correction with conventional interactive correction, such as speaking again, choosing from a list, and keyboard input. Results show that multimodal correction is faster than conventional correction without keyboard input, but slower than correction by typing for users with good typing skills. Furthermore, while users initially prefer speech, they learn to avoid ineffective correction modalities with experience. To extrapolate results from this user study we developed a performance model of multimodal interaction that predicts input speed including time needed for error correction. We apply the model to estimate the impact of recognition technology improvements on correction speeds and the influence of recognition accuracy and correction method on the productivity of dictation systems. Our model is a first step towards formalizing multimodal (recognition-based) interaction.

## Keywords

multimodal interaction, interactive error correction, quantitative performance model, speech and pen input, speech user interfaces.

## INTRODUCTION

As speech recognition technology matures, speech user interfaces have begun to replace traditional interfaces. For example, speech systems replace live human operators in automated call centers, and voice input is available as an alternative to keyboard input in automatic dictation systems. Speech recognition technology, however, comes with inherent limitations. Our research addresses the problem of recognition errors due to imperfect recognition. Assuming that recognition remains imperfect despite continued progress in recognition algorithms (even human recognition is imperfect), we investigate *interactive* error correction methods. Efficient and graceful error correction is crucial in the design of speech user interfaces (as noted, for example, in [2]).

We conducted an informal survey of interactive correction methods used in current speech recognition applications. These applications used the following four correction methods: repeating using continuous speech (from here on *respeaking*), typing, choosing from a list of alternative words, and clarification dialogues. (Clarification dialogues allow the user to make corrections within the context of a spoken dialogue.) What are the drawbacks of these methods? Previous research shows that both correction by respeaking and by choosing from a list can be ineffective in continuous speech applications [17]. Correction by typing assumes keyboard input. Keyboard input is not available in some applications, and it is effective only for users with good typing skills. Clarification dialogues are appropriate mostly for one category of speech user interfaces, so-called conversational speech applications [3]. Our work currently focuses on non-conversational applications.

Previous work hypothesized that error correction could benefit from switching multimodal [12, 13]. It is commonly believed that redundant use of several modalities contributes to the ease of human-to-human communication. Multimodal human-computer interfaces aim to benefit from redundant use of modalities in human-computer interaction in similar ways. Our research explores the benefits of multimodal interaction in the context of error correction. Since words that are confused by automatic recognition systems tend to be different across modalities, switching between modalities for correction should eliminate repeated recognition errors; but to-date, no empirical study with real recognition systems has confirmed this hypothesis.

To test this hypothesis, we have implemented multimodal interactive correction methods [10, 17]. We integrated multimodal correction in an automatic dictation system to build a prototype *multimodal* dictation system. Recognition output is displayed on the screen, and the user locates recognition errors by selecting misrecognized words. The user corrects by deleting, inserting, or replacing misrecognized words. In multimodal correction, there is a choice of different correction modalities:

CHI '99 Pittsburgh PA USA
Copyright ACM 1999 0-201-48559-1/99/05...$5.00

repeating input using continuous speech, (verbal) spelling, handwriting, and editing using gestures drawn on a touch-sensitive display (e.g., deleting words with an X or scratching gesture, or changing the position of the cursor with a caret gesture). This paper describes a user study of interactive multimodal error correction and presents a predictive performance model of multimodal, recognition-based interaction. *Recognition-based* means that user input must be interpreted using an automatic recognition system [13].

We begin by describing our empirical evaluation of interactive error correction on a dictation application. The main goal of the user study was to provide empirical evidence for the hypothesis that multimodal correction expedites error correction in speech user interfaces. The study shows that unimodal correction (using the same modality for input and correction) is ineffective, and that multimodal correction is effective. Furthermore, the study compares current interactive correction methods (with and without keyboard input) with multimodal correction methods. Finally, the study investigates which modalities users prefer by analyzing usage frequencies of different modalities. To test whether accuracy influences modality choice, we correlated usage frequency with modality accuracy. Our longitudinal analysis of usage frequencies shows that learning effects do occur.

We then describe a simple performance model of multimodal interaction that we apply to error correction in order to extrapolate the results of the user study. Predictive models are particularly useful in multimodal interfaces, since they abstract from current recognition performance. Previous work proposed a quantitative performance model for speech-only interfaces [11]. This model predicted task completion time using critical path analysis. The model accounts for imperfect recognition by modeling error correction as repetition of input. The authors reported a good match between model predictions and empirical data. To our knowledge, the model was never applied to multimodal interaction.

