PREDICTING HYPERARTICULATE SPEECH DURING HUMAN-COMPUTER ERROR RESOLUTION[1]

Sharon Oviatt, Margaret MacEachern and Gina-Anne Levow[2]

ABSTRACT

When speaking to interactive systems, people sometimes *hyperarticulate*— or adopt a clarified form of speech that has been associated with increased recognition errors. The goals of the present study were: (1) to establish a flexible simulation method for studying users' reactions to system errors, (2) to analyze the type and magnitude of linguistic adaptations in speech during human-computer error resolution, (3) to provide a unified theoretical model for interpreting and predicting users' spoken adaptations during system error handling, and (4) to outline the implications for developing more robust interactive systems. A semi-automatic simulation method with a novel error generation capability was developed to compare users' speech immediately before and after system recognition errors, and under conditions varying in error base-rate. Matched original-repeat utterance pairs then were analyzed for type and magnitude of linguistic adaptation. When resolving errors with a computer, it was revealed that users actively tailor their speech along a spectrum of hyperarticulation, and as a predictable reaction to their perception of the computer as an "at risk" listener. During both low and high error rates, durational changes were pervasive, including elongation of the speech segment and large relative increases in the number and duration of pauses. During a high error rate, speech also was adapted to include more hyper-clear phonological features, fewer disfluencies, and change in fundamental frequency. The two-stage CHAM model (Computer-elicited Hyperarticulate Adaptation Model) is proposed to account for these changes in users' speech during interactive error resolution.

KEYWORDS: hyperarticulation, linguistic adaptation, CHAM model, speech recognition errors, human-computer interaction

1. INTRODUCTION

Recognition errors and fragile error handling are regarded by many researchers as the main weakness of current speech technology, which degrade its performance and limit its commercial potential (Kamm, 1994; Rhyne and Wolf, 1993; Yankelovich, Levow and Marx, 1995). Although impressive and steadily improving, benchmark error rates reported for speech recognition systems still are too high to support many applications (ARPA '94 Proceedings; Cohen, 1996; Roe and Wilpon, 1994). As a result, the amount of time that users spend resolving errors can be quite substantial with current systems. Since recognition-based technology is inherently error-prone, graceful error handling clearly is an essential capability for any spoken language interface.

As speech applications emerge for mobile technology, system recognition errors can be expected to increase further. It is well known that laboratory assessments overestimate system performance in actual field settings by 20—30% (Gagnoulet, 1989; Karis and Dobroth, 1991; Spitz, 1991). Mobile usage conditions in natural settings will include public locations with variable noise levels, groups of interacting people, interruption of tasks and multi-tasking, stress and increased human performance errors. All of these factors are expected to increase variability in the speech signal, which could complicate its intelligibility and processability. In this respect, major improvements in error handling also will be important preparation for supporting viable speech interfaces on new mobile systems— which will constitute a far greater challenge to interface design than do present systems.

1.1 Designing for Error

In addition to concern over basic system error rates, a general problem with current systems is the mismatch between speech recognition algorithms and human models of miscommunication, which results in users being unable to predict when or why a system recognition failure will occur (Rhyne and Wolf, 1993). For example, the basis of system recognition errors often cannot be understood in terms of acoustic or other similarity dimensions that govern human errors (Wolf, 1990). In addition, although human misunderstandings usually resolve after one or two repeats, speech recognition errors can *spiral* unpredictably, resulting in the need for users to repeat input multiple times before resolution is achieved. These lengthy attempts at error resolution are frustrating and time-consuming for users, and can result in termination of an interaction (Oviatt and VanGent, 1996). Finally, recognizers frequently perform *worse* rather than better on users' repeated input, which also contrasts with human expectations of communication. With little basis for understanding the cause of recognition errors, people have no guidance in organizing their repeat input to resolve them reliably. In short, current speech recognition systems simply are not designed to be intuitive, nor do they leverage from empirical models of human spoken language in an effort to minimize errors and boost recognition rates.

Although "designing for error" has been advocated widely for conventional interfaces (Lewis and Norman, 1986), this concept has yet to be applied effectively to the design of recognition-based technology. 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 and Morgan, 1995; Rhyne and Wolf, 1993), so it is clear that future spoken language systems will need to be designed for error if they are to succeed commercially. To design for both avoidance and resolution of errors, one research strategy is to analyze human-computer interaction during recognition errors and to model users' speech during interactive error handling. The empirical models generated then can be used to guide the design of future systems with improved error handling characteristics.

