

## *Linguistic Adaptations During Spoken and Multimodal Error Resolution\**

SHARON OVIATT, JON BERNARD  
and GINA-ANNE LEVOW

*Oregon Graduate Institute of Science and Technology*

---

### KEY WORDS

*error resolution*

*hyperarticulation*

*linguistic contrast*

*multimodal interaction*

*spiral errors*

*spoken and  
multimodal interaction*

### ABSTRACT

Fragile error handling in recognition-based systems is a major problem that degrades their performance, frustrates users, and limits commercial potential. The aim of the present research was to analyze the types and magnitude of linguistic adaptation that occur during spoken and multimodal human-computer error resolution. A semiautomatic simulation method with a novel error-generation capability was used to collect samples of users' spoken and pen-based input immediately before and after recognition errors, and at different *spiral depths* in terms of the number of repetitions needed to resolve an error. When correcting persistent recognition errors, results revealed that users adapt their speech and language in three qualitatively different ways. First, they *increase linguistic contrast* through alternation of input modes and lexical content over repeated correction attempts. Second, when correcting with verbatim speech, they *increase hyperarticulation* by lengthening speech segments and pauses, and increasing the use of final falling contours. Third, when they hyperarticulate, users simultaneously *suppress linguistic variability* in their speech signal's amplitude and fundamental frequency. These findings are discussed from the perspective of enhancement of linguistic intelligibility. Implications are also discussed for corroboration and generalization of the Computer-elicited Hyperarticulate Adaptation Model (CHAM), and for improved error handling capabilities in next-generation spoken language and multimodal systems.

---

\* Acknowledgments: This research was supported by Grant No. IRI-9530666 from the National Science Foundation, and by grants, contracts, and equipment donations from Apple, GTE Labs, Intel, Microsoft, NTT Data, Southwestern Bell, and U.S. West. We especially thank Robert vanGent, Jon Lindsay, and Eric Iverson for adapting the simulation software to support these studies, Robert and Eric for acting as simulation assistants during data collection, and Robert, Jon, Karen Kuhn, and Yetunde Laniran for assistance with transcription, scoring, and preparation of figures. Thanks to Mark Fanty, Ed Kaiser, Terri Lander, Pieter Vermeulen, Karen Ward, and Lodewyk Wessels of CSLU for discussions and assistance with OGI's Speech Toolkit. Thanks also to Phil Cohen for helpful comments on an earlier manuscript draft. Finally, we are grateful to the people who generously volunteered their time to participate in this research.

Address for correspondence: Sharon Oviatt, Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, Oregon, 97291. E-mail: <oviatt@cse.ogi.edu> WWW site: <<http://www.cse.ogi.edu/~oviatt/>> Collaborators' respective affiliations: Linguistics Dept., University of Chicago, Chicago, IL.; Artificial Intelligence Laboratory, MIT, Boston, MA.

## INTRODUCTION

The speech community presently faces an impasse posed by recognition errors and fragile error handling. Among speech researchers, there is a growing awareness that stylistic variation in spontaneous interactive speech is associated with large degradations in recognition performance, compared with the high recognition rates reported for read speech under controlled laboratory conditions. This realization has come in part from recent experience with domains like the DARPA Switchboard corpus of spontaneous telephone dialogs, for which word error rates in recent benchmark tests have ranged from 30 to 40% (Martin, Fiscus, Fisher, Pallett, & Przybocki, 1997). Furthermore, concerted efforts to improve these recognition rates in 1996 and 1997 have resulted in minimal progress.

### *Speech style variation in real-world settings*

In a recent study, Weintraub and colleagues confirmed that word error rates vary directly with speaking style, such that the more natural the speech delivery the higher the recognition system's word-error rate (Weintraub, Taussig, Hunicke-Smith, & Snodgrass, 1997). For the same speakers and identical lexical content, word error rates increased from 29% during carefully read dictation, to 38% during a more conversationally read delivery, to 53% during natural spontaneous interactive speech. During spontaneous interaction, speakers are engaged in real tasks that involve planning, frequent errors and repairs, coordination with others, fluctuating emotional states, and other social and cognitive demands that actively generate variability in their spoken dialog (Banse & Scherer, 1996; Oviatt & Cohen, 1991; Oviatt, MacEachern, & Levow, 1998).

In addition to the general challenge of spontaneous speech, it has been acknowledged that laboratory assessments overestimate system performance in actual field settings by 20–30% (Gagnoulet, 1989; Karis & Dobroth, 1991; Spitz, 1991). Field and mobile usage conditions include public locations with variable noise levels, interruption of tasks and multitasking, stress and increased human performance errors. These and other factors associated with real-world usage patterns are expected to increase variability in the speech signal further, thereby complicating its intelligibility and processability for many desirable applications (Junqua, 1993; Lively, Pisoni, van Summers, & Bernacki, 1993; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988; Talkmitt & Scherer, 1986; Williams & Stevens, 1969).

In spite of these observations, there currently exists no clear set of diagnostics accounting for why recognition of spontaneous speech is so degraded and difficult to improve. However, different research efforts have begun to examine recognition errors in search of new insights and diagnostic information. One such recent focus has emphasized the phonetic variability typical of spontaneous human speech. In the challenging Switchboard corpus, Greenberg points out that frequently occurring words can be associated with dozens of distinct phonetic realizations— with the most common variant sometimes accounting for as little as 10–15% of total usage among the many forms observed (Greenberg, 1997a, 1997b). This variability is not modeled explicitly by most speech recognition systems, and efforts to process phonetic variation through multiple pronunciation dictionaries typically only capture a few forms involving partial variability (Greenberg, 1997a).

Other research has directly addressed the problem of system recognition errors by examining how users adapt their speech during interactive error handling. This research has emphasized that substantial durational changes occur when users *hyperarticulate*, or use a stylized and clarified form of pronunciation when addressing the computer. During both interpersonal and human-computer communication, speakers will adjust their speech dynamically toward a hyperarticulate style whenever they anticipate or actually experience a communication failure with their listener. Under these circumstances, they hyperarticulate to assist their listener in identifying the lexical content of their speech (Lindblom, 1990; Lindblom, Brownlee, Davis, & Moon, 1992).

In the case of correcting speech recognition errors, the miscommunication that occurs is a particularly powerful elicitor of hyperarticulate change in users' speech signal (Oviatt, MacEachern, & Levow, 1998). Hyperarticulate adaptations, which occur dynamically and sometimes abruptly, primarily include: (1) changes in pause structure toward more pauses and longer pausing, (2) elongation of speech segments, (3) suppression of disfluencies, (4) increase in hyper-clear phonological features, and (5) increase in final falling intonation contours. During correction of a stressed focal region within an utterance, speech adaptations also include: (6) expansion of pitch range, and (7) small increases in amplitude and maximum fundamental frequency on the focal repair region. These characteristics of users' hyperarticulate speech during system error handling are summarized in the Computer-elicited Hyperarticulate Adaptation Model (CHAM) (Oviatt, MacEachern, & Levow, 1998; Oviatt, Levow, Moreton, & MacEachern, 1998).

From the standpoint of building robust spoken language systems, hyperarticulate speech is problematic because it has been associated with elevated rates of system recognition failure (Shriberg, Wade, & Price, 1992). When people hyperarticulate in an effort to correct errors, recognition rates actually can degrade because their speech style departs from the training data upon which the recognizer was developed. As a result, hyperarticulate speech poses a hard-to-process form of signal variability, which is particularly difficult to process because its onset and offset tend to be abrupt. Since hyperarticulation can be both a *reaction* to system recognition failure, and a potential *fuel* for precipitating a higher system error rate — the net effect is that it has the potential to generate a *cycle of recognition failure* (Oviatt, 1996).

### *Designing for error in spoken interfaces*

User acceptance of speech technology is influenced strongly by the error rate, the ease of error resolution, the cost of errors, and their relation to users' ability to complete a task (Kamm, 1994; Frankish, Hull, & Morgan, 1995; Rhyne & Wolf, 1993). When speaking to current interactive systems, the amount of time that users spend resolving errors is substantial not only because of high word error rates, but also because multiple correction attempts often are needed before a repair is successful (Oviatt & vanGent, 1996; Swerts & Ostendorf, 1997). That is, speech recognition systems frequently produce *spiral errors*, or recognition errors that recur consecutively on the same content item. It also has been widely acknowledged that speech technology is by nature inherently error-prone (Rudnicky & Hauptmann, 1992). As a result, future spoken language systems need to be designed to handle recognition errors effectively if they are to perform in an adequately reliable manner and to succeed commercially.