The performance model presented in this paper predicts the interaction throughput of multimodal, recognition-based interaction. Throughput includes the time necessary for error correction. We apply the model to predict what correction accuracies are necessary to achieve certain correction speeds. For example, we predict what accuracy is necessary to beat speed of correction by typing for people with poor, average, and good typing skills. Using this model, we extrapolate the results of our user study to the performance of commercially available dictation recognizers. The model is validated using data from the user study. In conclusion, we discuss implications of this work for document creation using speech recognition, error correction in speech user interfaces, and future multimodal applications.

## EMPIRICAL EVALUATION OF A MULTIMODAL DICTATION SYSTEM

This section describes our empirical evaluation of interactive multimodal error correction in the context of a prototype multimodal dictation system. The user study pursued three research goals: providing empirical evidence for the effectiveness of multimodal correction, comparing multimodal with current interactive correction methods, and investigating user preferences between modalities.

### Experiment Conditions and Method

We used a prototype multimodal dictation system that enhances JANUS, a state-of-the-art large vocabulary speech recognizer [14], with multimodal interactive error correction. For handwriting and spelling recognition, we used specialized recognizers: Npen++ [9] and Nspell [6]. All recognizers were constrained to the same 20,000 word vocabulary. For gesture recognition, we enhanced a template matcher [15] with heuristics. More details are described in [17].

Participants read aloud either one or more sentences, which were chosen from newspaper text. After reading a sentence, the recognition result was displayed on the screen. Then, participants visually located recognition errors, selected them by tapping on the screen, and corrected them using one of the available correction methods. Which methods are available depended on the experimental condition. Participants were instructed to correct all errors, as quickly as possible (within the constraints of the condition).

Experimental conditions compared multimodal correction to conventional correction with and without keyboard input ("Keyboard&List" and "Respeak&List" condition, respectively). We considered correction by keyboard and mouse input separately in the "Keyboard&List" condition, differentiating participants based on their typing skill. Table 1 shows the experimental conditions by indicating for each experiment condition which of the different correction methods (shown in the first column) are available. We decided on a within-subject, repeated measures design to minimize the impact of the known high variation of recognition accuracy across users.

| Experimental Condition | Respeak & List | Multi-modal | Keyboard & List |
|---|---|---|---|
| Choose from List of Alternatives | X | X | X |
| Respeaking | X | X | |
| Spelling | | X | |
| Handwriting | | X | |
| Pen Gestures | X | X | |
| Keyboard / Mouse | | | X |

**Table 1: Experimental conditions, with the available correction methods shown as rows.**

Fifteen participants were recruited from the local campus community, five in each of the categories of typing skill (slow, average, and fast typist). Participants included students and administrative staff, they were balanced in gender, and most participants did not have any prior experience with speech-recognition software.

Before the study, participants learned to use the different correction modalities in a 45-60 minute long tutorial and practice session. After this session, all participants showed sufficient familiarity with the different correction methods on trial tasks. The participants then proceeded to the three experimental sessions, one for each condition. Three different sets of sentences were randomly assigned to the experimental sessions using a 3x3 Latin Square, to avoid order effects. After completing the experimental sessions, participants filled out a post-experimental questionnaire. The participants were asked which modality they perceived to be most efficient, and which modality they would prefer if all modalities had equal accuracy.

### Data Collection and Evaluation Measures

During experimental sessions, we collected data in two ways. First, the prototype multimodal dictation system created a time-stamped record of all spoken, written, and typed user interaction. This record was later manually annotated with the correct system response for each interaction, to assess recognition accuracies. For analysis of modality choice patterns, the record also contains for each recognition error the sequence of modalities used, until successful correction. All sessions were videotaped – the second method of data collection.

We measured performance at the level of a single input modality using the following three measures: input rate (i.e., how many words can a user enter per minute), system response time (i.e., how much time does automatic recognition require), and recognition accuracies (i.e., the probability of recognizing a word correctly). To distinguish between initial input and input occuring during correction of recognition errors, we use the term *correction accuracy* whenever we discuss recognition accuracy on correction input. In other words, the correction accuracy is the probability of success for the current correction attempt.

To assess performance at the task-level, we defined two measures. *Correction speed* is the average number of errors that can be successfully corrected per minute, including multiple correction attempts when necessary. For example, a correction speed of 6 cpm (corrections per minute) means that a user spends on the average 10 seconds to correct each misrecognized word. *System throughput* is the average number of words that can be sucessfully entered per minute, including the time necessary for the correction of recognition errors.

