ERROR RESOLUTION DURING MULTIMODAL HUMAN-COMPUTER INTERACTION*

*Sharon Oviatt** & Robert VanGent*

Center for Human-Computer Communication, Department of Computer Science
Oregon Graduate Institute of Science & Technology

ABSTRACT

Recent research indicates clear performance advantages and a strong user preference for interacting multimodally with computers. However, in the problematic area of error resolution, possible advantages of multimodal interface design remain poorly understood. In the present research, a semi-automatic simulation method with a novel error-generation capability was used to collect within-subject data before and after recognition errors, and at different *spiral depths* in terms of number of repetitions required to resolve an error. Results indicated that users adopt a strategy of switching input modalities and lexical expressions when resolving errors, strategies that they use in a *linguistically contrastive* manner to distinguish a repetition from original failed input. Implications of these findings are discussed for the development of user-centered predictive models of linguistic adaptation during human-computer error resolution, and for the development of improved error handling in advanced recognition-based interfaces.

1. INTRODUCTION

Recent research has indicated a strong user preference to interact multimodally with computers in a variety of different tasks, as well as clear performance advantages when interacting multimodally rather than unimodally (Oviatt, 1996 & in press). During multimodal interaction, people's choice to use a particular mode and to integrate multiple modes depends on: (1) the content being communicated (e.g., digit versus text), (2) the presentation format used (e.g., form-based versus unstructured), (3) the pragmatic function being expressed (e.g., original input versus error correction), and (4) relative speed advantages (Oviatt & Olsen, 1994; Oviatt, 1996 & in press; Rudnicky & Hauptmann, 1992). With respect to salient integration themes, people also prefer to use input modes for different but complementary functions, rather than redundantly (Oviatt & Olsen, 1994; Oviatt, 1996). The most influential factor accounting for multimodal integration appears to be *contrastive functional use of modes*. For example, in Oviatt & Olsen's (1994) research, people switched between spoken and pen-based input to designate some consequential shift in the content or functionality of their expression, which accounted for 57% of all patterned modality use observed in the data. In contrast, simultaneously speaking and writing the same propositional content was rare, accounting for less than 1% of all words expressed.

One major advantage of multimodal system design is its potential for improving both error avoidance and resolution. When free to interact multimodally, users often act upon good intuitions regarding the accuracy of a particular mode for conveying certain content. This is evident through their increased likelihood of writing rather than speaking a foreign surname, relative to other content (Oviatt & Olsen, 1994). A well-designed multimodal system that permits flexibility potentially can leverage from people's natural ability to use modes accurately and efficiently. Furthermore, in a telecommunications study, error analyses revealed that up to 86% of all task-critical errors could have been avoided simply by making a second input mode available to people, rather than only speech (Oviatt, in press). This indicates that the degree of error avoidance possible through multimodal interface design can be substantial.

With respect to error resolution, shifting input modes in a multimodal interface also would be likely to expedite recovery when a recognition error is encountered, since the confusion matrices differ for the same content when spoken versus written. During *error spirals* that are typical of recognition-based systems, which require people to repeat input multiple times before correct recognition occurs, switching to an alternate input mode may be particularly effective for shortcutting a string of repeated failures. To the extent that people are free to alternate modes when interacting multimodally, the likelihood of both avoiding and rapidly resolving errors therefore should be facilitated. Since poor error handling is widely acknowledged to pose a serious bottleneck to the successful commercialization of recognition-based technologies (Rhyne & Wolf, 1994), empirical research could be of assistance in exploring actual benefits of multimodal interfaces.

The goal of the present research was to examine how users strategically adapt and integrate their use of input modes and lexical expressions while resolving system recognition errors in a multimodal interface. In the present study, within-subject data were compared before and after system recognition errors, as well as at different spiral depths in terms of the number of repetitions required to resolve an error. It was hypothesized that users' rate of spontaneously switching modalities and lexical expressions would become increasingly elevated when attempting to resolve errors, and that they essentially would function in a *linguistically contrastive* manner to distinguish a repetition from original failed input. To explore *redundant use of modes* as another possible theme during error resolution, the frequency with which users simultaneously used both modes to emphasize or clarify information during repeats also was assessed. The long-term goal of this research is the development of a user-centered predictive model of linguistic adaptation during human-computer error resolution, as well as the development of improved error handling capabilities for advanced recognition-based interfaces.

2. METHOD

2.1. Subjects, Tasks, and Procedure

Twenty native English speakers, half male and half female, participated as paid volunteers. Participants represented a broad spectrum of different occupational backgrounds.

