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 two issues that will be crucial in the successful development of emerging multimodal user interfaces: how the problem of interpretation errors can be addressed, and how to evaluate such applications. We present results from a user study that compares multimodal correction methods with conventional correction methods (with and without keyboard input). The study confirms the hypothesis that multimodal correction expedites keyboard-less correction in speech user interfaces. Choice between modalities for experienced users is driven by correction accuracy; however there is initially a strong bias towards using speech. We also present a performance model of multimodal human-computer interaction that predicts input speeds including the time necessary for correction of recognition errors. We apply the model to important issues in speech user interfaces, such as what recognition accuracy is necessary to make multimodal input faster than typing, and we extrapolate results from the user study. To evaluate multimodal applications more effectively, a combination of user studies and model-based predictions is proposed.

Keywords

Multimodal interaction, evaluation methodologies, speech user interfaces, interactive error correction, quantitative performance model

1. INTRODUCTION

In contrast to traditional applications that rely on keyboard and mouse input, multimodal applications automatically interpret user input in the modalities that humans naturally employ to communicate. This paper addresses two issues in multimodal applications: interactive correction of inevitable recognition errors, and evaluation.

How to deal with system interpretation errors of user input is the main added challenge of multimodal interfaces, compared to traditional interfaces, since input has to be interpreted automatically. Previous work [1, 2] hypothesized that error correction in speech recognition applications should benefit from multimodal flexibility, and other work [3, 4] described multimodal interactive correction methods. This paper reports results of an extensive evaluation of interactive error correction in a dictation application. Related work suggested that automatic speech recognition technology could significantly increase productivity on dictation tasks [5, 6], but formal evaluations of dictation systems reported either only small productivity increases [7], or lack of user acceptance despite significant productivity increases [8]. Our results suggest that adequate error correction methods, in addition to recognition accuracy, is crucial to achieve high throughput in dictation systems. Within this paper, our study serves three main purposes. First, it provides empirical evidence for the hypothesis that multimodal correction expedites keyboard-less error correction in speech user interfaces. Second, data from the study allowed us to validate our performance model of multimodal human-computer interaction, which will be described in the next section. Finally, the study motivates the methodological contribution of this paper as described below.

Within the HCI field, three methodologies for evaluating interactive systems have been established [9, 10]: empirical techniques (or user-based approach), predictive models (or theory-based approach), and expert evaluation. Evaluation of multimodal applications poses a number of new challenges. First, the empirical methodology of benchmark tests on canned data, established in the fields that developed the necessary component technologies (e.g., continuous speech handwriting recognition), is not applicable. Technology-oriented evaluation measures (such as recognition accuracy) must be complemented by measures that capture qualitative issues such as productivity increase, perceived usefulness, and user satisfaction. Such issues are addressed in empirical user studies. But user studies are costly, and the results depend on interface implementation and performance of available recognizers. By using the WoZ methodology [11], arbitrary recognition accuracies can be simulated; but it is difficult to realistically simulate error behavior of automatic recognition systems.

We propose to complement empirical user studies with predictions from quantitative models for more effective evaluation of emerging multimodal interfaces. To this end, we present a performance model of multimodal recognition-based human-computer interaction that predicts input speeds based on a few basic parameters. To demonstrate the value of a combined evaluation, we show how model-based predictions complement user studies by abstracting from implementation details and performance of current recognizers.

The remainder of this paper is organized as follows. Section 2 presents our study that compares multimodal with conventional correction methods on a dictation task. It is shown that unimodal corrections are less effective than multimodal correction, how multimodal correction compares to current keyboard-less correction and typing in correction speed, and what factors drive user choices between modalities. Section 3 presents our predictive performance model of multimodal human-computer interaction that predicts task completion time based on a few basic parameters, such as recognition accuracy and speed. The performance model is applied to multimodal interactive error correction, and validated using data from the user study. The paper closes with a discussion and conclusions.