Although "designing for error" has been advocated widely for conventional interfaces (Lewis & Norman, 1986), this concept has not yet been applied effectively to the design of recognition systems involving speech, pen, or similar new input modes. To design for error, one research strategy is: (1) to analyze human-computer interaction during recognition errors, (2) to model users' speech and language during episodes of error handling, and (3) to use these empirical models to guide the design of future spoken language systems with improved error handling capabilities. The present research explores different types of linguistic adaptation that occur during the repair of speech recognition errors. It also analyzes signal adaptations as they unfold over time during spiral errors that require multiple corrections to resolve.

### *Designing for error in multimodal interfaces*

A different approach to resolving the impasse created by recognition errors is to design more flexible multimodal interfaces that incorporate speech as one of two or more input options. Previous research has established that users prefer to interact multimodally, and that their performance can be enhanced as well (Oviatt, 1996, 1997). When interacting multimodally, users tend to deploy input modes for different but complementary functions, or to engage in *contrastive functional use of modes* (Oviatt & Olsen, 1994; Oviatt, deAngeli, & Kuhn, 1997). That is, their use of one input mode rather than another typically designates some salient distinction in the content or functionality of their language. For example, when communicating multimodally users have been observed to speak subject and verb sentential constituents, but to use pen input for locative constituents (Oviatt, deAngeli, & Kuhn, 1997).<sup>1</sup> In comparison, people rarely use speech and writing to convey the same propositional content in a redundant manner (Oviatt & Olsen, 1994).

One major advantage of multimodal interface design is the potential for improving both the avoidance and resolution of errors in a way that leads to superior overall error handling. When free to interact multimodally, users will often select an input mode that they judge to be less error-prone for any given lexical content. For example, they are more likely to write than speak a foreign surname, relative to other content (Oviatt & Olsen, 1994). Users' language also tends to be briefer and simpler linguistically when they communicate multimodally rather than just using spoken language. This also would be expected to expedite the processing of their language and to improve system recognition rates (Oviatt, 1997; Oviatt, deAngeli, & Kuhn, 1997). In both of these respects, a well-designed multimodal system that permits flexibility can leverage from people's natural ability to use input modalities efficiently and accurately, and in a way that potentially avoids recognition errors.

With respect to improved error handling, one question that arises is whether people may use input modalities contrastively during error handling to distinguish a repair from their original failed input. Any tendency to shift input modes would be likely to expedite recovery during a recognition error, since the confusion matrices differ for the same content when spoken versus written. For this reason, switching to an alternate input mode could

---

<sup>1</sup> During the process of speech understanding, complementarity also has been emphasized between the cues provided by the acoustic signal and by lip movements (Stork & Hennecke, 1995).

be a particularly effective strategy for shortcutting a string of repeated failures during spiral errors. The present research analyzes users' linguistic adaptations when they are correcting errors multimodally, and assesses whether their rate of both modality and lexical shifting is higher during corrections than during baseline periods when recognition is proceeding smoothly.

### *Goals and predictions*

The present study generally was designed to investigate how users adapt their speech and language during human-computer error resolution. During data collection, users interacted with a simulated system using spoken and pen-based input as part of a multimodal interface. Within-subject data on different speech and language behaviors were then compared immediately before and after system recognition errors, and at different spiral depths in terms of the number of repetitions needed to resolve an error. Since users can and often do interact unimodally for periods of time even when using a multimodal interface (Oviatt & Olsen, 1994; Oviatt, deAngeli, & Kuhn, 1997), data were also available on how users correct errors during unimodal spoken interactions.

One goal of this research was to examine how input modes and lexical expressions are used when resolving errors multimodally. It was hypothesized that users' rate of spontaneously switching modalities and lexical expressions would increase as they persisted in attempting to resolve errors—essentially functioning in a *linguistically contrastive* way to distinguish a repetition from the original failed input. An additional subgoal was to explore whether redundant or simultaneous use of modes might increase during error resolution as a means of emphasizing or clarifying corrected information.

A second goal was to investigate whether hyperarticulate changes that occur on a users' first correction attempt would continue and perhaps increase in magnitude when further corrections are needed to resolve an error. To assess this possibility, a subset of matched utterance pairs was compared in which users continued speaking the same verbatim lexical content before and after the recognition failure. Based on previous findings involving hyperarticulation during initial corrections (Oviatt, MacEachern, & Levow, 1998), it was predicted that durational increases in speech segments and pauses would become magnified across further correction attempts, and that the likelihood of a final falling intonation contour also would increase across repetitions.

A third goal was to investigate whether the acoustic-prosodic characteristics of users' speech might become more or less variable across correction attempts during spiral errors. The specific goal was to explore corrections in which users spoke a verbatim repair in order to determine whether *variability* in duration, amplitude, and fundamental frequency might change systematically across repeated spoken corrections. In particular, constriction of variability in any of these acoustic-prosodic dimensions could function to create a constant backdrop against which other linguistic changes may become more salient.

A final goal was to explore how these three different types of linguistic adaptation are coordinated to clarify lexical meaning during critical periods of human-computer miscommunication. Since speech is a complex and finely-tuned tool, it was hypothesized that varied linguistic adaptations might well function in unison—in a manner analogous to different facial muscles jointly contracting and relaxing to express a smile. To investigate

this type of interplay, the present research analyzed a comprehensive set of measures with the goal of providing a broader and more holistic view of how different linguistic adaptations function together over time.

To summarize, this study was designed to investigate how users adapt their speech and language in qualitatively different ways during system error handling, including changes that involve:

- increased linguistic contrastivity
- increased hyperarticulation
- changes in acoustic-prosodic variability

The long-term goal of this research is the development of a predictive model of how users adapt their speech and language during human-computer error resolution. Such a model could provide detailed guidance to designers who strive to avoid or resolve errors more gracefully in the future systems they build. A further goal of this research is to establish a better understanding of the advantages and tradeoffs entailed in making the fundamental design choice to build a system multimodally, rather than pursuing a more traditional unimodal system design.

## METHOD

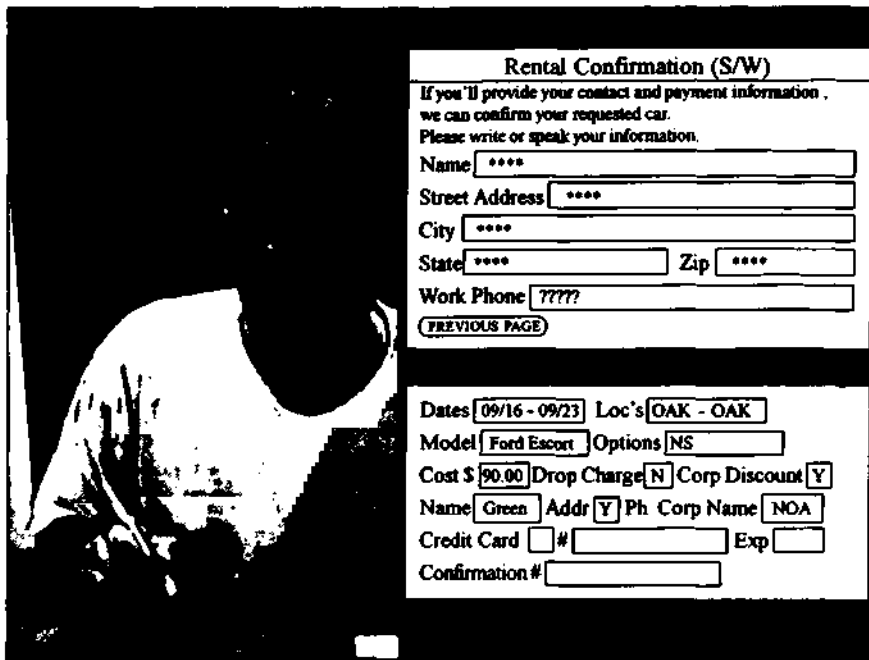
### *Subjects, tasks, and procedure*

Twenty native English speakers, half male and half female, participated as paid volunteers. Participants represented a broad range of occupational backgrounds, excluding computer science.