### Results

*Research Question 1: Why is multimodal correction more effective than unimodal correction?*

*Answer:* To confirm the hypothesis that unimodal correction (by repeating input in the same modality) is less accurate than multimodal correction, we calculated the correction accuracy for consecutive correction attempts *in the same modality*. Note that the counter for the correction attempt is reset after each switch of modality, even if the same recognition error is being corrected. For example, if a user corrected a recognition error in three attempts, using speech for the first two attempts, and handwriting for the final attempt, this final attempt is assigned to attempt category "1", being the first attempt after a switch of modality.

Figure 1 shows correction accuracies assuming that the original input was in speech. If users repeat input in speech, correction accuracy is much lower than if users switch to a different modality (40% for speech, 75% and 80% for handwriting and spelling). If multiple correction attempts are necessary, correction accuracy in successive attempts remains high if the user switches modality after each attempt. In terms of Figure 1, this means staying within category "1". An analysis of variance confirms that corrections in the same modality are significantly more difficult to recognize (F(2,6)=36.2, p<0.01).

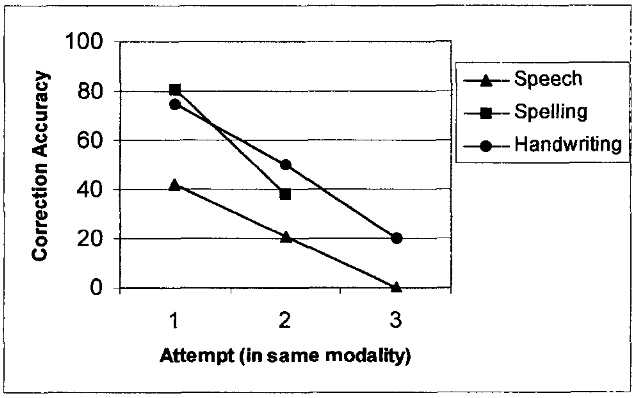

**Figure 1: Deterioration of accuracy of repeated correction attempts in the same modality**

*Research Question 2: How does multimodal correction compare with current interactive correction methods?*

*Answer:* Table 2 shows the correction speed in corrections per minute (in short *cpm*) for conventional keyboard-free correction ("Respeak&List"), correction by keyboard and mouse ("Keyboard&List"), and multimodal correction. Multimodal correction is faster than conventional keyboard-less correction by respeaking and choosing from alternatives (confirmed by post-hoc comparisons). The comparison to correction by keyboard and mouse input depends on the user's typing skill. We measured a range of correction speeds in the "Keyboard&List" condition, which corresponds to different typing skills. (The average typing rates of our participants on plain text were 23, 35, and 40 wpm for the slow, average, and fast typists, respectively). While multimodal correction is about as fast as correction by typing for users with average typing skills, it is slower for users with good typing skills.

Table 2 also shows a range of speeds for multimodal correction, which corresponds to different variations of multimodal correction. For experienced users, we measured 6.8 cpm, which is almost as fast as correction by typing for users with good typing skills. Speed of multimodal correction increases with experience because users learn to avoid ineffective correction methods.

| Correction Method | Correction Speed [cpm] |
|---|---|
| Respeak & List | 2.3 |
| Keyboard & List | 5.9 – 7.3 |
| Multimodal | 4.5 – 6.8 |

**Table 2: Speed of conventional and multimodal correction**

*Research Question 3: Which modality do users prefer?*

*Answer:* We analyzed the development of user preferences in the course of the experiment by estimating modality usage frequencies every forty correction interactions (which corresponds to one time unit in Figure 2 below) and by determining the correlation between usage frequency and correction accuracy. A positive correlation indicates that users prefer more accurate modalities.

Figure 2 shows how modality choice changed in the course of the experiment for one typical user. Corrections by handwriting were the most accurate modality for this user. This user clearly learns to prefer handwriting over the less accurate modalities of speech and choice from a list. In general, across all users, the correlation between usage frequency and correction accuracy becomes significantly more positive with experience ($F(2,4)=7.25$, $p<0.05$), i.e., users learn to prefer more accurate modalities.