2. EMPIRICAL EVALUATION OF A MULTIMODAL TEXT EDITOR

This section describes a user study that evaluates interactive multimodal error correction in the context of a prototypical multimodal text editor. Its main goal is to provide empirical evidence for the hypothesis that error correction in speech user interfaces benefits from multimodal flexibility. First, the study provides shows the ineffectiveness of unimodal correction and the effectiveness of multimodal correction. Second, the study provides insights useful for the designer of speech user interfaces by formally comparing conventional keyboard-less and keyboard-oriented correction with multimodal correction methods. Finally, results suggest that accuracy is an important factor in determining user preferences between different input modalities, that there is a bias towards using speech, that users learn to prefer more effective modalities with practice, and that the preferred (and most effective) modality differs across users.

2.1 Multimodal Interactive Error Correction

From a survey of commercial and published research systems, we identified four correction methods employed in current speech user interfaces: correction by respeaking, typing, choosing from a list of alternatives, or clarification dialogues. Previous research [4] shows that correction by respeaking and by choosing from a list is ineffective in continuous speech applications, and correction by typing relies on keyboard input. Clarification dialogues (correction in a spoken dialogue, similar to error resolution in human-human communication) are appropriate for conversational speech applications, which our research currently does not address. The multimodal interactive approach makes keyboard-less correction in speech user interfaces efficient for text input and data entry tasks. Multimodal error correction allows the user to switch modality for correcting misrecognized input. Instead of repeating input using continuous speech, it is repeated using either spelling or handwriting. Simple editing tasks are performed using gestures that are drawn on a touchsensitive display. The gestures are similar to marks used by text editing professionals [12]. For example, items are deleted by scratching them through, and the position of the insertion cursor is changed with a carret mark.

2.2 The Research Questions

The goals of providing empirical evidence for the ineffectiveness of unimodal correction, of comparing multimodal with conventional correction methods, and of determining what factors drive user preference between modalities, translate into the following specific research questions.

1. Research Question: Why is unimodal correction (by repeating input in the same modality) ineffective, and why is multimodal correction effective? Hypothesis: Recognizing corrections that are repetitions in the same modality is difficult. Recognition performance of (most) current recognizers deteriorates on corrections in the same modality, unless the recognition algorithms are modified. Switching modality significantly increases correction accuracy, compared to correction in the same modality.
2. Research Question: How does multimodal correction compare with current interactive correction methods? Hypothesis: With current recognition technology, multimodal correction is faster than unimodal correction by respeaking, but slower than correction by fast typing.
3. Research Question: Which correction modalities do users prefer, and which factor(s) drive user choice? Hypothesis: Users prefer the most effective correction modality.

2.3 Experiment Conditions and Method

Participants were instructed to read aloud sentences (chosen from the Wall Street Journal), and to correct all recognition errors that occur during automatic recognition of the read sentence. The experimental conditions differ by the methods available for error correction, thus comparing keyboard-less multimodal correction with conventional correction by respeaking, keyboard input, and choosing from a list of alternatives. The columns in Table 1 define the experimental conditions, indicating which of the different correction methods (shown as rows) are available in each experimental condition. Correction by respeaking is considered separately (in the "Respeak/Choose from List" condition) from correction by keyboard and mouse input ("Keyboard/List" condition), since the effectiveness of correction by typing varies in wide ranges depending on typing skill.

The experimental design is a within-subject, repeated measures design. Thus, the impact of the known high variation of recognition accuracy across users on comparing correction speeds can be minimized. Participants learned to use the different correction modalities in a 45 minute long tutorial and practice session that preceded the experimental sessions. As a test whether participants sufficiently learned all correction methods, the final step of this session required the participants to dictate and to correct a few sentences under experiment conditions. Fifteen participants were recruited from the local campus community, and systematically controlled for typing speed: five participants in each of the (self-reported) categories of slow, average, and fast typist. Most participants did not have any prior experience with speech-recognition software.

Correction Method	Respeak & List	Multi-modal	Keyboard & List
Choose from List of Alternatives	X	X	X
Respeaking	X	X	
Spelling		X	
Handwriting		X	
Pen Gestures		X	
Keyboard / Mouse			X

Table 1: Experimental Conditions

Evaluation measures include both quantitative and qualitative issues. Besides input rates and system response times, correction accuracy, correction speed, and dictation speed (as defined below) are the main quantitative measures. *Correction accuracy* refers to the success rate of a single correction attempt (i.e., the word accuracy on correction input). *Correction speed* is defined as the number of errors that can be successfully corrected per minute, and the *dictation speed* (or dictation system throughput) is defined as the number of words dictated to the system, including time necessary for correction of recognition errors. A post-experiment questionnaire addressed qualitative issues, including which modality users perceive to be most efficient, and which modality users prefer if all modalities had equal accuracy.