1.2 Hyperarticulation and the Cycle of Recognition Failure

Hyperarticulate speech refers to a stylized and clarified form of pronunciation, which has been observed informally in connection with the use of interactive speech systems. From the standpoint of building robust spoken language systems, hyperarticulate speech is problematic since it has been associated with elevated rates of recognition failure (Shriberg, Wade and Price, 1992). To the extent that people do hyperarticulate to speech systems, for example when correcting errors, then recognition rates would be expected to degrade as hyperarticulated speech departs further from the original training data upon which a recognizer was developed. This problem arises because the basic principle of automatic speech recognition is pattern matching of human speech with relatively static stored representations of subword units. Although recognition algorithms typically model phonemes and co-articulation effects, they do not tend to model dynamic stylistic changes in the speech signal that are elicited by environmental factors, such as hyperarticulate speech during miscommunication or Lombard speech during noise.

In particular, current speech recognizers invariably are trained on original error-free input, typically collected under unnatural and constrained task conditions. Realistic interactive speech usually is not collected or used for training purposes, which means that training is omitted on hyperarticulate speech during system error handling. As a result, to the extent that users hyperarticulate, this type of stylized speech presents a hard-to-process source of variability that threatens to degrade recognizer performance. In this sense, the inherent variability of natural interactive speech poses a serious problem for current speech technology, which is known to have special difficulty resolving recognition errors gracefully.

In short, hyperarticulate speech appears to be both a *reaction* to system recognition failure, as well as potentially *fueling* a higher rate of system errors. That is, hyperarticulation has the potential to generate a cycle of recognition failure. These factors appear to contribute to the presence of spiral errors in recognition-based systems (Oviatt and vanGent, 1996), which are a particularly adverse form of error from a usability standpoint.

The design of recognition technology also can contribute to this cycle of recognition failure and, in particular, to *clustering* of recognition errors. For example, one unfortunate property of Hidden Markov Models is the propagation of recognition error, such that a misrecognized word can cause others in its vicinity to be misrecognized too (Rhyne and Wolf, 1993). Likewise, language models based on conditional probabilities can propagate recognition errors, because an error forces the language model into an incorrect state and increases the likelihood of an error on subsequent words (Jelinek, 1985). To summarize, once an error has occurred, the properties of spoken language technology and users' reactive hyperarticulation both can play a role in perpetuating error— thereby complicating prospects for a graceful recovery.

1.3 Modeling Hyperarticulate Speech

Technologists developing speech and language-oriented systems often have asserted that "users can adapt" to computational constraints. As a result, they sometimes have attempted to rely on instruction, training, and practice to encourage users to speak in a manner that matches system processing capabilities. However, studies have demonstrated that giving users practice with a particular system cannot be counted on to improve the performance of a recognition system over time (Frankish, Hull and Morgan, 1995). Likewise, instructions to "speak naturally" do not reliably steer users' input to match the recognizer's training model in a manner that yields improved recognition rates (Shriberg, Wade, and Price, 1992). Finally, when improvement has been demonstrated in recognition rates following user training, it typically has been modest, and any limited effects have not been demonstrated to persist over time (Danis, 1989; Wolf, 1990). This raises concern about the practical utility of training as a long-term approach to managing speech interface design. Training also is intrusive and time-consuming, which generally would be expected to deter people from using a system.

The widely held technology-oriented view that human speech and language are adaptable needs to be modified to acknowledge that there are *constraints on learning*, and therefore adaptability occurs only within natural limits. Human speech involves highly automatized skills organized within modality-specific brain centers for reception and production (Caramazza and Hillis, 1991), and many of the features of human speech production are not under full conscious control— such as disfluencies, prosody, and timing. There are constraints on the extent to which even the most cooperative user can adapt his or her speech production to suit system limitations, such as the need to articulate with artificial pauses between words for an isolated word recognizer. Even when people can concentrate on changing some aspect of their speech, such as deliberate pausing, as soon as they become absorbed with a real task they quickly forget and slip back into a more natural and automatic style of delivery. As a result, it is unrealistic to expect that people can adapt all aspects of their speech to suit system limitations. Human speech is limited in its adaptability, and interface techniques that rely on specific modifications of natural speech patterns should not be assumed effective without closer examination.

One approach to the design of spoken language systems is to model the speech upon which a system must be built, and then to design spoken interface capabilities that leverage from these existing and strongly-engrained speech patterns. For example, recent research has identified disfluent language as a hard-to-process source of linguistic variability in speech to interactive systems, and has developed predictive models accounting for the rate of disfluencies during different types of human-computer interaction (Oviatt, 1995). One premise of this research is that knowledge of the cognitive factors that drive disfluencies makes it possible to design corresponding interface techniques that minimize their occurrence (Oviatt, Cohen and Wang, 1994; Oviatt, 1995). This basic modeling approach likewise could be used to model other difficult sources of variability in human speech to computers, such as hyperarticulation, thereby providing a basis for effective interface design. However, to date it remains unclear precisely what the definition of hyperarticulation is in the context of human-computer interaction.