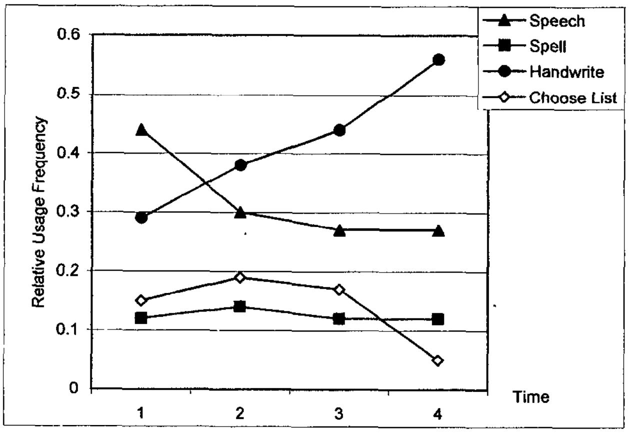

**Figure 2: Usage frequencies of different modalities for a typical user. The time axis represents the duration of the experiment (~1 hour)**

The initial bias towards speech is consistent with data from the post-experimental questionnaire, in which participants indicated that they would prefer speech if it had the same accuracy as other modalities.

## PERFORMANCE MODEL OF MULTIMODAL RECOGNITION-BASED INTERACTION

In speech-based, and more generally in multimodal interfaces, the performance of the recognition systems has a huge impact on overall interface performance. To extrapolate the results of our user study of interactive error correction to future recognition performance, we developed a simple performance model of multimodal interaction that predicts interaction throughput as a function of recognition performance. We estimated the basic model parameters based on data from our user study and applied the model to multimodal error correction.

### The Performance Model

Our performance model of recognition-based multimodal human-computer interaction predicts interaction throughput. We chose interaction throughput as the performance variable because a rational user can be expected to prefer methods that minimize effort, and time is the most important factor determining user effort. Since our definition of throughput includes the time necessary to correct any recognition errors, this measure combines time factors and recognition accuracy into a single performance measure. The following paragraphs describe the model in the context of error correction in dictation applications to derive predictions for the correction speed; but it can be generalized to other situations as well.

The model uses four basic parameters: recognition accuracy, input rate, recognition speed, and overhead time. The recognition accuracy $WA(m)$ is defined as the probability of recognizing a word (or more generally, an input item) correctly using modality $m$. (In the context of error correction, we use the term correction accuracy $CA(m)$.) The input time $T_{input}(m)$ is the average time to input a word in modality $m$ and is measured in seconds per word. We denote its inverse, the input rate (or speed), as $V_{input}(m)$ (e.g., speaking and handwriting rate). The speed of recognition is captured in the real-time factor $R(m)$. It indicates how many times longer than real-time automatic recognition in modality $m$ takes. For example, R=1 means recognition finishes at the same time as user input, without any delay. Lastly, all other times necessary to complete an interaction in modality $m$ are summarized in the overhead time $T_{Overhead}(m)$, which is measured in seconds per correction attempt. The overhead includes the time to plan or select an appropriate interaction method and the time to initiate an interaction, such as moving the hand to the screen to write or gesture on it. Hence, the overhead time depends both on modality and interface implementation.

We model a recognition-based multimodal interaction by the following steps: the user plans the interaction, chooses a modality, provides the necessary input, waits for the system to interpret the input, and finally decides whether correction is necessary.

How much time does such a multimodal interaction require? The steps of planning, choosing the modality, and the preparation of the actual input correspond to the overhead time. Then, user input in modality *m* and its automatic interpretation takes *R(m)* times $T_{input}(m)$ seconds. We therefore model a single interaction with the following simple linear additive relationship:

$$T_{Attempt}(m)=T_{Overhead}(m)+R(m)T_{Input}(m)$$

**Equation 1: Basic Decomposition of Time per Interaction into Overhead, Input, and System Response Time**

Based on this estimate for the time for one correction attempt, the correction speed is the quotient of 60 seconds and the total time to correct an error. Since error correction attempts occur sequentially, the average total time is the product of the number of attempts and the time per attempt. Denoting the average number of corrections attempts until success in modality *m* as *N(m)*, the correction speed can therefore be estimated as:

$$V_{Correct}(m)=\frac{60\text{sec}}{N(m)T_{Attempt}(m)}$$

**Equation 2: Factorization of Correction Speed into Time per Interaction and Interaction Attempts**

Assuming a constant recognition accuracy across repeated correction attempts (a simplifying assumption, as Figure 1 showed), the average number of interaction attempts until success can be developed into a geometric series, and the expected average number of correction attempts can be calculated as *N(m)=1/CA(m).*

To apply the model, some of its parameters are replaced by standard estimates, while other parameters correspond to the independent variables of the problem under question. For example, to predict the correction speed as a function of correction accuracy, we replace input rates by standard estimates, set the overhead times and real-time factors to certain values, and use correction accuracy as independent variable.