2.4 Results

Hypothesis 1: Unimodal correction is less effective than multimodal correction.

Figure 1 shows the correction accuracy across several attempts at correcting the same recognition error. As can be seen, correction accuracy dramatically deteriorates when input is repeated in the same modality, while accuracy is much higher if modality is switched for correction. An analysis of variance confirms that repeated corrections are significantly more difficult to recognize ($F(2,6)=36.2$, $p<0.01$).

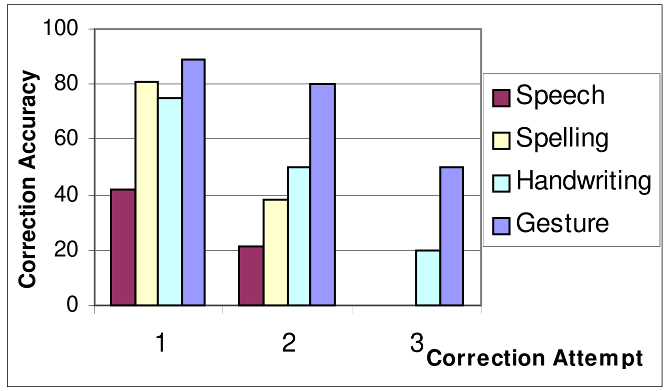

Figure 1: Deterioration of Correction Accuracy on Repeated Corrections

Hypothesis 2: Multimodal correction is faster than conventional keyboard-less correction (by respeaking and choosing from alternatives), but slower than correction by typing for users with good typing skills.

Table 2 shows the correction speed in corrections per minute [cpm] for multimodal correction, conventional keyboardless correction ("Respeak&List"), and correction by keyboard and mouse ("Keyboard&List"). The range of speeds for multimodal correction corresponds to different multimodal correction methods: multimodal correction with and without partial word correction (4.5 and 5.0 cpm), multimodal correction by spelling or handwriting and choosing from a list (5.3 and 5.2 cpm), and multimodal correction for experienced users (6.8 cpm). The range of speeds in the "Keyboard&List" condition corresponds to different typing skills (23-40wpm). Post-hoc pairwise comparisons confirm that multimodal correction is significantly faster than conventional keyboardless correction ("Respeak&List"), but slower than correction with keyboard input for fasts typists (>40 wpm).

Correction Method	Correction Speed [cpm]
Respeak & List	2.3
Multimodal	4.5 – 6.8
Keyboard & List (typing speed: 23 wpm – 40 wpm)	5.9 – 7.3

Table 2: Speed of Conventional and Multimodal Correction

Hypothesis 3: Users intuitively prefer speech, but with practice they learn to prefer the most accurate correction modality. The preferred modality differs across users.

We analyzed modality usage patterns in the course of the experiment by estimating modality usage frequencies every 40 correction interactions, and by determining the correlation between usage frequency and correction effectiveness (defined as rank with respect to correction accuracy, i.e. rank 1 for the most effective modality). Figure 2 shows how these correlations develop with increasing duration of the experiment, across different correction modalities. As can be seen, the correlation becomes more positive, i.e., with increasing experience, users learn to prefer more effective modalities. This effect is significant ($F(2,4)=7.25$, $p<0.05$). The negative correlation for the speech modality shows that speech is used despite repeated evidence for its ineffectiveness. Especially in the first correction attempt, speech is frequently used initially. The fact that most users would prefer speech if it had the same accuracy (as indicated in the post-experiment questionnaire) may explain this bias towards using speech. Finally, different users prefer different modalities, depending on which modality proves to be most effective for them.