1.4 Speech Adaptations to Risk Populations

Although literature on hyperarticulate speech to computers currently is lacking, some guidance is available from related research on how people routinely adapt their speech during human-human exchanges when they expect or experience a comprehension failure from their listener. In the linguistic and psychological literature on interpersonal speech, a variety of listener and situational factors have been associated with variations in speaking style. For example, systematic modifications have been documented in parents' speech to infants and children (Ferguson, 1977; Fernald et al., 1989; Garnica, 1977), in speech to the hearing impaired (Picheny et al., 1986), and in speech to nonnative listeners (Ferguson, 1975; Freed, 1978). Even young preschool children actively adapt their speech to accommodate perceived listener characteristics (Shatz and Gelman, 1973). Systematic speech modifications also have been observed in noisy environments (Hanley and Steere, 1949; Junqua, 1993; Schulman, 1989; Summers et al., 1988), in environments involving heavy workload or that precipitate psychological stress (Brenner et al., 1985; Lively et al., 1993; Tolkmitt and Scherer, 1986; Williams and Stevens, 1969), and when speakers are asked to "speak clearly" in laboratory settings (Cutler and Butterfield, 1990 and 1991; Moon, 1991; Moon and Lindblom, 1994).

The specific hyperarticulate adaptations observed in these cases have differed depending on the target population and situational context. For example, speech adaptations to infants often include elevated pitch, rapid pitch excursions, expanded pitch range, and stress on new vocabulary content— features that assist in gaining and maintaining infants' attention and that subserve teaching functions (Ferguson, 1977; Fernald et al., 1989; Garnica, 1977). In the case of communication with hearing-impaired individuals, speech is higher in amplitude and fundamental frequency, longer in duration, and contains hyper-clear phonological features (e.g., increased segment insertions, decreased burst elimination) (Picheny et al., 1986). Research on speech to nonnative listeners has emphasized simplification of the lexicon and grammar (Ferguson, 1975; Freed, 1978), but also has revealed a speech style that is briefer and more clearly articulated (Freed, 1978).

1.5 Speech Adaptations in Adverse Environments

Speakers likewise respond to noise by dynamically and sometimes abruptly modifying their speech in accord with the "Lombard effect" (Lombard, 1911). Adaptation to a noisy environment involves an increase in vocal effort that manifests itself as more than simple amplification of the speech signal. It also includes change in articulation of consonants, and increased duration and pitch of vowels (Junqua, 1993; Schulman, 1989). In addition, the adaptations observed in Lombard speech have sometimes included variability due to gender effects (Junqua, 1993). Under conditions of high workload, speakers typically increase both amplitude and variability in amplitude, while simultaneously speaking at a faster rate and with decreased pitch range (Lively et al., 1993). When speakers are stressed by their environment, noteworthy adaptations include an increase in fundamental frequency and change in pitch variability (Brenner et al., 1985; Tolkmitt and Scherer, 1986; Williams and Stevens, 1969).

Finally, when simply instructed by an experimenter to "speak extra clearly," linguistic researchers have found phonological change toward hyper-clear articulation, restriction in the magnitude of duration-dependent "vowel undershoot," and increased amplitude, pitch, and duration (Moon, 1991; Moon and Lindblom, 1994). Other studies involving similar instructional manipulation have discovered that word boundaries are marked by selective insertion and lengthening of pauses in hyper-clear speech, especially before weak syllables (Cutler and Butterfield, 1990 and 1991). In this latter literature involving laboratory-based instructional manipulation, it should be qualified that speakers typically have not had a natural listener, nor have they engaged in interactive speech toward the achievement of a goal.

Clearly, the interpersonal dynamics associated with this spectrum of very different populations and circumstances vary, even though all of them can be viewed as "high risk" communications. Although they share some features in common, the acoustic-prosodic and phonological features observed in these different cases nonetheless are defined by distinct profiles (see Uchanski et al. (1996) for discussion of speech to the hearing impaired vs. in a noisy environment; see Freed (1978) for discussion of speech to nonnative speakers vs. children). Speakers' expectations about the likely cause of communication failure in each of these cases appears to influence the hyperarticulate characteristics of their speech. However, the relation between speakers' model of a listener and the manner in which they tailor their speech is a topic that is poorly understood. Likewise, little currently is known about the form and magnitude of hyperarticulatory change during human-computer interaction.

1.6 The Spectrum of Hyperarticulation: When and Why Speech is Adapted

Based on experimental phonetics data, Lindblom and colleagues maintain that human speech is highly plastic. That is, the relation between the speech signal and intended phonemes is a highly variable one (Lindblom, 1996), which is not entirely captured by positing a constant mapping between phonemes and physical acoustic or phonetic characterizations, nor by factoring in local coarticulation effects. Speaking style, ranging from conversational to hyper-clear, also contributes substantially to natural variability in the speech signal.