A "Service Transaction System" was simulated that could assist users with conference registration and car rental transactions. The application content focused on verbal-temporal and numeric information, including proper names and scheduling information. After a general orientation, people were shown how to enter information into a graphic interface by using a stylus to click-to-speak or else write directly onto highlighted areas of a form that was displayed on a Wacom LCD tablet. 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. During these *failure-to-understand* errors (i.e., *rejections*, rather than substitutions or insertions), the system informed the user of its failure to recognize what the user's input meant, so it was not necessary for the user to detect the error. In this case, participants were instructed to try again by re-entering their information in the same slot until system feedback was correct.

A form-based interface was used during data collection so that the locus of system errors would be clear to users. The typical utterance length elicited in this interface varied between 1 and 13 words, with most input involving fragments that averaged 2 to 3 words in length. Figure 1 illustrates a user receiving error feedback after speaking his phone



**Rental Confirmation (S/W)**

If you'll provide your contact and payment information, we can confirm your requested car. Please write or speak your information.

Name

Street Address

City

State  Zip

Work Phone

---

Dates  Loc's

Model  Options

Cost \$  Drop Charge  Corp Discount

Name  Addr  Ph Corp Name

Credit Card  #  Exp

Confirmation #

*"six three one,  
oh two oh seven"*

**Figure 1**

User receives error feedback after speaking his phone number during a car rental transaction.

number during a car rental transaction. To successfully resolve an error, the simulation was programmed so that participants had to repeat their input between one and six times, thereby simulating spiraling in recognition-based systems.

Users were told that the system was a well-developed one with an extensive vocabulary and processing capabilities, so they could express things as they liked and not worry about what they could and could not say. They were advised that they could speak normally, work at their own pace, and just concentrate on completing their transaction. They also were told that if for any reason the computer did not understand them they would know immediately, because it would respond with "?????" feedback. If this occurred, they were instructed that they always would have the opportunity to re-enter their input. Users completed approximately 30 input fields for each of six tasks, which took 30 to 60 minutes for the whole session.

Following their session, all users were interviewed. These face-to-face interviews were conducted by the experimenter, who was a third party separate from the system developer. Users were asked whether they had encountered any errors, what they were like, whether there were more errors when speaking or writing, whether they were able to resolve errors satisfactorily, what they did that resolved errors most effectively, and so forth. Questions were worded neutrally, and participants were encouraged to speak freely and at length about their impressions.

Finally, participants were debriefed about the goals of the research. All participants reported that they had believed the system was a fully functional or "real" one.

*Semiautomatic simulation method*

A semiautomatic 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, with an average delay of 0.4 seconds between a subject's input and system response. Rapid simulation response was emphasized during software design, since it was judged to be an important prerequisite for collecting high quality data on human speech to computers.

To support research specifically on errors, a random error generation capability was developed that could simulate different types of system recognition error, different error base-rates, and different realistic properties of speech recognition errors (e.g., spiraling). This error generation capability was designed to be preprogrammed and controlled automatically so that, for example, errors could be distributed randomly across all task content. For the present study, the error generation software was adapted to deliver failure-to-understand or rejection errors, which were presented in the context of a simulated system that delivered both a low and a high overall base-rate of errors.

Another goal was to simulate a credible system interaction, so that users would be motivated to make themselves understood by what they perceived to be a real system. One shortcoming of previous linguistic studies on clear speech has been the procedural artificiality of simply asking people to "speak clearly" while reading a list — a situation with no natural communication analog, and no premium on intelligibility. With an adequately realistic simulation, it was believed that the type and magnitude of spoken adaptations during error correction would have more ecological validity and therefore utility for system design. Further technical details of the present simulation method have been provided elsewhere (Oviatt, Cohen, Fong, & Frank, 1992).

*Research design*

A within-subject repeated measures design was used to collect data on users' speech and language during error correction episodes with the computer. The main independent variables of interest were *correction status* (i.e., original vs. repeated input), as well as the *spiral depth* of the correction being attempted (i.e., 1–6 repeats required to resolve an error). All 20 participants completed 12 subtasks, two within each of six tasks. Half of these involved a low base-rate of system errors (i.e., 6.5% of input slots) and half a high one (i.e., 20% of slots), with the order counterbalanced across subjects.<sup>2</sup> Each of the three high error-rate conditions included six simulated errors, and each of the three low error-rate conditions included two simulated errors. In total, data were collected on 24 simulated errors for every subject, with each of the six spiral depths represented four times per subject. Overall, 480 simulated error interactions and 1,680 individual attempts at error resolution were available for analysis.

<sup>2</sup> Data for all serial repeat positions represent an average across high and low error-rate tasks.



### *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; data were summarized and analyzed by trained linguists during each error episode.

Dependent measures included: (1) the alternation of input modality (i.e., speech, writing, or both) from one spiral repeat position to the next between an original utterance and its sixth repetition. When input modality remained the same, then (2) change in lexical content was assessed. When input was spoken and involved verbatim repetition of lexical content,<sup>3</sup> then indices of the following acoustic-prosodic characteristics were assessed: (3) duration of speech and pauses, (4) amplitude, (5) fundamental frequency, and (6) intonation contour. A final set of dependent measures focused on interview data involving (7) subjects' self-reported modality preference and perception of recognition errors.

Data on each of the acoustic-prosodic measures only were compared for matched utterance pairs in which *the same speaker spoke the same lexical content* immediately adjacent in time. The speech segments for utterance pairs meeting these qualifications were digitized for any given error episode up until the point at which either a lexical or mode switch occurred. For all such matched utterance pairs involving verbatim spoken input, software from the OGI Speech Toolkit was used to align word boundaries automatically and label each utterance. Most automatic alignments were then hand-adjusted further by an expert phonetic transcriber. The ESPS Waves + signal analysis package was used to analyze amplitude and frequency, and both Waves and OGI Speech Toolkit were used for duration.

*Input modality.* The likelihood of switching input modes between writing and speech was summarized during original input and each of the 1–6 serial positions during a spiral error. For a given serial repetition, the likelihood summarized represents change between the preceding and present serial position. From these data, the average rate of switching modalities per 100 turns was summarized at each serial position. In addition, the percentage of words spoken, written, or simultaneously spoken and written out of the total words communicated was tabulated during original input and for each serial repetition. Finally, the percentage of participants who preferred speech versus written input was also summarized.

*Lexical content.* The likelihood of an alternation in lexical expression (e.g., "dc" 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 preceding 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 total number of words for which spelled letters were spoken was tabulated.

*Duration.* The following were summarized for original input and each of the 1–6 serial repeat positions: (1) average total utterance duration, (2) average speech segment duration

<sup>3</sup> Verbatim repetition of lexical content means that there were no audible lexical deviations in spoken content. For example, if the user said "two oh five" and then "two zero five," or "Susan Johnston" and then "Um, Susan Johnston," such cases were excluded from those scored as verbatim identical.

(i.e., total utterance duration minus pause duration), (3) average pause duration for multiword utterances, and (4) average number of pauses for multiword utterances. In addition, the percentage of utterances involving a  $\pm 20\%$  change in duration from the preceding to present serial position also was assessed as an index of durational variability.

*Amplitude.* Maximum intensity was computed at the loudest point of each utterance using ESPS Waves +, and then was converted to decibels. Values judged to be extraneous nonspeech sounds were excluded. The average amplitude was computed for original input and each of the 1–6 serial positions, as well as the percentage of all repetitions involving a  $\pm 2$  dB change from the preceding to present serial position.

*Fundamental frequency.* Spoken input was coded for maximum F0, minimum F0, and F0 range during original input and each of the 1–6 serial repetitions. Increases or decreases in maximal pitch range were analyzed as an index of change in pitch variability. The fundamental frequency tracking software in ESPS Waves + was used to calculate values for voiced regions of the digitized speech signal. Pitch minima and maxima were calculated automatically by program software, and then adjusted to correct for pitch tracker errors such as spurious doubling and halving, interjected nonspeech sounds, and extreme glottalization affecting  $\leq 5$  tracking points.

*Intonation contour.* The intonation contour of subjects' input was judged to involve a final rise, final fall, or no clear change. Each matched original-repeat utterance pair then was classified as: (1) Rise/Rise, (2) Rise/Fall, (3) Fall/Fall, (4) Fall/Rise, or (5) Unscorable (i.e., one or both utterances containing no clear change). The average percentage of final falling contours per user was summarized for original input and each of the serial repeat positions. In addition, the likelihood of switching final intonation contour from one serial repeat position to the next (i.e., represented by categories 2 and 4) versus holding it the same (categories 1 and 3) also was summarized for each repeat position. The likelihood of a shift for a given repeat position represented change from the preceding to present position.