A "Service Transaction System" was simulated that could assist users with conference registration and car rental transactions. After a general orientation, people were shown how to enter information using a stylus to click-to-speak or write directly on active areas of a

* This research was supported by Grant No. IRI-9530666 from the National Science Foundation.

** First author: Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science & Technology, P.O. Box 91000, Portland, Oregon, 97291 (oviatt@cse.ogi.edu; http://www.cse.ogi.edu/~oviatt/) Second author: Computer Science Department, Stanford University.

form displayed on a Wacom LCD tablet. During data collection, they were free to use either input mode whenever they wished. As input was received, the system interactively confirmed the propositional content of requests by displaying typed feedback in the appropriate input slot. For example, if the system prompted with **Car pickup location:**_____ and a person spoke "San Francisco airport," then "SFO" was displayed immediately after the utterance was completed.

In the case of simulated errors, the system instead responded with "????" feedback to indicate its failure to recognize input. In this case, subjects were instructed to re-enter their information into the same slot until system feedback was correct. Each simulated error required 1-6 repeats before error resolution was successful, thereby simulating spiraling in recognition-based systems. A form-based interface was used during data collection so that the locus of system errors would be clear to users. During orientation, users were familiarized with the type of errors they encountered during testing.

People were asked to concentrate on completing their transaction accurately. They were told that the system was a well developed one with extensive processing capabilities, so they could speak and write normally, express things as they liked, and work at their own pace. If they spontaneously inquired about whether they could correct errors by speaking spelled letters, changing input mode, or some other method, they were told that they could make corrections any way they wished as long as the final transaction was correct. After the session was completed, a post-experimental interview was conducted in which people were asked about system errors.

2.2. Semi-automatic Simulation Method

A semi-automatic simulation technique was used as a tool for collecting high-fidelity data on people's spoken and written input during system error handling. Using this technique, people's input was received by an informed assistant, who performed the role of interpreting and responding as a fully functional system would. The simulation software provided support for rapid subject-paced interactions, which averaged 0.4 second delay between a subject's input and system response. Technical details of the simulation method have been provided elsewhere (Oviatt et al., 1992), although its random-error generation capability was adapted for this study to simulate the appropriate base-rates and properties of recognition errors.

2.3. Research Design

The research design was a within-subject factorial that included the following independent variables: (1) Base-rate of system errors (Low- 6.5% of input slots; High- 20% of slots), and (2) Spiral depth (1-6 repeats required to resolve error). All 20 subjects completed 12 subtasks, two per condition. Half of these involved a low base-rate of errors and half a high one, with the order counterbalanced across subjects. Each of the three high error-rate conditions included six simulated errors, one apiece requiring 1–6 repetitions to resolve. Each of the three low error-rate conditions included two simulated errors, with each of the 1-6 spiral depths represented once. In total, data were collected on 24 simulated errors for each of the 20 subjects, which involved 480 simulated error interactions and 1,680 individual attempts at error resolution.

2.4. Data Coding and Analysis

Speech input was collected using a Crown microphone placed directly behind the tablet, and pen input was captured and printed automatically in the appropriate slots of the interface. All human-computer interaction was videotaped and transcribed for analysis purposes. Dependent measures included: (1) modality preference, simultaneous use of modes, and modality alternation, (2) lexical alternation, interjection of novel lexical content, and (3) subjects' reported perception of recognition errors.

Modality Preference: The total number of words spoken versus written of the total communicated was tabulated both during original input and error resolution.

Simultaneous Use of Modes: The total number of words simultaneously spoken and written of the total communicated was tabulated during error resolution.

Modality Alternation: The likelihood of switching input modes from writing to speech, or vice versa, was summarized for each of the 1-6 serial positions during resolution of a spiral error. For each given serial position, the likelihood represents change from the preceeding to present serial position. The average rate of switching modalities per 100 words also was summarized during both original input and error resolution.

Lexical Alternation: The likelihood of an alternation in lexical expression (e.g., "wdc" to "District of Columbia") was summarized for each of the 1-6 positions during resolution of a spiral error, with each likelihood representing the probability of a change from the preceeding to present serial position. Since lexical content can be modality-specific and a shift in modality can prompt a coincident shift in lexical content, for the present analyses only within-mode lexical changes were tabulated. In addition, the likelihood of introducing new lexical content not yet expressed in a given error spiral was summarized for each serial repetition. Finally, the total number of words for which spelled letters were spoken was tabulated during error resolution.

Perception of Recognition Errors: Major themes were summarized, as well as the percentage of subjects reporting specific beliefs about: (1) the type of errors encountered, (2) the causal basis of errors, (3) effective ways to resolve errors, and (4) subjective reactions to errors.