Lindblom and colleagues have argued that speakers make a moment-by-moment assessment of their listener's need for explicit signal information, and they actively adapt their speech production to the perceived needs of their listener in a given communicative context (Lindblom, 1990 and 1996; Lindblom et al., 1992). This adaptation varies along a continuum from *hypo- to hyper-clear speech*. Hypo-clear speech is conversational, relaxed, and contains phonological reductions. A hypo-clear speech style requires minimal expenditure of articulatory effort by the speaker, and instead relies more on the listener's ability to fill in missing signal information from knowledge. In contrast, hyper-clear speech is a clarified style of articulating that requires more effort. It is designed to achieve ideal target values for the acoustic form of vowels and consonants, thereby relying less on listener knowledge.

Along this spectrum of hypo- to hyper-clear articulatory effort, speakers trade off between economizing effort and achieving intelligibility. When a speaker perceives no particular threat to their listener's ability to comprehend them, he or she typically economizes by relaxing articulatory effort (Lindblom, 1996). The result is hypo-clear speech, which represents the default speaking style. On the other hand, when a threat to comprehension is anticipated, as in a noisy environment or when a listener's hearing is impaired, the speaker will adapt their speech toward hyper-clear to deliver more explicit signal information. In this sense, phonetic signals are actively modulated by the speaker to complement their listener's perceived speech processing ability and world knowledge. The effect of these adaptations is to assist the listener in identifying the signal's intended lexical content. Lindblom believes that speakers operate on the principle of supplying *sufficient discriminatory information* for the listener to comprehend their intended meaning, while at the same time striving for articulatory economy. To summarize, within Lindblom's framework, the speech signal and its phonetic gestures are modulated and tuned adaptively in accordance with on-line communicative demands.

In accord with these theoretical notions, there is corroborating evidence that adaptation toward hyperarticulate speech improves intelligibility by listeners. For example, in a variety of studies involving normal and impaired listeners, the following hyper-clear speech characteristics have been associated with improved intelligibility— increased duration of speech segments and pauses, slower speaking rate, increased duration of vowels and clearer differentiation of the vowel space with respect to formant values, more distinct VOT distributions for voiced and voiceless consonants, increased amplitude and reduced variability in amplitude, increased consonant-to-vowel amplitude ratio, and increased pitch and expansion of pitch range (Bond and Moore, 1994; Chen, 1980; Cutler and Butterfield, 1990; Gordon-Salant, 1987; Lively et al., 1993; Moon, 1991; Payton et al., 1994; Picheny et al., 1985; Uchanski et al., 1996). These studies confirm that listeners can recover lexical content more successfully when speech is adapted toward the hyperarticulate end of the spectrum, rather than being conversational. However, there can be differences in the intelligibility advantage of hyperarticulate speech under different circumstances— for example, hyperarticulate speech produced to accommodate the hearing impaired, compared with that produced in a noisy environment (Uchanski et al., 1996). Finally, extreme circumstances can elicit adaptations such as shouting that actually degrade intelligibility (Pickett, 1956).

1.7 The Computer as "At-Risk" Listener

All of the above research focuses on adaptations of interpersonal speech during anticipated or actual communication errors. At present, the type and magnitude of speech adaptations during human-computer interaction simply is not known. It is difficult to predict how users might adapt their speech to a computer "partner" during failure, since the communication model that people adopt for speaking intelligibly to different "at-risk" human listeners and in adverse environments is so poorly understood. Each of these types of communication involves different interpersonal dynamics, and would be expected to differ in terms of the speaker’s model of probable causes of communication failure— which in turn could influence the nature of adapted signal characteristics.

One relevant question is: *What is speakers’ impression of their computer listener’s ability to successfully extract lexical content from a speech signal?* Although we know that humans respond to computers as social agents in certain ways (Nass, Steuer and Tauber, 1994), especially during spoken interaction, nonetheless there are sources of human failure that speakers may not expect when addressing a computer. For example, users may not expect that they need to work to attract and maintain the computer’s attention, nor that a computer in a quiet office environment would have difficulty with basic audibility. On the other hand, speakers may be concerned with whether a computer can segment rapid speech, whether it can interpret specific sounds as words, and whether it shares a common vocabulary with them. They also may have concerns about whether a computer can interpret them reliably, or whether it can approximate the competence of a human listener— even one considered to be "at risk." Since communication bandwidth is limited with a computer (e.g., precluding gestures, headnodding, and other nonverbal cues), another question that arises is whether speakers may exaggerate hyperarticulate speech to a computer as compensation for loss of bandwidth. Although it is possible that people may adapt their speech to a computer similarly to a human listener, the exact profile of linguistic adaptation is not known.