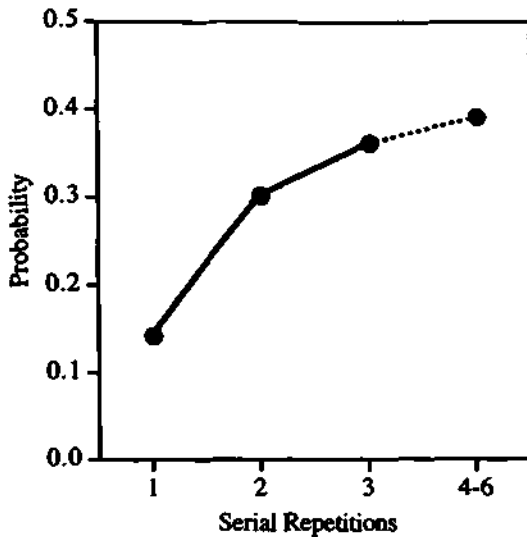
*Self-reported perception of errors.* The percentage of participants who reported different specific beliefs about: (1) the causal basis of errors, and (2) effective ways to resolve errors was summarized.

*Reliability.* For all measures reported except amplitude, lexical expression, and modality, 15–20% of the data were second-scored, with attention to sampling across repeat positions. For discrete classifications that required second scoring, the inter-rater reliabilities on number of pauses exceeded 87%, and the reliability for final intonation contour exceeded 81%. The acoustic-prosodic measures were scored by linguists familiar with the dependent measures and relevant software analysis tools. For fundamental frequency, the inter-rater reliabilities for minimum and maximum F0 were an 80% match with less than a 10Hz departure. For duration, pause length was an 87% match with less than a 26 ms departure, and speech duration an 80% match with less than 52 ms departure.

## RESULTS

### *Input modality*

Speech was the preferred input mode, with 81.5% of the total 10,600 words spoken and



**Figure 2**

Probability of switching input modes from preceding to present serial repetition during a spiral error (solid line marks statistically significant changes).

18.5% written during baseline input when no errors were occurring. During error resolution, spoken language dropped to 70% of all words, while written input increased to 30%.

In addition, redundant use of modes was rare during error handling. Overall, only 0.7% of all words were simultaneously spoken and written.

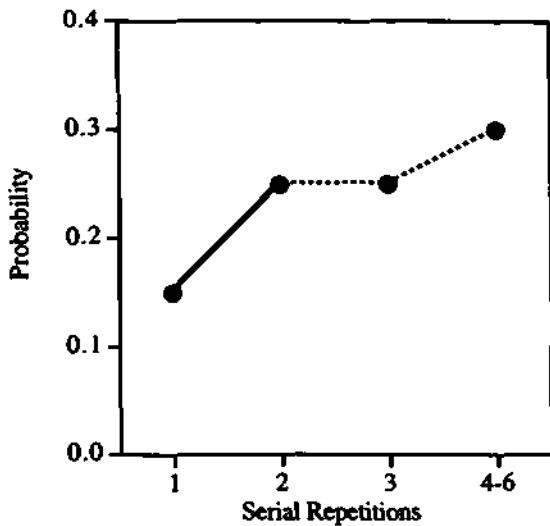
When no errors were encountered, the base-rate probability of spontaneously shifting modalities from speech to writing, or vice versa, was .048. As illustrated in Figure 2, the average likelihood of switching input modes during error correction increased to .14 on the first repetition, .30 on the second, .36 on the third, and .39 on the fourth to sixth.<sup>4</sup> Wilcoxon Signed Ranks tests confirmed that mode shifting increased significantly during error handling between the first and second repetitions,  $z=3.43$  ( $N=17$ ),  $p<.0005$ , one-tailed, and again between the second and third repetitions,  $z=1.70$  ( $N=20$ ),  $p<.05$ , one-tailed. Beyond the third repetition, no significant further change in mode shifting was evident. Overall, the rate of mode shifting increased in magnitude by +179% between original input and its peak rate during the fourth to sixth spiral repetitions.

#### *Lexical content*

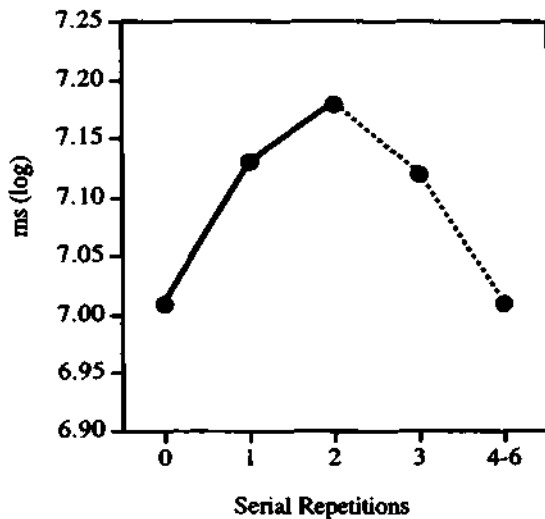
During error resolution, only once did a user ever spontaneously spell a word. Of the total words communicated during error handling, the rate of spontaneous spelling therefore was a negligible 0.06%.

As illustrated in Figure 3, the probability of alternating lexical content from one repetition to the next increased from .15 between original input and the first repetition, to .25 between the first and second repetitions, .25 between the second and third, and .30 between the third and fourth to sixth. Further analysis using Wilcoxon Signed Ranks tests

<sup>4</sup> The total number of data points available per repetition varied for different dependent measures. Since there were diminished data on the deeper spiral positions, repetitions 4-6 were collapsed and presented as an average for all dependent measures. For modality and lexical shifts, maximum data were available (i.e., total  $N=480$  for first correction attempts, 400 for second, 320 for third, and 480 for the average of the fourth to sixth). For acoustic-prosodic dependent measures that were restricted to analyses of matched spoken utterances with identical lexical content, the total  $N$ s ranged lower (i.e., 280 for first attempts, and diminishing correspondingly over repetitions).

**Figure 3**

Probability of switching lexical content from preceding to present serial repetition during a spiral error (solid line marks statistically significant changes).

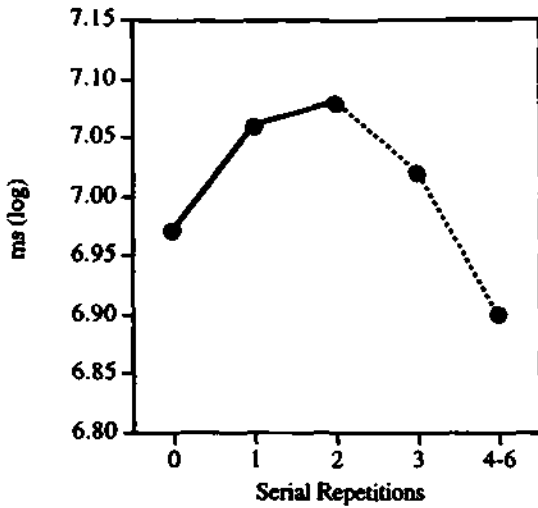
**Figure 4**

Average total utterance duration during original input and each serial repetition of a spiral error (solid line marks statistically significant changes).

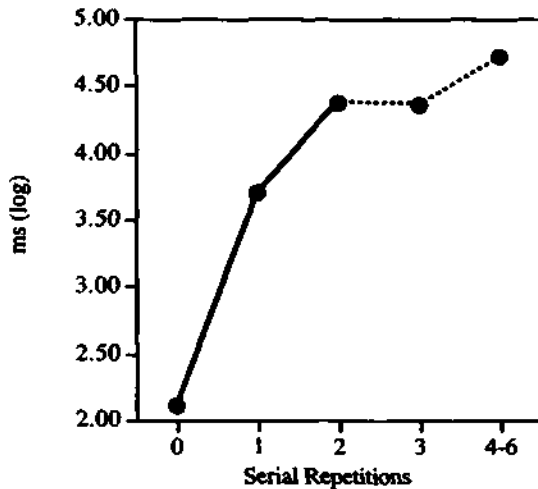
revealed that the probability of a within-mode lexical alternation on the second repetition was significantly greater than on the first,  $z=2.56$  ( $N=20$ ),  $p<.01$ , one-tailed, although increases in lexical alternation beyond the second repetition were not significant. Overall, rate of lexical alternations increased +100% between their initial assessment and peak rate on the fourth to sixth repeat positions.