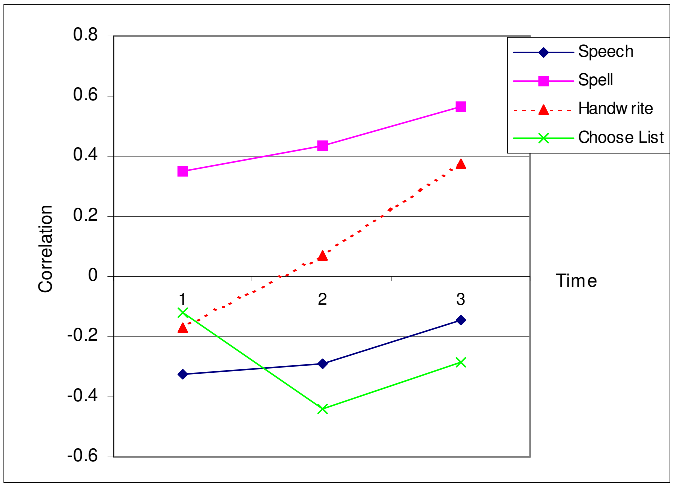

Figure 2: Correlation between modality usage frequency and correction effectiveness over time (i.e. with increasing experience). Positive correlation means effective modalities are preferred.

Our study thus suggests a refinement of hypothesis 3: while users eventually learn to prefer the most effective correction modality, they initially prefer speech.

3. PERFORMANCE MODEL OF MULTIMODAL RECOGNITION-BASED INTERFACES

With different input methods available in multimodal interfaces, predicting which method users will prefer becomes an important question for the designer of such applications. Predictive quantitative models are particularly useful in multimodal interfaces that rely on imperfect recognition, to help abstract from dependence on current recognizer performance.

Although the difficulty of evaluating recognition-based interfaces has been recognized in previous work [13], to date there has been only one attempt to develop a predictive performance model for speech-based interfaces [14]. This model predicted task completion time using critical path analysis. The model accounted for imperfect recognition by modeling recognition errors as repetition of input. The model was validated on three tasks: registration number entry, surveillance, and map interaction. A good, albeit not always statistically significant match between model predictions and empirical data was reported. The drawback of this model is that dependency of task completion time on modality, recognizer, implementation specific factors is not explicitly modeled.

3.1 Our Performance Model

Our performance model of recognition-based multimodal human-computer interaction predicts input speed. Input speed is chosen as main performance variable because a rational user can be expected to prefer methods that minimize the effort spent on interacting with the system, and time is the most important factor determining user effort. The core idea is to combine time factors with recognition accuracy into a single performance measure by measuring the time until successful completion of input, including the time necessary to correct any recognition errors. A decomposition of input completion time (including correction time) models the dependence of completion time on implementation and recognition explicitly. The following paragraphs describe the formulas of our performance model in its simplest instantiation, adequate to predict input and correction speeds in sequential recognition-based multimodal interaction, setting the stage for its application to our example application of a multimodal text editor.

The model uses four basic parameters: recognition accuracy, input rate, recognition speed, and overhead time. The recognition accuracy $WA(m)$ indicates the success rate for inputting an information item correctly using input method (i.e., modality) m. In dictation applications, the elementary information item is typically a word. The input time $T_{input}(m)$ indicates the time necessary to convey an input item to the application in m, and is the inverse of the input rate $V_{input}(m)$ (e.g., speaking and handwriting rate). The recognition speed is captured in the real-time factor $R(m)$ that measures how many times longer than real-time it takes to recognize input automatically in modality m. Finally, all other times necessary to complete an interaction in m are summarized as overhead time $T_{Overhead}(m)$. The overhead includes the time to plan or select an appropriate interaction method, and the time to initiate an interaction (e.g., moving the hand to the screen to write or gesture on it).

How can input speed in an interaction method be predicted based on these parameters? One recognition-based multimodal interaction consists of the following steps: the user plans the interaction, chooses an interaction method, provides the necessary input, waits for the system to interpret the input, and finally decides whether correction is necessary. We model such an 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 time to complete one interaction, the correction speed (in corrections per minute) is estimated as:

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

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

N(m) is an estimate of the average number of attempts necessary to successfully complete input in modality *m*. Here, the assumptions of sequential interaction and correction of recognition errors by repeating input come into play. If the recognition accuracy is assumed to be constant across multiple 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 can be calculated as *N(m)=1/WA(m).*

Before the model can be applied, its basic parameters have to be estimated. Recognition accuracy and speed are the standard performance parameters for any recognition system and can be obtained from specifications (if an off-the-shelf recognizer is used), or otherwise they can be measured. Modality input rates have to be measured once; for novel interaction methods, the input speed can be measured in pilot experiments. Finally, the overhead time strongly depends on the interface implementation. However, rough values can be estimated early in the design of a multimodal application (in pilot tests) and refined later on whenever necessary. Table 3 shows the estimates for word-level correction that were measured in our user study for corrections using continuous speech, spelling, handwriting, and typing.