1.8 Goals and Predictions of the Study

The goal of the present research was to identify the type and magnitude of linguistic adaptations that occur during human-computer interaction involving error resolution. The specific goals of this study were: (1) to develop a flexible simulation method for collecting data on speech and language during system error handling, (2) to provide a comprehensive analysis of acoustic, prosodic, and phonological adaptations in speech during error resolution, (3) to construct a user-centered predictive model of linguistic adaptation during human-computer error resolution, and (4) to generate implications for improved error handling capabilities in next-generation spoken language and multimodal systems. It was hypothesized that users' repetitions during error resolution would be adapted toward clear speech acoustic-phonetic features, including higher amplitude and fundamental frequency, greater frequency range, longer duration, more clear-speech phonological features and fewer disfluencies. To make these assessments, within-subject data were examined for matched utterance pairs in which speakers repeated the same lexical content immediately before and after a simulated recognition error. Speech data also were compared during both a high and low base-rate of errors to assess the possibility of magnified hyperarticulation effects during high error-rate conditions. The long-term goal of this research is the development of a user-centered predictive model of linguistic adaptation during human-computer error resolution, and the development of improved error handling capabilities for spoken language and multimodal interfaces.

1.9 Simulation Method for Research on Errors

One general purpose of this research was to devise a flexible simulation method for supporting varied studies on user responding during system errors— a method that could be adapted easily to examine different aspects of error handling. The simulation developed for this purpose was an adapted version of a semi-automatic simulation method previously outlined by Oviatt et al. (1992). Using this technique, people's spoken input was received by an informed assistant, who performed the role of responding as a fully functional system. The simulation software provided support for rapid subject-paced interactions, which averaged 0.4-second delay 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. This error generation capability was designed to be pre-programmed 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 errors, which were presented at both low and high error base-rates.

Another goal of the present research was to make the simulation 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 analogue, and no particular premium on intelligibility. With an adequately realistic simulation, it was believed that the magnitude of any spoken adaptations during error correction would be more representative of those with a real system.

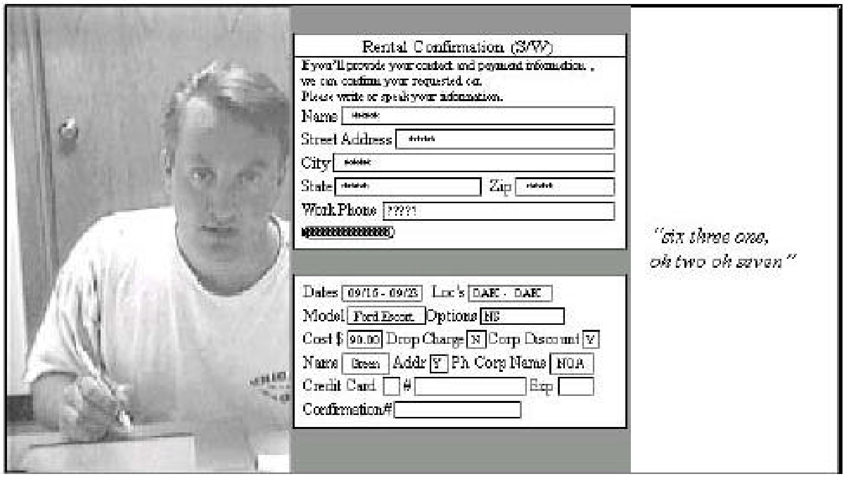
2. METHOD

2.1 Participants, 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. 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 form 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, 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.



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 09/16 - 09/23 Loc's OAK - OAK
Model Ford Escort Options NS
Cost $ 90.00 Drop Charge N Corp Discount Y
Name Green Addr Y Ph Corp Name NOA
Credit Card □ # □ Exp □
Confirmation# □

Figure 1. User receives error feedback after speaking his phone number during a car rental transaction.

A form-based interface was used during data collection so that the locus of system errors would be clear to users. Figure 1 illustrates a user receiving error feedback after speaking his phone number during a car rental transaction. To successfully resolve a simulated error, the simulation was programmed so that the participant 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 didn't understand them they would know immediately, because it would respond with "**????**" If this occurred, they were instructed that they always would have the opportunity to re-enter their input.

Following their session, all users were interviewed and then debriefed about the nature of the research. All participants reported that they believed the system was a fully functional one.