| | Speech | Spelling | Handwriting |
|---|---|---|---|
| **Input Rate $V_{input}$ [wpm]** | 47 (5) | 26 (6) | 18 (4) |
| **CA [%]** | 36 (23) | 80 (17) | 86 (6) |
| **Realtime Factor R** | 2.6 | 1.5 | 1.3 |
| **$T_{Overhead}$ [sec./correction]** | 5.4 (2.1) | 4.3 (0.7) | 3.5 (1.1) |

**Table 3: Model parameters for multimodal error correction. The widths of 95% confidence intervals are shown in parentheses.**

How can the model parameters be estimated? Recognition accuracy and speed are standard performance parameters for any recognition system and easily measured. Modality input rates have to be measured once; for standard input modalities (such as handwriting or typing), they can be found in the literature. Finally, overhead times depend on interface implementation and modality.

We divided the data from our fifteen participants into a training set (to estimate model parameters, cf. Table 3) consisting of nine participants and a test set consisting of six participants, two in each category of typing skill. Table 3 shows estimates measured on the training set for input rate (in words per minute), correction accuracies *CA* (in %), realtime factors, and overhead times (in seconds per correction). Some of these estimates will be used for predictions in following subsections.

Some readers may know that current spelling and handwriting recognizers report accuracies of 90% and more on standard benchmark tasks, raising the question why we measured much lower speeds for multimodal correction in our study? The performance of current recognizers is lower on correction input than on standard benchmarks because corrections are more difficult to recognize. More details can be found in [17].

### Application to Interactive Multimodal Error Correction

This section applies our model to the following three questions about interactive error correction in a multimodal dictation system:

1) How does correction speed depend on recognition accuracy and modalities, and how does this affect the speed of multimodal versus unimodal correction?
2) What recognition accuracy is necessary to beat typing in correction speed?
3) What is the total system throughput of a multimodal dictation system as a function of dictation accuracy and error correction?

#### *Correction Speed with Imperfect Recognition*

Correction speed depends on the performance of available recognizers and on the modality. To predict correction speed as a function of recognition performance and modality *m*, $T_{Attempt}(m)$ in Equation 2 is replaced by Equation 1, and we used estimates for input rates as shown in Table 3. We then assumed recognition in real-time for all modalities (R=1) in anticipation of faster computers. Finally, to normalize for implementation specific differences in the overhead time across modalities (as shown in Table 3), we set $T_{Overhead}=3.0$ seconds for all modalities, which is more optimistic than the measured values.

Figure 3 shows that at best, with 100% recognition accuracy, correction by respeaking achieves 24 corrections per minute (cpm), and correction by handwriting 15 cpm. This compares favorably to correction by typing for users with good typing skills (>12 cpm).

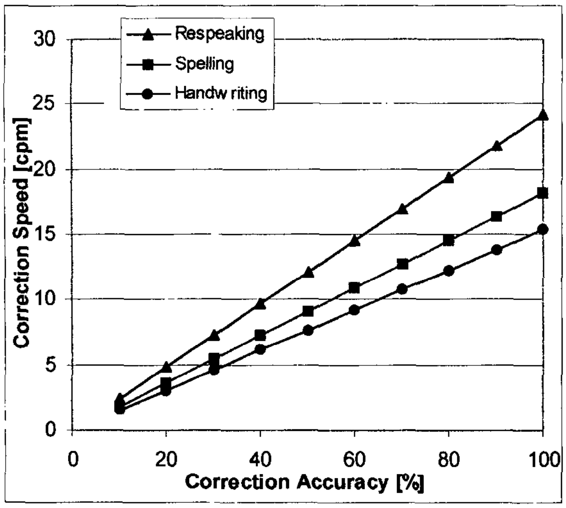

**Figure 3: Predicted Correction Speed for Repeating in Continuous Speech, Spelling, and Handwriting**

Furthermore, we can use Figure 3 to predict under what conditions unimodal correction by speech could be as efficient as multimodal correction. Since speech is the fastest modality for text input, speech would also be the most effective correction modality in a dictation system, if recognition was accurate enough. For example, multimodal corrections by spelling are 80% accurate with current recognizers (cf. Table 3). Figure 3 predicts that corrections by speech would be faster if they were more than 60% accurate, across repeated correction attempts. While our recognizer achieved only 36% accuracy on speech corrections (cf. Table 3), adapting the speech recognizer on correction input can significantly increase correction accuracy [16].