Reliability: For all measures reported, over 20% of the data were second-scored, with all inter-rater reliabilities exceeding 90%.

3. RESULTS

3.1. Modality Preference

Speech generally was preferred as an input mode, with 81.5% of 10,600 words spoken and 18.5% written during original input. During error resolution, however, spoken language dropped to 70% of all words conveyed, while written input increased to 30%.

3.2. Simultaneous Use of Modes

During error resolution, only 0.7% of all words were simultaneously spoken and written. Since the rate of simultaneously speaking and writing previously was reported to be 0.5% during non-error interactions (Oviatt & Olsen, 1994), there is no evidence that people use redundancy of input as a means of clarifying or emphasizing during error resolution.

3.3. Modality Alternation

During non-error interactions, the base-rate of spontaneously shifting modalities from speech to writing, or vice versa, was 4.8 per 100 words. However, the rate of mode shifts increased 225% to 15.7 during error resolution, significant by paired t, $t = 6.40$ ($df = 19$), $p < .001$, one-tailed. Figure 1 illustrates that the average likelihood of switching input modes also increased over repetitions, with a slight drop between repetitions 5 and 6. After an initial *no-switch strategy* during the first repetition, the likelihood of mode shifting ranged

between .29-.40 and averaged .35, or 1 switch every third repetition. An analysis revealed that the probability of mode shifting on the first repetition averaged .14, significantly lower than the .35 probability of a shift on repetitions 2-6, paired $t = 7.23$ ($df = 19$), $p < .001$, two-tailed. This 149% increase in the likelihood of a mode shift between the first and all subsequent repeats confirms the existence of a qualitatively different no-switch strategy on the first repetition. Finally, although correction of errors in this study was not contingent on users' strategy of mode switching, nonetheless their likelihood of shifting input modes persisted strongly, with only a .05 decrease from beginning to end of the session.

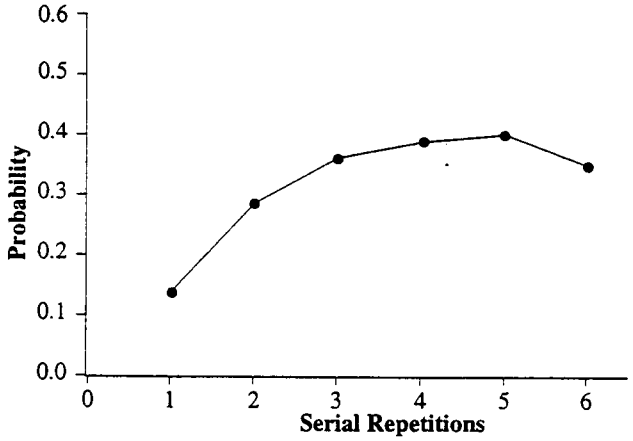

Figure 1: Probability of switching input modes at each serial repetition during a spiral error (1-6)

3.4. Lexical Alternation

The likelihood of introducing novel lexical content, which had not appeared previously in the same spiral error, averaged .15, and was relatively stable across serial repeats. In contrast, after an initial *no-switch strategy* during the first repetition, the probability of alternating lexical content from one repetition to another trended upward, ranging between .21-.37, and averaging .23, or one shift every fourth repetition. An analysis revealed that the probability of a within-mode lexical alternation on the first repetition averaged .14, significantly lower than the .23 likelihood of a shift on repetitions 2-6, paired $t = 3.55$ ($df = 19$), $p < .002$, two-tailed. This 63% increase in the likelihood of a lexical shift between the first and all subsequent repetitions again confirms the presence of a qualitatively different no-switch strategy on the first position.

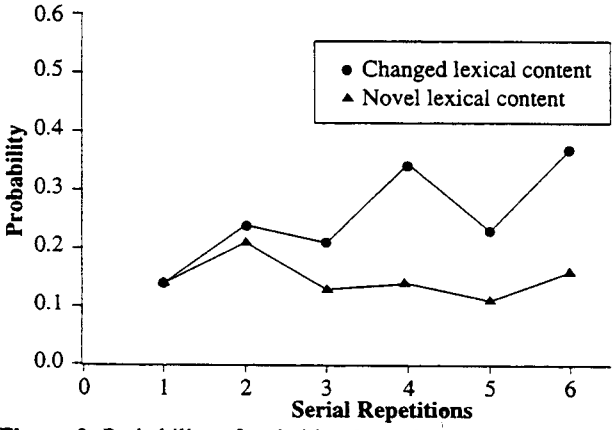

Figure 2: Probability of switching lexical content and introducing novel lexical content at each spiral repetition (1–6)

Essentially, switching input modality and lexical content both were predominant during error resolution after the first repetition, and both intensified over serial repetitions. The cumulative likelihood of alternating either input mode or lexical content is illustrated in Figure 3. After the initial repetition, which represented a no-switch strategy, subsequent repeats involved a large and increasing probability of shift along one or both linguistic dimensions, with an average cumulative likelihood of .52.