2.2 Research Design

The research design was a within-subject factorial that included the following independent variables: (1) Error status of speech (Original input; Repeat input after error), (2) Base-rate of system errors (Low— 6.5% of input slots; High—20% of slots). All 20 participants completed 12 subtasks, half involving a low base-rate of errors and half a high one, with the order counterbalanced across subjects. In total, data were collected on 480 simulated errors, of which over 250 involved the same speaker repeating identical

lexical content during the first repetition of a repair attempt3. For these matched utterance pairs, original input provided a baseline for assessing and quantifying the degree of change along linguistic dimensions of interest.

2.3 Data Coding and Analysis

Speech input was collected using a Crown microphone, and all human-computer interaction was videotaped and transcribed. The speech segments of matched utterance pairs involving original input and first repetitions were digitized, and software was used to align word boundaries automatically and label each utterance. Most automatic alignments then were hand-adjusted further by an expert phonetic transcriber. The ESPS Waves+ signal analysis package was used to analyze amplitude and frequency, and the OGI Speech Tools were used for duration measurements. For the present acoustic-prosodic and phonetic analyses, only the first repair was compared with original input, although disfluency rates were based on all spoken repetitions that occurred during error resolution.

Duration. The following were summarized: (1) total utterance duration, (2) total speech segment duration (i.e., total duration minus pause duration), (3) total pause duration for multi-word utterances in which at least one pause was present, and (4) average number of pauses per subject for multi-word utterances. No attempt was made to code pauses less than 40 msec in duration. Due to difficulty locating their onset, utterance-initial voiceless stops and affricates were arbitrarily assigned a 20-msec closure, and no pauses were coded as occurring immediately before utterance-medial voiceless stops and affricates.

Rate of Speech. The rate of speech was calculated in milliseconds (msec) per syllable by dividing total utterance duration by the total number of syllables.

Amplitude. Maximum intensity was computed at the loudest point of each utterance using ESPS Waves+, and then was converted to decibels (dBs). Values judged to be extraneous non-speech sounds were excluded.

Fundamental Frequency. Spoken input was coded for maximum F0, minimum F0, F0 range, and F0 average. 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 non-speech sounds, and extreme glottalization affecting = 5 tracking points. To avoid skewing due to line noise, only voiced speech within the coder-corrected F0 range was used to calculate F0 mean.

Intonation Contour. The final rise/fall intonation contour of subjects' input was judged to involve a rise, 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. The likelihood of switching final intonation contour from original input to first repetition (categories 2 and 4) versus holding it the same (categories 1 and 3) then was analyzed. In the case of a shifting contour from original to repeated input, the likelihood of changing from a rising to falling contour versus a falling to rising one also was evaluated.

Phonological Alternations. Phonological changes within original-repeat utterance pairs that could be coded reliably by ear without a spectrogram were categorized as either representing a shift from conversational-to-clear speech style, or vice-versa. The following contrasting categories were coded: (1) released and unreleased plosives, (2) unlenited coronal plosives and alveolar flaps, and (3) presence versus absence of segments. Alveolar flaps, deleted segments, and unreleased stops were considered characteristic of conversational speech, whereas unlenited coronal plosives, undeleted segments, and audibly released stops were indices of clear speech4. A focus was placed on identifying uncontroversial phonological changes with respect to the conversational-to-clear speech continuum, and those that could be reliably coded by ear without access to a spectrogram. For example, cases of glottalization and glottal stop insertion were not included due to known difficulty with reliability when coding by ear (Eisen, Tillmann and Draxler, 1992).

Disfluencies. Spoken disfluencies were totaled for each subject and condition during original spoken input as well as errors (i.e., including all 1-6 repeats), and then were converted to a rate per 100 words. The following types of disfluencies were coded: (1) content self-corrections, (2) false starts, (3) repetitions, and (4) filled pauses. For further classification and coding details, see Oviatt (1995).

Self-Reported Perception of Recognition Errors. The percentage of subjects reporting specific beliefs about the causal basis of errors, and effective ways to resolve errors was summarized from post-experimental interviews.

Reliability. For all measures reported except amplitude, 10% to 100% of the data were second-scored, with attention to sampling equally across conditions. Acoustic-prosodic and phonological alternation measures were scored by linguists familiar with the dependent measures and relevant software analysis tools. For discrete classifications, such as number of pauses, disfluencies, and phonological alternations, all inter-rater reliabilities exceeded 87%. For phonological alternations, only cases agreed upon by both scorers were analyzed. For fundamental frequency, the inter-rater reliability for minimum F0 was 90% with a 0 hz departure, and for maximum F0 80% with 3 hz departure. For duration, pause length was an 80% match with less than a 50 msec departure, and total utterance duration an 80% match with less than 40 msec departure.