### *Duration*

The total utterance duration (log transformed) across serial positions averaged 7.01 log ms for original input, 7.13 log ms for the first repetition, 7.18 for the second, 7.12 for the third, and 7.01 for the average of the fourth to sixth repetitions, as illustrated in Figure 4. Paired  $t$ -tests on log transformed data confirmed significant increases in total utterance duration between original input and the first repetition, paired  $t=8.33$  ( $df=126$ ),  $p<.001$ , one-tailed, and again between the first and second repetition, paired  $t=2.84$  ( $df=125$ ),  $p<.003$ , one-tailed, with no additional significant changes in utterance duration on the third or subsequent repetitions,  $t_s<1$ . That is, the apparent drops in Figure 4 at the deeper spiral positions also were not statistically reliable ones.

**Figure 5**

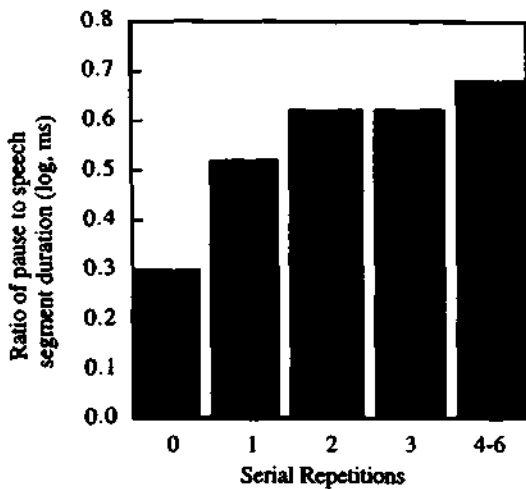
Average total speech segment duration during original input and each serial repetition of a spiral error (solid line marks statistically significant changes).

**Figure 6**

Average total pause duration during original input and each serial repetition of a spiral error (solid line marks statistically significant changes).

Follow-up analyses of speech segment durations (log transformed) revealed averages of 6.97 log ms for original input, and 7.06, 7.08, 7.02, and 6.90 log ms, respectively, for the first through fourth to sixth repetitions, as illustrated in Figure 5. Significant increases occurred in speech duration (log transformed) between original input and the first repetition, paired  $t = 7.09$  ( $df = 126$ ),  $p < .001$ , one-tailed, with a marginal further increase on the second repetition, paired  $t = 1.62$  ( $df = 125$ ),  $p < .055$ , one-tailed, but no significant changes occurred thereafter,  $ts < 1$ . The total relative change in actual speech segment durations between original input (1356 ms) and the peak duration observed on the second repetition (1512 ms) was +12%. Once again, the apparent drops at the deeper spiral positions were not statistically reliable.

Similar analyses of change in total pause duration (log transformed) revealed averages of 2.11 log ms for original input, and 3.70, 4.38, 4.37, and 4.72 log ms, respectively, for the first through fourth to sixth repetitions, as illustrated in Figure 6. Significant increases occurred in pause duration (log transformed) between original input and the first repetition, paired  $t = 6.25$  ( $df = 84$ ),  $p < .001$ , one-tailed, and between the first and second repetitions, paired  $t = 2.15$  ( $df = 84$ ),  $p < .02$ , one-tailed, with no further changes thereafter. In addition,



**Figure 7**

Increasing ratio of pause to speech segment duration (log, ms) during original input and each serial repetition of a spiral error.

analysis of change in the total number of pauses replicated the findings for pause duration. The average number of pauses in multiword utterances increased from 0.67 during original input, to 1.08 on the first repetition, 1.43 on the second, 1.67 on the third, and 1.03 on the average of the fourth to sixth repetitions. Wilcoxon Signed Ranks tests again confirmed a significant increase in number of pauses between original input and the first repetition,  $z=3.18$  ( $N=13$ ),  $p<.001$ , one-tailed, and between the first and second repetitions,  $z=2.62$  ( $N=11$ ),  $p<.005$ , one-tailed, with no significant increases thereafter. Overall, the maximum relative change in total number of pauses between original input and the peak number observed on the third repetition was +149%. Likewise, the maximum relative change in actual pause durations between original input (154 ms) and peak duration (361 ms) was +134%.

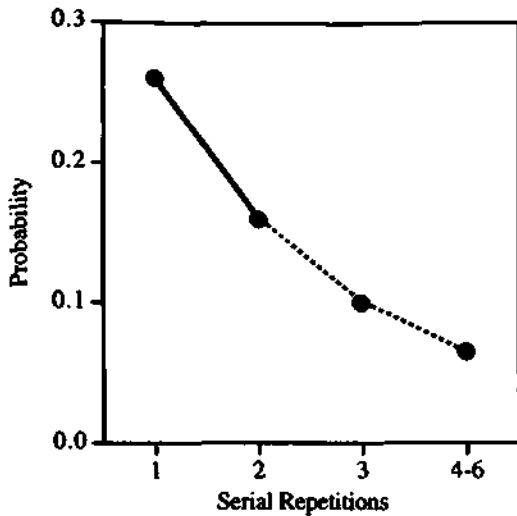
Figure 7 illustrates that there was an increasing ratio of pause to speech segment duration from original input throughout the serial repetitions of a spiral error. That is, pauses became selectively more elongated than the speech segment over repetitions.

Finally, the probability of a change in utterance duration of  $\pm 20\%$  or greater was .35 on the first repetition, .31 on the second, .28 on the third, and .26 on the fourth to sixth repetition. Wilcoxon Signed Ranks tests revealed no significant changes in this measure of durational variability across serial repetitions.

### *Amplitude*

The maximum amplitude level across serial repetitions averaged 70.8 dB for original input, 70.8 dB on the first repetition, 70.9 dB on the second repetition, 71.1 dB on the third repetition, and 71.1 on the fourth to sixth repetitions. Paired *t*-tests revealed no significant change in amplitude level as a function of serial position.

However, analysis of the percentage of shifts in amplitude of  $\pm 2$  dB or greater from the preceding to present position revealed a pattern of declining variability in amplitude. Figure 8 illustrates that the ratio of amplitude shifts of this magnitude or greater decreased from .26 on the first repetition, to .16 on the second, .10 on the third, and .065 on the fourth to sixth repetitions. Wilcoxon Signed ranks tests revealed a significant decline in amplitude variability between the first and second repetitions,  $z=2.73$  ( $N=13$ ),  $p<.006$ , two-tailed, with no further significant reductions thereafter. In short, while absolute



**Figure 8**

Probability of  $\pm 2$  dB shift or greater in amplitude from preceding to present serial repetition of a spiral error (solid line marks statistically significant changes).

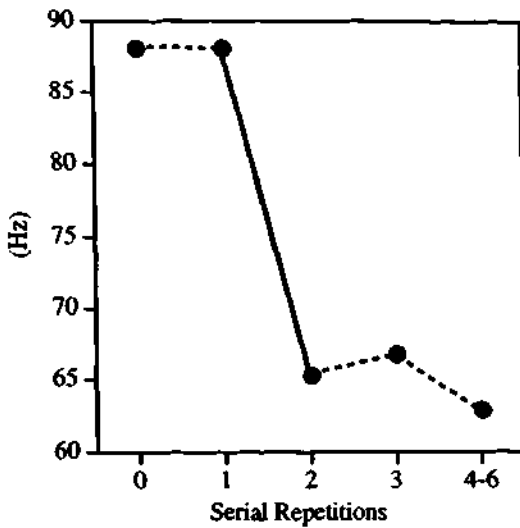
amplitude levels remained constant, variability in the signal's amplitude was suppressed while users persisted in resolving errors after their first correction attempt. Overall, variability in amplitude decreased  $-75\%$  from its baseline through maximum drop on the fourth to sixth repetitions.

#### *Fundamental frequency*

The maximum F0 levels during original input and across repetitions averaged 210.5 Hz for original input, 208.1 on the first repetition, 193.0 on the second repetition, 197.8 on the third repetition, and 176.4 on the fourth to sixth repetitions. Paired *t*-tests revealed a significant decrease in maximum F0 (log transformed) between the first and second repetitions, paired  $t = 6.38$  ( $df = 124$ ),  $p < .001$ , two-tailed, with no other significant changes before or after this point.