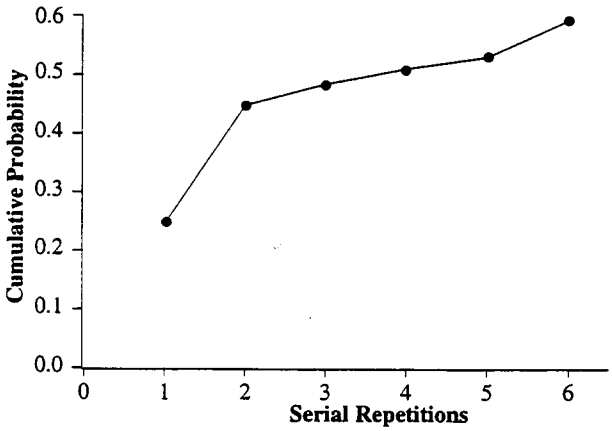

Figure 3: Cumulative probability of changing input mode or lexical content at each serial repetition during a spiral error (1-6)

Although people spontaneously shifted input modes and lexical expressions with high frequency, in only one instance did a subject ever spontaneously speak the spelled letters of a word during error resolution, accounting for 0.06% of all words communicated during error resolution.

3.5 Self-reported Perception of Errors

It was common for users to develop the belief that a particular type of content was more error-prone than others on the basis of just one or two errors they happened to encounter ("It really had trouble with credit card number digits"), even though simulated errors were distributed randomly over task content. However, they did not tend to report believing that one input mode was more error-prone than the other. In fact, for typical users who spoke most of their input, the absolute number of errors encountered when speaking would have been higher than writing, although the ratio of errors to total input in the two modes was equivalent. In this sense, users appeared more hypersensitive to the type of content represented during errors than to modality of expression.

Although errors were delivered randomly, post-experimental interviews revealed that users typically posited a cause for errors that involved self-attribution of blame (e.g., "Oops, I must not have been clear enough"). Although errors were not contingent on input, people nonetheless believed strongly that they could influence the resolution of errors. Of the causal theories expressed, the majority focused on physical and linguistic characteristics of users' own language. More specifically, users reported that: (1) speaking more slowly (55%), and (2) shifting input modality (50%) were their most effective means of resolving errors. Occasionally, users attributed the cause of errors to mechanistic factors, such as system timing or slot-specific failures (e.g., "I think the zip code line was stuck, so I went back and reentered the state name before trying it again").

Users also expressed considerable frustration with repeated errors ("Sometimes it was stubborn as a mule"). They liked the flexibility of being able to shift modes when a repeat error occurred, which they said alleviated their frustration. When using a mode-switch strategy, people frequently reported that their preferred pattern was

to provide information twice in one mode, then switch to the other (e.g., "I'd give it two chances, and then I'd switch").

4. CONCLUSIONS

In a multimodal interface, users' natural inclination is to shift input modes and lexical expressions as they attempt to resolve system errors. Furthermore, the likelihood of using these strategies is accentuated over repeated resolution attempts. On the first repetition following an error, an initial *no-switch strategy* is evident in which users tend to repeat the same lexical content within the same mode. On this first repeat, their speech is adapted toward a hyperarticulated style, which involves elongation of pause and speech segments, as well as increased *clear-speech* phonological adaptations and decreased disfluencies (see Oviatt et al., 1996, this volume). On all subsequent repetitions, users' likelihood of switching modes and lexical expressions averaged .36 and .28, respectively, with a greater than 50% probability of change along one or both dimensions. Compared with non-error interactions, users' rate of spontaneously switching modes increased over 3-fold during error resolution, or 225%. These results are consistent with the view that *contrastive functional use of language* is a predominant theme in human multimodal integration (Oviatt & Olsen, 1994), and one that may be particularly salient when users attempt to distinguish a repetition from original failed input during error resolution.