3. RESULTS

The following sections summarize acoustic-prosodic and phonological dimensions of change found in hyperarticulate speech during error resolution. The relation between these data and speakers' self-reports about the nature of system errors also are outlined. Spoken utterances in this corpus tended to be brief fragments averaging two to three words, with all input ranging from 1-13 words in length.

3.1 Duration

When the error rate was low, total utterance duration averaged 1544 msec and 1802 msec during original and repeat input, a gain of +16.5%, a significant increase by paired t test on log transformed data, t = 7.05 (df = 49), p < .001, one-tailed. When the base-rate of errors was high, the total utterance duration averaged 1624 msec during original input, increasing to 1866 msec during repeat input, a gain of +15%, which again was significant by paired t test on log transformed data, t = 10.71 (df = 208), p < .001, one-tailed. No significant differences were found simply as a function of error base-rate.

Speech Segment Duration. Analyses revealed an increase in the total speech segment from an average of 1463 msec during original input to 1653 msec during repeat input when the error rate was low, a +13% gain, significant by paired t test on log transformed data, t = 7.44 (df = 50), p < .001, one-tailed. During a high error-rate, it also increased from 1515 msec during original input to 1686 msec during repeat input, an +11.5% gain, significant by paired t test on log transformed data, t = 10.20 (df = 215), p < .001, one-tailed. Again, no significant differences were revealed as a function of error base-rate.

Pause Duration. The total pause duration of multi-word utterances increased significantly from an average of 112 to 209 msec between original and repeat input when the error rate was low, an +86.5% gain, significant by paired t test on log transformed data, t = 2.87 (df = 22), p < .005, one-tailed, and it again increased significantly from an average of 159 msec during original input to 261 msec during repeat input when the error rate was high, a +64% gain, significant by paired t test on log transformed data, t = 6.97 (df = 81), p < .001, one-tailed. No significant difference was revealed due to error rate.

To test for elongation of individual pauses (i.e., independent of interjecting new ones), original and repeat utterance pairs matched on total number of pauses were compared for total pause length. This analysis confirmed that pauses were elongated significantly more in repeat utterances, paired t = 1.71, (df = 27), p < .05, one-tailed.

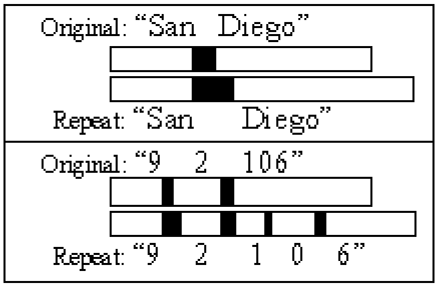


Figure 2: Pause elongation (top) and pause interjection (bottom) in matched original-repeat utterance pairs.

Number of Pauses. The average number of pauses per subject in multi-word utterances increased from 0.49 during original input to 1.06 during repeated speech when the error rate was low, a +116% gain, which was significant by Wilcoxon Signed Ranks test, z = 2.52 (N = 12), p < .006, one-tailed. During a high error base-rate, it increased from 0.57 to 0.95 during repetitions, or +67%, again significant by Wilcoxon, z = 3.03 (N = 16), p < .001, one-tailed. Figure 2 illustrates the general changes in pause structure in a typical utterance pair taken from the present corpus. It shows the average relative gains in total pause duration (+75%) and total number of pauses (+92%) in relation to average speech segment elongation (+12%).