	Speech	Spell-ing	Hand-writing	Typ-ing
Input Rate V_{input} [wpm]	47 (5)	26 (6)	18 (4)	17 (7)
WA [%]	36 (23)	80 (17)	86 (6)	84 (5)
Realtime Factor R	2.6	1.5	1.3	1.0
Overhead $T_{Overhead}$ [s/error]	5.4 (2.1)	4.3 (0.7)	3.5 (1.1)	2.6 (0.9)

Table 3: Model Parameters for Interactive Multimodal Error Correction (95% confidence intervals in parentheses)

3.2 Application to Interactive Multimodal Error Correction

To illustrate how to model multimodal recognition-based interaction, and what kind of useful predictions can be derived, this section applies our model to interactive correction of recognition errors in a multimodal text editor. Multimodal correction was described in Section 2.1. We apply the performance model to the following three questions:

1) Predicting the correction speed as a function of recognition accuracy in different modalities
2) Predicting the recognition accuracy necessary to beat typing in input and correction speed. Since keyboard input is one baseline to which multimodal interaction is compared in text production tasks, the comparison to keyboard input is of particular interest.
3) Predicting the total system throughput of a multimodal text editor, depending on the accuracy of the dictation recognizer and on the error correction methods used.

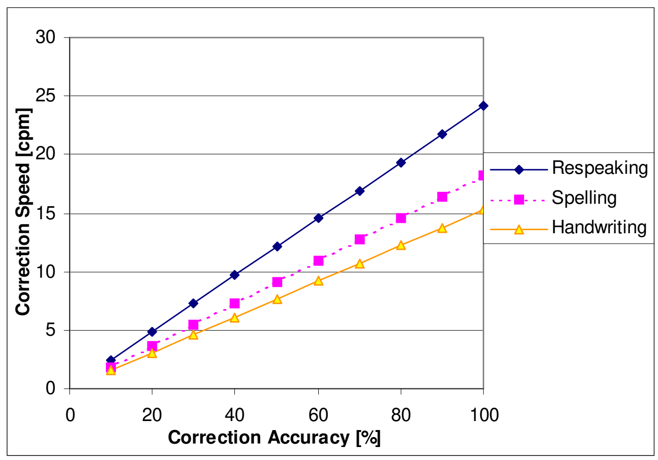

Figure 3: Predicted Correction Speed for Repeating in Continuous Speech, Spelling, and Handwriting

3.2.1 Correction Speed with Imperfect Recognition

We assume correction is performed in a sequential manner by repeating the initial input. Estimates for correction speed can be computed by substituting $T_{Attempt}(m)$ in Equation 2 with Equation 1, and by using the basic performance variables for the different correction methods under consideration, as shown in Table 3. To eliminate some obvious implementation-dependent variation in recognition speed and overhead, we assume recognition in realtime for all modalities (R=1), and an overhead time of $T_{Overhead}$ =3.0 seconds per correction attempt. Figure 3 plots the correction speed over the recognition accuracy for repeating in continuous speech, spelling, and handwriting. The model predicts that at best, with 100% recognition accuracy, correction by respeaking achieves 25 corrections per minute (cpm), and corection by handwriting 15 cpm.