Minimum F0 levels averaged 120.8 Hz during original input, 118.2 on the first repetition, 124.7 on the second repetition, 129.1 on the third repetition, and 118.3 on the fourth to sixth repetitions. Paired *t*-tests uncovered a correspondingly significant increase in minimum F0 (log transformed) between the first and second repetitions, paired  $t = 3.58$  ( $df = 124$ ),  $p < .001$ , two-tailed, with no significant change before or after this transition point.

Pitch range during original input and the first repetition averaged 88.1 and 88.0 Hz, but then dropped to 65.3 on the second repetition, 66.7 on the third, and 62.8 on the fourth to sixth repetitions. Analysis of pitch range revealed a significant decline in range between the first and second repetitions, paired  $t = 7.28$  ( $df = 121$ ),  $p < .001$ , two-tailed, with no significant change before or afterwards. Figure 9 illustrates this substantial constriction in pitch range after the first repetition, which then remained narrower during later repetitions. In short, users actively suppressed variability in their speech signal's fundamental frequency when trying to resolve errors after a first attempt. The total relative decline in pitch range from original input through its maximum drop on the fourth to sixth repetitions was  $-29\%$ . This decrease in pitch range was derived from both a decrease in maximum pitch and an increase in minimum pitch between the first and second repetitions.

**Figure 9**

Pitch range during original input and each serial repetition of a spiral error (solid line marks statistically significant changes).

### *Intonation contour*

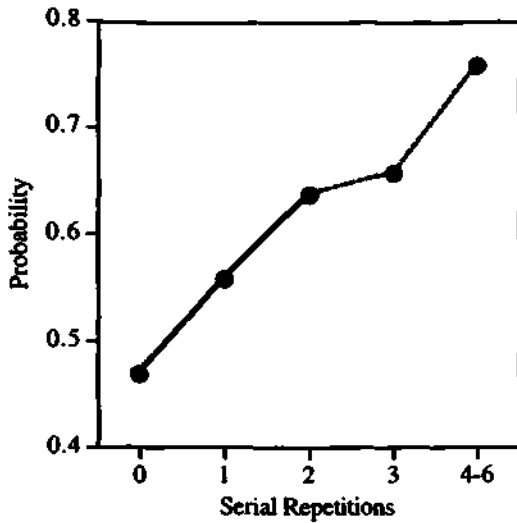
The probability of shifting final intonation contour from a rising to falling pattern, or vice versa, averaged .11 between original input and the first repetition, .145 between the first and second repetitions, .115 on the third repetition, and .13 on the fourth to sixth repetitions. This tendency simply to alternate final intonation contour did not change significantly across serial repetitions.

The likelihood of an utterance ending in a final falling contour increased from .47 during original input, to .56 on the first repetition, .64 on the second repetition, .66 on the third, and .76 on the fourth to sixth repetitions, as illustrated in Figure 10. Wilcoxon Signed Ranks tests confirmed that the likelihood of a final falling contour increased significantly from original input to the first repetition,  $z=2.32$  ( $N=12$ ),  $p<.015$ , one-tailed, and again between the first and second repetitions,  $z=1.84$  ( $N=9$ ),  $p<.035$ , one-tailed, but no significant change occurred thereafter. To summarize, speakers increasingly used an emphatic final falling contour as they persisted in attempting to correct stubborn errors. Overall, the magnitude of relative change in final falling contours between original input and peak rate on the fourth to sixth repeat positions was +62%.

### *Pitch range reduction in utterances with final falling contours*

Since the significantly increased rate of using final falling contours (i.e., from 47% of utterances during original input to 76% by the fourth to sixth repetition) could have influenced the significant constriction found in pitch range, a reanalysis of pitch range was conducted only on the subset of utterances with a final falling contour. For this more controlled subset, a significant decline in pitch range again was evident between the first and second repetition, paired  $t=4.01$  ( $df=59$ ),  $p<.001$ , one-tailed, with no other positions showing any significant change. Pitch range during original input and the first repetition averaged 88.5 and 87.4, but then dropped to 71.3 on the second repetition, 71.1 on the third repetition, and 61.0 on the fourth to sixth repetitions. These findings replicate the original results, confirming that pitch variability was actively suppressed when users persisted in correcting errors after their first try. Therefore, the constriction in pitch range cannot be accounted for simply by the transition toward a more uniform pattern of final





**Figure 10**

Probability of final falling intonation contour during original input and each serial repetition of a spiral error (solid line marks statistically significant changes).

falling contours over later repetitions. The total relative decline in pitch range between original input and the fourth to sixth repetition was  $-31\%$  for this controlled subset, which was similar in magnitude to  $-29\%$  for the entire data set.

#### *Self-reported perception of errors*

Although errors were delivered randomly, postexperimental 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 linguistic characteristics of users' own language. Most users reported that: (1) speaking more slowly (53% of users), and (2) shifting input modality (50% of users) were their most effective means of resolving errors. Occasionally, users attributed the cause of errors to mechanistic factors, such as timing or slot-specific failures (e.g., "I think the zip code line was stuck, so I went back and re-entered 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"). These reports are consistent with the average observed likelihood of approximately one mode shift for every three repetitions.

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 that they happened to encounter ("It really had trouble with credit card number digits"), even though the simulated errors were distributed randomly over task content. However, users did not tend to believe that one input mode was more error-prone than the other, in spite of the fact that when they mainly spoke their input the absolute number of speech errors was higher than writing.<sup>5</sup>

<sup>5</sup> The ratio of errors during speech and pen input was about 1.5.

**TABLE 1**Summary of types and magnitude<sup>†</sup> of linguistic change during error resolution

<i>Type of linguistic change</i>	<i>Magnitude change</i>
1. Increased Linguistic Contrast	
Mode shifts increase	+179%
Lexical shifts increase	+100%
2. Increased Hyperarticulation	
Number of pauses increases	+149%
Total pause time increases	+134%
Final falling intonations increase	+62%
Speech segments increase	+12%
3. Decreased Linguistic Variability	
Amplitude swings decrease	-75%
Pitch range decreases	-29%

<sup>†</sup> All changes listed are statistically significant. Magnitude change listed for each feature is the total relative change from its baseline through the repeat position of maximum change.

In this sense, users appeared more sensitive to the content of errors than to the modality in which errors occurred.

#### *Summary of linguistic adaptations during error resolution*

Table 1 summarizes three qualitatively different types of linguistic adaptation that occurred when users persisted in their attempts to correct system errors. The first type of change involved the use of *increased linguistic contrast* per se as a means of distinguishing a corrected utterance from the original failed one. In Figures 2 and 3 it is evident that users accelerated in their use of both modality and lexical switching during error resolution. Of these two mechanisms, users' strongest predilection was to switch input modalities. In fact, a +179% average increase occurred in users' likelihood of mode shifting between their baseline and peak rate while resolving errors. In comparison, a +100% relative increase was observed in lexical shifts. Both modality and lexical switching are examples of linguistic alternation involving a discrete and relatively high-level linguistic feature.

The second type of change involved an increase in a linguistic feature known to occur during hyperarticulation. Examples of *increased hyperarticulation* include the +149% increase in number of pauses, +134% increase in total pause time, +62% increase in the percentage of utterances involving a final falling contour, and +12% increase in speech segment duration. These adaptations in durational and intonational characteristics, illustrated in Figures 4, 5, 6, and 10, all represent a shift along the spectrum of hyperarticulation toward more deliberate, emphatic, and hyper-clear speech.

The third type of change observed during error handling involved a *reduction of linguistic variability* in continuous acoustic-prosodic variables. Figures 8 and 9 clarify that variability in both amplitude and pitch was actively suppressed during error resolution, with total drops in their variability of -75% and -29%, respectively. This constriction of their typical variability in spontaneous speech may have functioned to create a constant backdrop against which hyperarticulate changes would appear more salient.

## DISCUSSION

When a system makes a recognition error, the miscommunication that occurs can be a forceful elicitor of linguistic adaptation in users. Results from this study indicate that human language changes in at least three different ways during system error handling. When users are free to interact multimodally, they actively alternate input modes and lexical content in a way that increases linguistic contrast. However, in a multimodal interface users can and often do interact unimodally for periods of time. When they make corrections using verbatim speech, users hyperarticulate by lengthening speech segments and pauses, while also increasing their use of a final falling intonation contour to emphasize finality. Perhaps to further clarify these hyperarticulate changes, users also actively suppress variability in their amplitude and fundamental frequency during these spoken corrections.