In users' minds, shifting modalities evidently is a natural and effective option for resolving errors. People uniformly expressed frustration with repeated errors, and 50% reported that changing modes was their preferred means of resolving them. Users also reported that they liked the flexibility of being able to change input modes, which relieved their frustration when trying to resolve repeated failures. A few people described a preferred rhythm of two within-mode resolution attempts, followed by switching modes. Since error resolution was not contingent on mode switching, people's persistence in alternating modes at high levels and their description of mode shifting as effective indicate the extent to which this contrastive strategy is an engrained linguistic pattern.

Given that modalities and lexical expressions are used in a linguistically contrastive manner, and that they involve a finite set of choices, one implication of this research is that repeated language should become less variable due to process of elimination. For example, if the user of a pen/voice system is speaking when the recognizer fails, then pen input becomes more predictable by default if communication modes are being deployed contrastively to distinguish corrections. One effect of this *restricted linguistic variability* is that language modeling may be simplified, and users' input potentially easier to accommodate technologically, to the extent that predictability is enhanced during multimodal error resolution.

Since a high rate of spoken disfluencies has been associated with elevated sentential planning demands, task-critical errors, and task completion time (Oviatt, 1995; Oviatt, 1996 & in press), it can reflect undesirable levels of cognitive load. In comparison with non-error interactions, the rate of disfluent language in the present study actually *decreased* 53% during error resolution (Oviatt et al. 1996, this volume), at the same time that users dramatically *increased* their alternation of input modalities and lexical content. Under the circumstances, this significant drop in disfluencies during error resolution does not support the conjecture that users' active alternation of modes or lexical expressions incurred any excessive cognitive load during multimodal error resolution.

Within a multimodal interface, it was extremely rare for people to spontaneously speak spelled letters, with only one instance of it occurring in 1,680 attempts at error resolution in the present corpus. Clearly, spoken spelling was not a preferred or natural form of error resolution in this context, in comparison with switching modes to write letters. In spite of speculation that redundant use of input modes would be a dominant pattern during multimodal communication, there likewise was no evidence that people use simultaneous, redundant spoken and written input as a technique for emphasis or clarification during error resolution. In fact, less than 1% of error correction language involved simultaneous use of modes. However, redundant use of modes might become a common integration pattern during collaborative use of multimodal systems, especially when one party assumes a tutorial role.

The linguistic and self-report findings in this research confirm the viability and ease of error resolution within a multimodal interface, and the potential of a multimodal interface to support more graceful error handling and overall robust recognition. These findings also emphasize the importance of multimodal flexibility from users' viewpoint, both in alleviating their frustration with the errors that inevitably will occur in recognition-based systems, and by providing an alternative to the within-mode hyperarticulation that has been documented during error resolution (Oviatt et al., 1996, this volume).

5. REFERENCES

1. Oviatt, S. L. Multimodal interactive maps: Designing for human performance, *Human-Computer Interaction,* in press.

2. Oviatt, S. L. Toward multimodal support of interpreted telephone dialogues, in *The structure of multimodal dialogue II* (ed. by M. M. Taylor, F. Ne'el, and D. G. Bouwhuis), John Benjamins: Amsterdam, in press.

3. Oviatt, S. L. Multimodal interfaces for dynamic interactive maps, in *Proceedings of Conference on Human Factors in Computing Systems: CHI '96,* New York, ACM Press, 1996, 95-102.

4. Oviatt, S. L. Predicting spoken disfluencies during human-computer interaction, *Computer Speech and Language*, 1995, 9, 1, 19-35.

5. Oviatt, S. L., Cohen, P. R., Fong, M. W., & Frank, M. P., A rapid semi-automatic simulation technique for investigating interactive speech and handwriting, *Proceedings of the International Conference on Spoken Language Processing* (ed. by J. Ohala et al.), University of Alberta, 1992, vol. 2, 1351-1354.

6. Oviatt, S. L., Levow, G.A., MacEachern, M. & Kuhn, K. Modeling hyperarticulate speech during human-computer error resolution, *Proceedings of the International Conference on Spoken Language Processing*, 1996, in press.

7. Oviatt, S. L. & Olsen, E. Integration themes in multimodal human-computer interaction, *Proceedings of the International Conference on Spoken Language Processing,* 1994, vol. 2, 551-554.

8. Rhyne, J. R. & Wolf, C. G. Recognition-based user interfaces, in *Advances in Human-Computer Interaction,* Vol. 4 (ed. by H. R. Hartson & D. Hix), Ablex Publishing Corp.: Norwood, N. J., 1993, 191-250.

9. Rudnicky, A. I. & Hauptmann, A. G. Multimodal interaction in speech systems, in *Multimedia Interface Design* (ed. by M. M. Blattner & R. B. Dannenberg), Addison-Wesley: Menlo Park, Ca., 1992, 147-171.