3.2.2 Comparing Multimodal Correction with Typing

Using the dependency of the correction speed in different correction modalities on correction accuracy as derived in the first subsection, the question of what correction accuracy is necessary to beat a certain input speed can be answered easily. Figure 4 predicts the word accuracy necessary to beat typing in correction speed, assuming real-time recognition in all modalities. From our user studies we know that fast non-secretarial typists can correct up to 12 errors per minute using keyboard and choice from the N-best list. To reach this level of correction speed, accuracy for corrections by repeating in continuous speech would have to be recognized more than 50% accurate. Corrections by spelling would have to be 70% accurate, and corrections by handwriting 85% accurate. While current spelling and handwriting recognizers already surpass such accuracy levels on standard benchmark tasks, recognizing correction input is more difficult than standard benchmarks, and the assumption of constant correction accuracy across multiple correction attempts does not hold (see Figure 1). In our user study we therefore measured lower speeds (than 12 cpm) for multimodal correction.

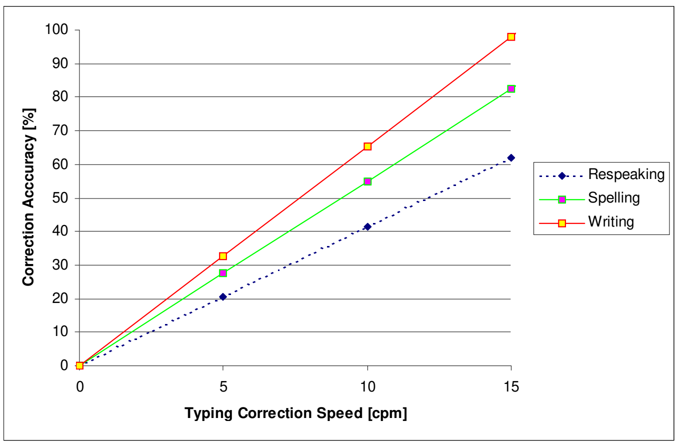

Figure 4: Repair Accuracies to beat Typing in Correction Speed Dictation Speed including Error Correction

3.2.3 Throughput of Dictation Systems

The most important time-related measure in evaluating the productivity of a dictation system is system throughput, i.e. the speed of text input including the time required to correct any recognition error(s). This is in contrast to commercial vendors of dictation systems who exclude correction time when they report dictation input speeds of up to 180 wpm. A simple formula for this total dictation speed can be derived as follows.

A user with speaking rate $V_{Input}(dictate)$ [wpm] can dictate $wordN= V_{Input}(dictate) * 1\ minute$ words in one minute. The speech dictation recognizer will take $T_1=R(m)*1min$ to interpret this input. The average number of errors when recognizing these words at accuracy $WA(dictate)$ is $errorN=wordN*(1-WA(dictate))$. The correction of these recognition errors using correction method m requires $T_2=errorN*T_{Correct}\ (m)$, with the inverse $T_{Correct}\ (m)$ of the correction speed in modality m $V_{Correct}\ (m)$. The overall time to complete an input interaction, including correction time, is $T= T_1 + T_2$. Putting the pieces together yields a formula that predicts the throughput of a dictation system based on dictation accuracy and speed, and speed of error corrections.

To assess the potential productivity gain of multimodal input methods for text input tasks, we compare the system throughput of a multimodal text editor (i.e., first dictate text, then correct multimodally) with a conventional listening typewriter (i.e., first dictate, then correct using keyboard and choice from list) and with a standard text editor (i.e., type the whole text). To this end, we extrapolate results from our user study to a dictation recognizer that achieves 90% accuracy in real-time using the formula derived in the previous paragraph. This performance is comparable to what current commercial dictation systems achieve. We measured lower dictation accuracies of ~75% in our study since we didn't adapt the speech recognizer to each participants voice (a method that significantly increases dictation accuracy), to keep the length of experimental sessions within acceptable limits.

Since typing speed obviously has a large impact on this comparison, the results are tabulated across different typing skills. Since the experiment did not cover very slow typists, results for the slow category are based upon predictions from the performance model. Table 4 compares the predicted system throughput of a multimodal text editor (first dictate, then correct multimodally without keyboard input) with a conventional listening typewriter (first dictate, then correct using keyboard and choosing from alternatives), and a text editor (type the whole text). As can be seen, multimodal text editor achieves text input rates that compare favorably to fast (non-secretarial) typing of 40 wpm - without requiring any typing. Thus, users with poor typing skills can be as efficient as good typists if they used a multimodal text editor.