In this study, the opportunity was created to observe all three of these different types of linguistic adaptation during repeated error correction. Under this circumstance, it was possible to examine the coordinated interplay of these linguistic changes over time as a miscommunication unfolded. Typical users initially responded to system recognition errors by repeating their spoken input verbatim but in a hyperarticulated form. When their first correction did not resolve the error, then by the second attempt about half of users changed the input modality or lexical content of their language. Those who instead continued speaking in a hyperarticulated style, by the second repetition also simultaneously reduced the variability in their amplitude and fundamental frequency.

### *Mode shifting: The preferred contrastive mechanism*

When resolving system errors in a multimodal interface, users' natural inclination is to shift input modes and, to a somewhat lesser extent, lexical expressions. Overall, a +179% relative increase was observed in users' rate of mode shifting between their initial and peak rate during error resolution, and a +100% increase was observed in lexical shifting. Furthermore, a substantial six-fold increase was observed in the average rate of mode shifting during error resolution, in comparison with the baseline rate of spontaneous mode shifting during nonerror interactions. This increased use of linguistic contrast appeared to function as a means of distinguishing a corrected utterance from the original failed one. Since recognition errors represent an aversive experience, these relatively high-level forms of linguistic alternation may have been motivated in part by avoidance of initial strategies associated with an error.

Users' rate of applying these contrastive strategies accelerated substantially over repeated attempts at error resolution. However, the largest increases occurred on the second repetition, when two-thirds of the total increase in both mode and lexical shifting was observed. On the first correction attempt, users typically displayed an initial *no-switch strategy* in which they repeated the same lexical content using the same mode as their original input. At this point, they adapted their speech toward a hyperarticulate style, which is discussed in detail in the next section and in related research (Oviatt, MacEachern, & Levow, 1998). By users' second repetition, however, a strategy shift appeared involving a rapid increase in modality and lexical shifting. From this second repetition on, the likelihood of switching modes and lexical expressions averaged .35 and .27, respectively — or one in three corrections for mode shifting, and one in four for lexical change. Overall,

the cumulative likelihood of change along one or both dimensions exceeded 50% on all correction attempts after the first repetition. These results are consistent with the view that *contrastive functional use of language* is a dominant theme during multimodal integration (Oviatt & Olsen, 1994), and one that is particularly salient when users attempt to distinguish a repetition from original failed input during error handling.

In users' minds, shifting modalities 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. They also liked the flexibility of being able to change input modes, which relieved their frustration when correction attempts were not effective immediately. Since errors are aversive, mode shifting appeared to provide a form of high-level linguistic escape, similar to the way speakers will switch topics or even conversational partners when a miscommunication persists and becomes socially uncomfortable. A few people described a preferred rhythm of two within-mode resolution attempts, followed by a modality switch. Since error resolution was *not contingent* on mode switching, users' persistently high levels of changing modes and their reports of its effectiveness reveal the strength and ingrained nature of this contrastive strategy.

In spite of speculation that redundant use of input modes is a dominant pattern during multimodal communication, in this study people did not simultaneously speak and write to emphasize or clarify corrected information. In fact, less than 1% of error correction language involved simultaneous use of both modes. This finding is consistent with previous reports of a 0.5% rate of simultaneous mode use during nonerror interactions (Oviatt & Olsen, 1994). However, it is possible that redundant use of input modalities might become a more common integration pattern during collaborative use of multimodal systems, especially during tutorial interactions.

The linguistic data and self-report findings in this research support the ease of error handling and potential for more robust recognition within a multimodal interface. From the users' viewpoint, the flexibility of a multimodal interface alleviates subjective frustration with repeat errors, which are characteristic of recognition-based systems. The ability to shift input modes after an error is a natural strategy for avoiding the error and contrasting it with the correction. Finally, since the confusion matrices differ for the same propositional content when spoken versus written, the users' tendency to increase mode shifting during spiral errors typically would be a very effective strategy for undercutting a string of repeated system failures — which could be critical to supporting a robust and viable recognition-based interface. In short, the present research has identified concrete reasons for superior error handling in a multimodal interface, compared with a more traditional unimodal one.

#### *Hyperarticulation during persistent error handling*

The primary hyperarticulate features during error correction episodes included a +149% increase in number of pauses, +134% increase in total pause time, and +12% increase in the duration of speech segments. During error correction over time, rather dramatic increases in pause duration were observed on users' first and second repetitions — with pause time nearly doubling at each of these points. Although significant additional lengthening was not observed after the second repetition, pause duration continued its upward trend through the fourth to sixth repetitions.

In contrast, increases in speech segment durations were more modest in magnitude, with a downward trend but no statistically reliable changes after the second repetition. Since the speech segment constituted most of the total utterance duration, these changes were mirrored in the apparent U-shaped curve observed in total utterance duration. Although neither speech nor total utterance duration actually decreased significantly in the deeper spiral positions, nonetheless the possibility exists that the apparent U-shaped functions may be real, since sample size and experimental precision were somewhat reduced on deeper spirals. To summarize, the increases in pause duration were larger and more persistent during error handling than the changes observed in speech segments. However, all of the durational increases associated with hyperarticulation plateaued somewhat after users' second repetition.

The type and general magnitude of these durational effects in pausing and speech segments is consistent with other recent reports of hyperarticulate change during human-computer error resolution, which have investigated different types of substitution and rejections errors (Oviatt, MacEachern, & Levow, 1998; Oviatt, Levow, Moreton, & MacEachern, 1998). The present durational findings also corroborate the Computer-elicited Hyperarticulate Adaptation Model (CHAM) (Oviatt, MacEachern, & Levow, 1998). However, since durational increases clearly continue beyond users' first correction attempt, one implication of the present results is that the total magnitude of durational increases can be considerably larger than has been reported previously. Furthermore, users increase duration in a phased and graduated manner when they perceive that further reinforcements are needed due to continued system failure. Given the large size of durational changes overall, these data emphasize the need for modeling of *relative durational information*, ideally in a manner specific to a user and the history of a given correction episode. Since current interactive speech systems frequently do not resolve errors successfully after a single correction, algorithms that are optimized to process relative durational changes over successive correction attempts could effectively improve error handling and overall recognition rates.

When people corrected system errors, their use of final falling intonation contours increased, as has been reported previously (Oviatt, MacEachern, & Levow, 1998). Furthermore, the likelihood of final falling contours became considerably more accentuated over repeated correction attempts — with an overall increase of +62% between baseline and the fourth to sixth repetition. This increased use of the final falling contour appears to be an emphatic technique by which users can underscore the intended finality of an error correction subdialog. This pattern is consistent with previous research indicating that a final falling contour and reduction in pitch are the strongest cues used to produce finality judgments during human-human speech (Swerts, Bouwhuis, & Collier, 1994).

#### *Reducing signal variability to magnify hyperarticulate change*

In previous research on hyperarticulation during system error handling, adaptations in fundamental frequency and amplitude have been negligible, with two exceptions. First, even on an initial correction attempt, it is clear that users increase final falling intonation contours to emphasize the intended finality of their error correction subdialogs (Oviatt, MacEachern, & Levow, 1998). Second, users expand pitch range on the stressed focal repair region of an utterance, and to a lesser extent also increase amplitude (Oviatt, Levow, Moreton, & MacEachern, 1998).

When users persisted in correcting errors in this study, a new pattern emerged. Both

amplitude and fundamental frequency became substantially *reduced* in variability across repetitions. Overall, variability in amplitude level was suppressed  $-75\%$  between users' first and fourth to sixth repetitions, and pitch range constricted by  $-29\%$ . In the case of pitch range, an abrupt drop occurred on the second repetition — after which users held their pitch constant within this narrower range throughout the rest of the error episode. Furthermore, analyses revealed that this constriction in pitch range was *completely independent* of the increases in final falling intonation contour that also occurred over repetitions.