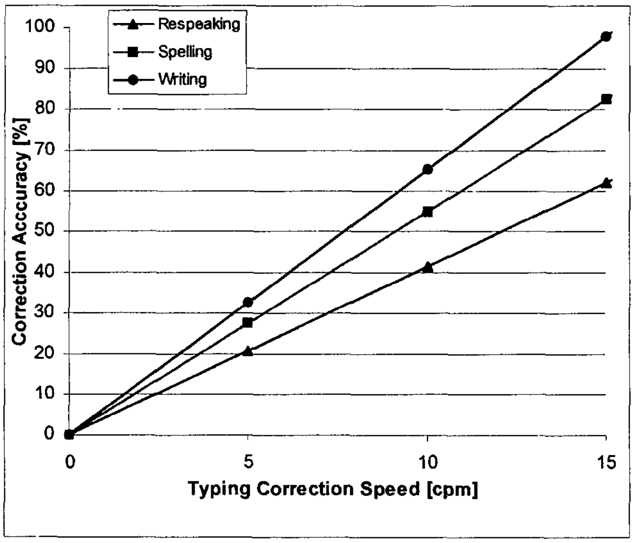

**Figure 4: Repair accuracy to beat typing in correction speed**

### *Comparing Multimodal with Typing Correction*

To compare multimodal correction and correction by typing, we answer the following question: Which correction accuracy is necessary to beat typing in correction speed, across different typing skills?

This question can be easily answered using our model by comparing the speed of multimodal correction as a function of correction accuracy with the speed of correction by typing, as shown in Figure 4.

For example, fast non-secretarial typists can correct up to 15 errors per minute using keyboard and choice from the N-best list (as measured in our study). To reach this correction speed, accuracy for corrections by repeating in continuous speech would have to be recognized at more than 65% accuracy. Corrections by spelling would have to be 85% accurate, and corrections by handwriting almost 100% accurate. Hence multimodal correction would beat correction by typing even for users with good typing skills if correction accuracy could be further improved.

### *Throughput of Dictation Systems*

Moving beyond the issue of error correction, this section discusses implications on the overall text production process. To assess the potential productivity gain of multimodal input methods, we first apply the model to predict system throughput as a function of dictation accuracy and error correction, and then compare the system throughput of three text production methods: a *multimodal dictation system* (i.e., first dictate text, then correct multimodally without any keyboard input), a *conventional dictation system* (i.e., first dictate, then correct using keyboard and choosing from alternatives) and a *standard text editor* (i.e., type the whole text). Note that our usage of the term throughput is different from some commercial vendors of dictation systems who exclude the time necessary for correction.

Our performance model can be applied to predict the throughput of dictation systems as follows. Text production with a dictation system consists of three steps: dictation, automatic interpretation of spoken input, and correction of recognition errors. How much time do these steps require? A user with speaking rate $V_{Input}(dictate)$ (in wpm) dictates $wordN = V_{Input}(dictate)$ * *1 minute* words in one minute. Then, the speech recognizer needs $T_1=R(m)*1min$ to interpret the dictation input. During automatic interpretation of the dictation input at accuracy $WA(dictate)$, on the average $errorN=wordN*(1-WA(dictate))$ recognition errors occur. The correction of these recognition errors using correction method *m* requires $T_2=errorN*T_{Correct}$ *(m)* seconds, where $T_{Correct}$ *(m)* is the inverse of the correction speed $V_{Correct}$ *(m)* (as derived in Equation 2). The total time to input *wordN* words including correction time is thus $T= T_1 + T_2$, leading to a simple formula for the throughput as function of correction method and dictation accuracy.

Figure 5 shows the system throughput for different text production methods. We extrapolate results from our user study, which were achieved with a dictation accuracy of 75%, to current commercial dictation recognizers that achieve 90% accuracy in real-time. Commercial recognizers achieve higher accuracies by adapting the speech recognizer to the user's voice. We decided not to adapt our recognizer to each participant to keep the length of experimental sessions within acceptable limits.

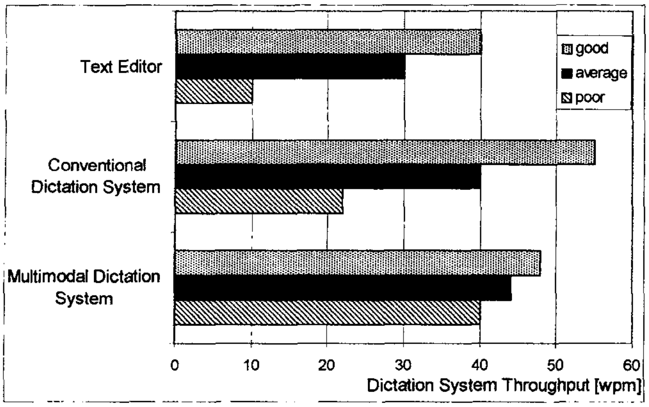

**Figure 5: Predicted throughput for different text production methods, across typing skills, for 90% dictation accuracy.**

Since typing speed obviously has a large impact on this comparison, the results are tabulated across different typing skills. For the multimodal dictation system, "poor" refers to novice users and "good" refers to experienced users. Since the experiment did not cover very slow typists, results for the slow category are based upon predictions from the performance model. As can be seen, a multimodal dictation system compares favorably to fast (non-secretarial) typing of 40 wpm - without requiring any keyboard input. For users with good typing skills, a conventional dictation system is still the most efficient text production method.