Text Production Method	**Predicted System Throughput [wpm]**
Multimodal Text Editor	40-48
Conventional Listening Typewriter, slow typist	22-34
Conventional Listening Typewriter, average typist	40-48
Conventional Listening Typewriter, fast typist	>52
Text Editor, slow typist	5-15
Text Editor, average typist	23-35
Text Editor, fast typist	>40

Table 4: Productivity for text production methods, across different typing skills

3.3 Performance Model Validation

We validated our performance model using data from the empirical evaluation of the multimodal text editor. To this end, the data from fifteen participants was divided into a training set consisting of nine participants that was used to estimate model parameters (shown in Table 1), and a test set consiting of six participants that was used to obtain model predictions on unseen data. We validated the performance model on two levels: first, predictions of correction speeds (as expressed in Equation 2 and 1), and second, the prediction of dictation system throughputs. As measure of the goodness of fit for our model, we use the average absolute error of model predictions, as suggested in [15].

To validate predictions of the correction speed based on our performance model, Table 5 compares the correction speed predictions with the measured values (averaged across the appropriate subsets of the six participants in 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), and thus within reasonable range for such empirical models.

Correction Method	$V_{Correct}$ measured	$V_{Correct}$ predicted	(Signed) Error
Multimodal	4.5	3.7	-18%
Keyboard & List (23 wpm typing)	5.9	6.2	5%
Keyboard & List (35 wpm typing)	6.2	7.0	13%
Keyboard & List (40 wpm typing)	7.3	7.2	-1%

Table 5: Validation Correction Speed Predictions

Correction Method	Measured Throughput	Predicted Throughput	(signed) Error
Multimodal	16 wpm	14.8 wpm	-1%
Keyboard & List (23 wpm typing)	16.3 wpm	16.7 wpm	8%
Keyboard & List (35 wpm typing)	18.4 wpm	18.1 wpm	-2%
Keyboard & List (40 wpm typing)	25.0 wpm	18.4 wpm	-26%

Table 6: Validation of dictation system throughput predictions

To validate the formula that predicts the dictation system throughput (see Section 3.2.3), Table 6 compares the system throughput as measured in the user study with model predictions. We used the following values that were measured in the user study: dictation accuracy *WA(dictate)=75%,* dictation speed V_{Input} *(dictate)=133 wpm,* and the correction speeds of different methods (see Table 2 in Section 2.4). As can be seen, there is a good match; the average absolute error is 18% across all categories (N=18). The high deviation for correction by typing in the fast category is probably due to an outlier in that category (only two participants in that subset of test set).

In summary, the performance model of recognition-based multimodal human-computer interaction can be applied to a number of interesting issues in multimodal error correction, and model predictions fit well with the empirical data.

4. DISCUSSION

Error Correction in Speech User Interfaces

The multimodal approach is most applicable to non-conversational applications with a graphical user interface that support data entry tasks. If speech is the only modality available (e.g., in telephony applications), multimodal flexibility can be exploited by switching between different speech modalities, such as continuous, discrete, and spelled speech.

Second, our study considered text input (dictation tasks) only. On other tasks, the trade-off between modalities is different. For example in entry of numerical data, handwriting digits is about as fast as speech (unlike in text input tasks). Multimodal input technologies offer the flexibility to switch the modality depending on the task.

Our study also provided insight into what factors determine user choice when multiple modalities are available. With practice, users learn to avoid ineffective modalities and prefer effective modalities. However, we observed a bias towards using speech, especially in the first correction attempt, despite repeated evidence for its ineffectiveness as correction modality. The bias towards speech was confirmed by the post-experimental questionnaire.

Evaluation Methodology for Multimodal Applications

Our performance model is currently formalized for sequential multimodal interaction, but it could be extended to model multimodal applications with simultaneous use of modalities, e.g., using critical path analysis techniques, similar to Mellor's approach [14].

Second, the predictions that were derived for correction speed and total system throughput assume correction by repeating input. This assumption is motivated by the chosen task: text input, or more generally data entry. For such tasks, it is natural and intuitive to correct by repeating. However, our model can be extended to other correction methods by modeling the initial input separately from correction input in Equation 1.