The decreased variability evident in amplitude and fundamental frequency reveals that users were actively reducing and holding these acoustic-prosodic dimensions constant. These changes may have functioned to create a more constant backdrop against which hyperarticulate changes were easier to perceive. Reductions in variability of this type do not appear to have been documented clearly in past literature on hyperarticulation, although these findings are consistent with claims based on a single speaker that high variability in amplitude can be associated with poor signal intelligibility (Bond & Moore, 1994). In future research, the issue of whether reductions in pitch range and amplitude variability do heighten the intelligibility of hyperarticulated speech should be investigated further.

The constriction of pitch range observed in this study underscores the complex functioning of fundamental frequency during error correction. In previous research on spontaneous human-human and human-computer speech, pitch range generally has been reported to increase rather than decrease during correction of errors (Ayers, 1994; French & Local, 1986; Oviatt, Levow, Moreton, & MacEachern, 1998). However, previous pitch range expansions have been observed during repair of wrong content on a focal utterance region, and observations have been limited to the first correction attempt. In contrast, the substantial decrease in pitch range found in this study was during *global* rather than focal utterance repairs, and during a *series of correction attempts*. The parameters of fundamental frequency during error correction appear to be particularly sensitive to the precise context of language use, including the history of a repair and the location of the repair within an utterance.

The reduction in variability of these continuous acoustic-prosodic features resulted in greater stability and therefore *predictability* of signal characteristics — which has the potential to enhance intelligibility for human listeners, as well as simplifying language modeling and recognition for machines. Likewise, when modes and lexical expressions that involve a finite set of choices are used in a highly contrastive manner during error correction, then language becomes more predictable due to process of elimination. For example, if a person is speaking when the recognizer fails and they use input modes contrastively following system failure, then the probability of pen input becomes higher by default. In this sense, both of these different types of linguistic adaptation entail a pattern of increasingly constrained language use during error correction, which should be exploited during future interface design to support more robust recognition.

Previous literature on hyperarticulation has too often focused on isolated linguistic measures and individual utterances, rather than examining related linguistic features more comprehensively and as part of a stream of meaningful dialog. In addition, metrics that represent linguistic variability per se have been largely neglected. Using a broader approach to evaluating linguistic change, the present findings have clarified that at least three different types of adaptation co-occur during error resolution when we observe the process unfolding over time. Furthermore, two of these three types of linguistic adaptation do not involve simple

unidirectional change in an absolute measure of hyperarticulation, but rather: (1) *increased linguistic contrastivity* per se of certain high-level linguistic characteristics, and (2) *decreased linguistic variability* of other continuous acoustic-prosodic features. In summary, a dynamic and coordinated interplay was uncovered among different linguistic adaptations as users clarify lexical meaning during critical periods of human-computer miscommunication.

First received: October 10, 1997; revised manuscript received: March 5, 1998;  
accepted: May 25, 1998

## REFERENCES

- AYERS, G. (1994). Discourse functions of pitch range in spontaneous and read speech. *Working Papers in Linguistics*, Ohio State University, Columbus OH, 44, 1–49.
- BANSE, R., & SCHERER, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70 (3), 614–636.
- BOND, Z. S., & MOORE, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, 14, 325–337.
- FRANKISH, C., HULL, R., & MORGAN, P. (1995). Recognition accuracy and user acceptance of pen interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*, (CHI'95) (pp. 503–510). New York: ACM Press.
- FRENCH, P., & LOCAL, J. (1986). Prosodic features and the management of interruptions. In C. Johns-Lewis (Eds.), *Intonation in Discourse* (pp. 157–180). San Diego, CA: College-Hill Press.
- GAGNOULET, C. (1989). Voice replaces dial in new public phones. *International Voice Systems Review*, 1 (1).
- GREENBERG, S. (1997a). On the origins of speech intelligibility in the real world. In *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation*. Baltimore, Maryland.
- GREENBERG, S. (1997b). The Switchboard Transcription Project. In *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation*. Baltimore, Maryland.
- JUNQUA, J. C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93 (1), 510–524.
- KAMM, C. A. (1994). User interfaces for voice applications. In D. B. Roe & J. Wilpon (Eds.), *Voice communication between humans and machines*, (pp. 422–442). Washington, DC: National Academy Press.
- KARIS, D., & DOBROTH, K. M. (1991). Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal of Selected Areas in Communications*, 9 (4), 574–585.
- LEWIS, C., & NORMAN, D. A. (1986). Designing for error. In D. A. Norman & S. W. Draper, (Eds.), *User-centered system design*, (pp. 411–432). Hillsdale, NJ: Lawrence Erlbaum.
- LINDBLOM, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal, (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht: Kluwer.
- LINDBLOM, B., BROWNLEE, S., DAVIS, B., & MOON, S. J. (1992). Speech transforms. *Speech Communication*, 11 (4 and 5), 357–368.
- LIVELY, S. E., PISONI, D. B., van SUMMERS, W., & BERNACKI, R. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America*, 93 (5), 2962–2973.
- MARTIN, A., FISCUS, J., FISHER, B., PALLETT, D., & PRZYBOCKI, M. (1997). System descriptions and performance summary. In *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation*. Baltimore, Maryland.
- OVIATT, S. L. (1996). User-centered design of spoken language and multimodal interfaces, *IEEE Multimedia*, 3 (4), 26–35.

- OVIATT, S. L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12 (1 & 2), 93–129.
- OVIATT, S. L., & COHEN, P. R. (1991). Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 5 (4), 297–326.
- OVIATT, S. L., COHEN, P. R., FONG, M. W., & FRANK, M. P. (1992). A rapid semiautomatic simulation technique for investigating interactive speech and handwriting. In J. Ohala. (Ed.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 2, pp. 1351–1354). Alberta, Canada: University of Alberta Press.
- OVIATT, S. L., DEANGELI, A., & KUHN, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '97)* (pp. 415–422). New York: ACM Press.
- OVIATT, S. L., LEVOW, G., MORETON, E., & MacEACHERN, M. (1998). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of the Acoustical Society of America*, 104 (5), 3080–3098.
- OVIATT, S. L., MACEACHERN, M., & LEVOW, G. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24, 87–110.
- OVIATT, S. L., & OLSEN, E. (1994). Integration themes in multimodal human-computer interaction. In K. Shirai & S. Furui (Eds.), *Proceedings of the International Conference on Spoken Language Processing, Vol. 2* (pp. 551–554). Yokohama, Japan: Acoustical Society of Japan.
- OVIATT, S. L., & vanGENT, R. (1996). Error resolution during multimodal human-computer interaction. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the International Conference on Spoken Language Processing, Vol. 1* (pp. 204–207). Wilmington, DE: University of Delaware and A. I. du-Pont Institute.
- RHYNE, J. R., & WOLF, C. G. (1993). Recognition-based user interfaces. In H. R. Hartson & D. Hix (Eds.), *Advances in human-computer interaction, Vol. 4* (pp. 191–250). Norwood, NJ: Ablex Publishing Corporation.
- RUDNICKY, A. I., & HAUPTMANN, A. G. (1992). Multimodal interaction in speech systems. In M. M. Blattner & R. B. Dannenberg (Eds.), *Multimedia interface design* (pp. 147–171). Menlo Park, CA: Addison-Wesley.
- SHRIBERG, E., WADE, E., & PRICE, P. (1992). Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 49–54). San Mateo, CA: Morgan Kaufmann Publishers.
- SPITZ, J. (1991). Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, San Mateo, CA: Morgan Kaufmann Publishers.
- STORK, D. G., & HENNECKE, M. E. (Eds.) (1995). *Speechreading by humans and machines*. New York: Springer-Verlag.
- SUMMERS, W. V., PISONI, D. B., BERNACKI, R. H., PEDLOW, R. I., & STOKES, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84, 917–928.
- SWERTS, M., BOUWHUIS, D. G., & COLLIER, R. (1994). Melodic cues to the perceived finality of utterances. *Journal of the Acoustical Society of America*, 96 (4), 2064–2075.
- SWERTS, M., & OSTENDORF, M. (1997). Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22 (1), 25–41.
- TOLKMITT, E. J., & SCHERER, K. R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology*, 12, 302–312.
- WEINTRAUB, M., TAUSSIG, K., HUNICKE-SMITH, K., & SNODGRASS, A. (1997). Effect of speaking style on LVCSR performance. In *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub5E Evaluation*. Baltimore, Maryland.
- WILLIAMS, C. E., & STEVENS, K. N. (1969). On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, 40, 1369–1372.



Copyright of Language & Speech is the property of Kingston Press Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.