### Performance Model Validation

We validated our performance model by comparing model predictions with results of our empirical evaluation. As measure of the goodness of fit for our model, we use the average absolute error of model predictions, as suggested elsewhere [7].

| Correction Method | $V_{Correct}$ measured | $V_{Correct}$ predicted | Signed Error |
|---|---|---|---|
| Multimodal | 4.5 | 3.7 | -18% |
| Keyboard & List ("slow" typists) | 5.9 | 6.2 | 5% |
| Keyboard & List (average typists) | 6.2 | 7.0 | 13% |
| Keyboard & List ("fast" typists) | 7.3 | 7.2 | -1% |

**Table 4: Validation correction speed predictions**

Table 4 compares the correction speed predictions with the measured values, averaged across the appropriate subsets of the test set. The average absolute error is 17% for multimodal correction (N=12) and 12% for correction using keyboard and list (N=6, two test participants in each of the three categories of typing skill). These absolute errors are within reasonable range for such empirical models – despite the simplifying assumptions of the model as presented here. Predictions of dictation system throughput (input speed including error correction) match empirical data equally well (cf. [17]).

## 4. DISCUSSION

We first raise several concerns about the external validity of our study and show how model predictions alleviate them. We then discuss implications of this research on dictation and other (multimodal) speech recognition applications.

### *Validity of Results*

As key result of the user study we found that accuracy decreases in repeated correction attempts unless modality is switched. This observation appears to generalize across modalities, and across state-of-the-art recognition systems [17]. The magnitude of this effect, and thus whether multimodal correction is faster than unimodal correction, depends on the recognition system used. For current recognition systems, our study showed that there is a gain in using multimodal correction. But if accuracy was significantly improved (by using different recognition algorithms on correction input, cf. [16]), unimodal correction by respeaking could outperform multimodal correction. Model predictions help to decide whether multimodal correction is beneficial.

A second external validity concern is the influence of implementation details. We argue that they do not change our main results. Furthermore, effects of implementation modifications can be estimated using our performance model. For example, halving the overhead time (i.e., the time spent on locating errors in recognition output and on starting a correction) for each modality would increase the speed of multimodal correction to the level of fast unskilled typing – without any further improvement in recognition accuracy! Overhead time could be significantly reduced, for instance, by automatically highlighting recognition errors.

Finally, our study did not control for how users switch between dictation and error correction in the overall process of text production. While this may have a significant impact on the absolute text production speeds, it does not affect the comparison of different correction methods – the focus of our study.

### *Implications for Dictation Systems*

Related work suggested that automatic speech recognition technology could significantly increase productivity on dictation tasks [4, 5]. However, formal evaluations of dictation systems reported either only small productivity increases [1], or lack of user acceptance despite significant productivity increases [8].

Our results suggest that not only high recognition accuracy, but also adequate error correction is crucial to ensure high text production speeds in dictation systems. Furthermore, the productivity gain of dictation systems may be smaller than widely assumed. First, most potential users of dictation systems have good typing skills, and our results showed that for skilled typists, the productivity gain of dictation systems is rather modest. Second, studies suggest that for creation of documents, not input speed, but the skill required to compose text is the main limiting factor [4].

### *Implications for other Speech Recognition Applications*

Our study explored the trade-off between speed and accuracy of different modalities only for text input. The most efficient input modality depends not only on input speed and accuracy, but also on the task. For example, for entry of numerical data, handwriting digits is about as fast as speech. We believe that the flexibility to change modality depending on the task is a great advantage of future multimodal input technologies.

Furthermore, applications other than dictation may limit which alternative modalities are available. However, error correction even benefits from just one alternative modality. If speech is the only modality available (e.g., in telephone applications), the speech user interface designer should consider switching between different speech modalities, such as continuous, discrete, and spelled speech.