Third, categorizing our model within the taxonomy of evaluation methods [9], it is an example for a quantitative predictive model. It is however not backed by a cognitive theory, like for example the keystroke model or GOMS.

5. CONCLUSIONS

This paper provided insights for designers of speech recognition applications. We compared multimodal and conventional interactive correction methods for non-conversational speech recognition applications. The user study of a multimodal text editor provided empirical evidence for the hypothesis [1] that multimodal flexibility expedites error correction in speech user interfaces. Our analysis suggests further refinements of this hypothesis: Multimodal correction is attractive as long as either accuracy of recognizing corrections by respeaking remains low, or as long as speed of conventional input methods (e.g., by typing) is low. The speed of conventional input methods may be low due to lack of typing skill, or due to task and application constraints (e.g., text entry on small handheld devices). 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.

This paper proposed a combination of model-based evaluation with user studies as an evaluation methodology for emerging multimodal applications. A performance model of multimodal human-computer interaction was presented that abstracts from dependency on interface implementation and performance of current recognition technology. Predictions from such a model can effectively complement an empirical evaluation of multimodal applications, and they increase the external validity of empirical results by abstracting from the performance of available recognition systems. Thus, the effects of design changes and recognition performance improvements can be evaluated without having to iterate over costly user studies. Rich data from a user study can validate model predictions, and addresses qualitative issues such as usefulness and user satisfaction.

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 Navy or the U.S. Government.

Thanks to my colleagues of the Interactive Systems Laboratories at Universität Karlsruhe and Carnegie Mellon University who made this research possible by developing the continuous speech, connected letter and cursive handwriting recognizers used in our prototype. Thanks also to David Novick for useful suggestions, and to all participants of the user studies.

REFERENCES:

[1] S. Oviatt and R. VanGent, "Error Resolution During Multimodal Human-Computer Interaction," presented at International Conference on Spoken Language Processing, Philadelphia (PA), 1996.

[2] J. R. Rhyne and C. G. Wolf, "Recognition-Based User Interfaces," in *Advances in Human-Computer Interaction*, vol. 4, H. R. Hartson and D. Hix, Eds. Norwood (NJ): Ablex Publishing, 1993, pp. 191-212.

[3] B. Suhm, B. Myers, and A. Waibel, "Interactive Recovery from Speech Recognition Errors in Speech User Interfaces," presented at International Conference on Spoken Language Processing, Philadelphia (PA), 1996.

[4] B. Suhm, "Empirical Evaluation of Interactive Multimodal Error Correction," presented at IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara (CA), 1997.

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

[6] J. D. Gould, J. Conti, and T. Hovanyecz, "Composing Letters with a Simulated Listening Typewriter," *Communications of the ACM*, vol. 26, pp. 295-308, 1983.

[7] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, and S. Scarci, "Experimenting Natual-Language Dictation with a 20000-Word Speech Recognizer," presented at VLSI and Computer Peripherals, 1989.

[8] J. Lai and J. Vergo, "MedSpeak: Report Creation with Continuous Speech Recognition," presented at International Conference on Computer-Human Interaction CHI, Atlanta (USA), 1997.

[9] M. Sweeney, M. Maguire, and B. Shackel, "Evaluating user-computer interaction: a framework," *International Journal of Man-Machine Studies*, vol. 38, pp. 689-711, 1993.

[10] S. Balbo, J. Coutaz, and D. Salber, "Towards Automatic Evaluation of Multimodal User Interfaces," presented at Conference on Intelligent User Interfaces, 1993.

[11] N. Dahlbäck, L. Joensson, and L. Ahrenberg, "Wizard of Oz Studies - why and how," LiTH-IDA-R-92-19, 1992.

[12] C. G. Wolf and P. Morrel-Samuels, "The use of hand-drawn gestures for text editing," *International Journal of Man-Machine studies*, vol. 27, pp. 91-102, 1987.

[13] C. Baber and K. S. Hone, "Modeling error recovery and repair in automatic speech recognition," *International Journal of Man-Machine Studies*, vol. 39, pp. 495-515, 1993.

[14] B. Mellor and C. Baber, "Modelling of Speech-Based User Interfaces," presented at European Conference on Speech Communication and Technology, Rhodes (Greece), 1997.

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