## CONCLUSIONS

This paper provides useful insights for designers of speech (and multimodal) user interfaces. Our study showed that multimodal correction is faster than conventional correction without keyboard input. Furthermore, we showed that recognition accuracy has a significant influence on user choice between modalities: with practice, users learn to avoid ineffective modalities in favor of more effective modalities. Our research suggests that multimodal input methods are particularly attractive for applications that do not allow fast keyboard input (e.g., small mobile devices), and for users with poor typing skills.

The performance model of multimodal human-computer interaction presented in this paper is a first step towards formalizing multimodal interaction. We showed how predictions from such a model help answer important design decisions in speech user interfaces, effectively complementing results from empirical evaluations. Future work may generalize the model to provide a general framework for multimodal interaction.

## ACKNOWLEDGMENTS

This research was sponsored by the DARPA under the Department of Navy, Office of Naval Research under grant number N00014-93-1-0806. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

Thanks to my colleagues of the Interactive Systems Laboratories at Universität Karlsruhe and Carnegie Mellon University who developed the continuous speech, spelling and cursive handwriting recognizers used in our prototype. Sincere thanks also to David Novick, Chloe Meadows, and to all participants of the user studies.

## REFERENCES:

1. Alto, P., *et al.* "Experimenting Natual-Language Dictation with a 20000-Word Speech Recognizer," in *VLSI and Computer Peripherals.* 1989. IEEE Computer Society Press. **2**: pp. 78-81.

2. Baber, C., Stammers, R.B., and Usher, D.M., "Error correction requirements in automatic speech recognition," in *Contemporary Ergonomics*, E.J. Levesey, Editor 1990, Taylor and Francis. London.

3. Gibbon, D., Moore, R., and Winski, R., eds. *Handbook of Standards and Resources for Spoken Language Systems.* 1997, Mouton de Gruyter: Berlin, New York.

4. Gould, J.D., "How Experts Dictate." *Journal of Experimental Psychology: Human Perception and Performance*, 1978. **4**(4): pp. 648-661.

5. Gould, J.D., Conti, J., and Hovanyecz, T., "Composing Letters with a Simulated Listening Typewriter." *Communications of the ACM*, 1983. **26**(4): pp. 295-308.

6. Hild, H., *Buchstabiererkennung mit neuronalen Netzen in Auskunftssystemen.* Fakultät für Informatik Fredericiana, 1997, Karlsruhe. 216 pages.

7. Kieras, D.E., Wood, S., D., and Meyer, D.E., "Predictive Engineering Models Based on the EPIC Architecture for a Multimodal High-Performance Human-Computer Interaction Task." *ACM Transactions on Computer-Human Interaction*, 1997. **4**(3): pp. 230-275.

8. Lai, J. and Vergo, J. "MedSpeak: Report Creation with Continuous Speech Recognition," in *International Conference on Computer-Human Interaction CHI.* 1997. Atlanta (USA). **1**: pp. 431-438.

9. Manke, S., Finke, M., and Waibel, A. "NPen++: A Writer Independent, Large Vocabulary On-Line Cursive Handwriting Recognition System," in *International Conference on Document Analysis and Recognition.* 1995. Montreal.

10. McNair, A.E. and Waibel, A. "Improving Recognizer Acceptance through Robust, Natural Speech Repair," in *International Conference on Spoken Language Processing.* 1994. Yokohama (Japan). **3**: pp. 1299-1302.

11. Mellor, B. and Baber, C. "Modelling of Speech-based User Interfaces," in *European Conference on Speech Communication and Technology.* 1997. Rhodes (Greece): ESCA. **4:** pp. 2263-2266.

12. Oviatt, S. and VanGent, R. "Error Resolution During Multimodal Human-Computer Interaction," in *International Conference on Spoken Language Processing.* 1996. Philadelphia (PA). **2**: pp. 204-207.

13. Rhyne, J.R. and Wolf, C.G., "Recognition-Based User Interfaces," in *Advances in Human-Computer Interaction*, H.R. Hartson and D. Hix, Editors. 1993, Ablex Publishing. Norwood (NJ). pp. 191-212.

14. Rogina, I. and Waibel, A. "The JANUS Speech Recognizer," in *ARPA Workshop on Spoken Language Technology.* 1995. Austin (TX). Morgan Kaufmann. pp. 166-169.

15. Rubine, D., "Specifying Gestures by Example." *ACM Journal on Computer Graphics*, 1991. **25**(4): pp. 329-337.

16. Soltau, H., 1998. Personal Communication.

17. Suhm, B., *Multimodal Interactive Error Recovery for Non-Conversational Speech User Interfaces.* PhD, Computer Science Department, Fredericiana University, 1998, Karlsruhe.