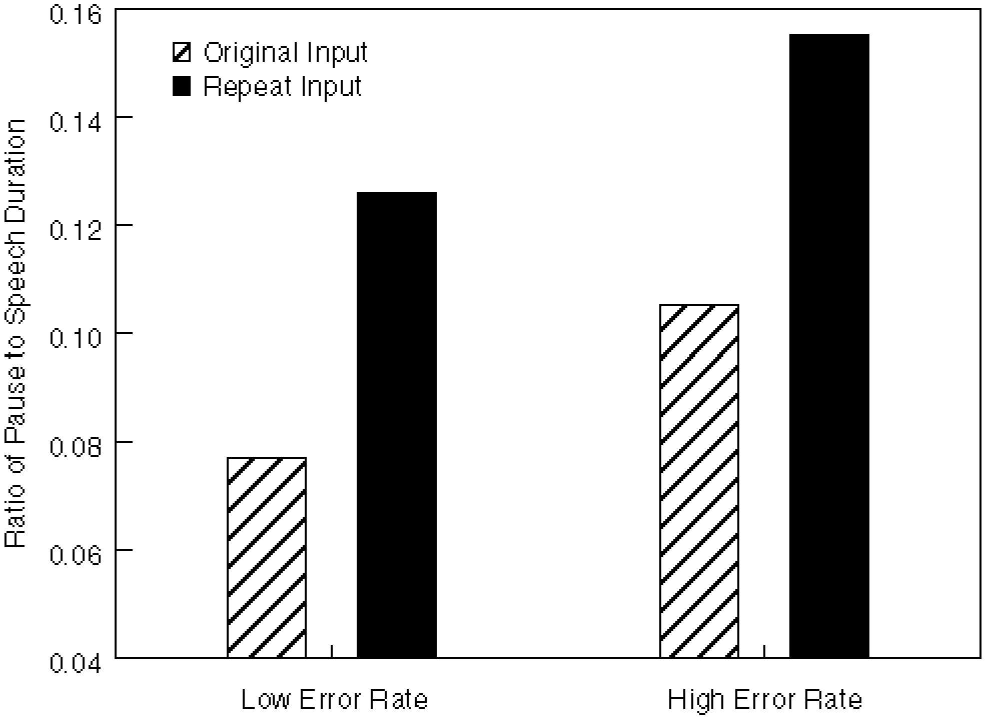


Figure 3: Ratio of total pause duration to speech segment duration in original and repeated speech.

Figure 3 illustrates the ratio of total pause duration to speech segment duration during original versus repeated input in both a low error rate (left) and a high one (right). During a low error rate, the pause length of original input was 8% that of speech, whereas during repeated input it increased to 13%. Likewise, during a high error rate, the pause length of original input was 11% that of speech, increasing to 16% during repetition. This figure clarifies that pause lengthening represented a greater proportional change during error resolution than did lengthening of speech.

Individual Differences in Pause Structure.

Since there were substantial durational effects during repetition, especially in the number and length of pauses, individual differences also were examined in these linguistic features. Analysis of between-subject differences indicated that average pause length during original input ranged from 22 to 497 msec. The average pause length during repetitions as well as average gain in pause length from original to repeated input were more variable— ranging from 222 to 1,008 msec and +1% to +1,232%, respectively. For a given user, the correlation between average pause length during their original input and that during repetitions likewise was not consistent, r = + .34 (N = 11), N.S. That is, pause duration was a highly variable linguistic feature of hyperarticulate speech both between and within users.

With respect to pause insertion, analysis of between-subject differences indicated that the average number of pauses per utterance during original input ranged from 0.08 to 1.00. The average number of pauses during repetitions as well as average gain in number of pauses from original to repeated input varied more substantially— from 0.17 to 3.41 and –50% to +525%, respectively. Despite this variability between users, there was a highly significant positive correlation between a given person's number of pauses during original input and those during their repetitions, r = + .74 (N = 16), p < .001. In fact, the strength of predictive association between the number of pauses during original and repeated input for the same person was $r2CU$ = 0.54. That is, 54% of the variance in the number of pauses inserted by a user during error handling could be predicted by knowing that user's pause rate during baseline spoken input.

These data indicate that change in pause structure during system error resolution is a variable linguistic phenomenon, especially between different users. However, individual users are consistent in the number of pauses they interject when repeating, compared with those during their baseline speech.

When speakers hyperarticulated during repetition, it is noteworthy that 56% of them exceeded the upper bounds of the range for number of pauses that were typical among users during original input. In addition, 33% surpassed the upper bound of the range for pause length that was typical during original input. That is, hyperarticulate adaptation of these pause phenomena very frequently did not fall within the normal variability that was observed among different users during baseline speech.

3.2 Rate of Speech

The rate of speech decreased significantly from an average of 298 to 348 msec per syllable during original and repeated input during a low base-rate of errors, paired t = 6.21 (df = 47), p < .001, one-tailed, and it also decreased significantly from an average of 300 to 347 msec per syllable under high base-rate conditions, t = 9.73 (df = 198), p < .001, one-tailed. Overall, speaking rate decreased an average of –16% during error resolution— which corresponds with elongation reported in average utterance duration.

3.3 Amplitude

The maximum intensity averaged 70.9 dB and 71.2 dB during original and repeat input when the error base-rate was low, and 71.2 dB and 71.0 dB when it was high. Paired t tests on original versus repeat speech revealed no significant change in intensity in either the low or high error conditions, paired t = 1.14 (NS) and t = 1.59 (NS), respectively.

Since errors were generated randomly and were not contingent on users' resolution strategies, it is possible that speakers could have initially altered their intensity but then abandoned this strategy as the session progressed. However, when intensity levels were re-examined for speech samples collected only during the beginning of the session, once again no significant amplitude change was found in either the low or high error-rate condition, t < 1 and t = 1.12 (NS), respectively.

3.4 Fundamental Frequency

Pitch Maximum. The maximum F0 did not differ significantly between original and repeat input in either the low or high error conditions, by paired t test on log transformed data, t = 1.58 (NS, one-tailed) and t = 1.35 (NS), respectively. Analyses (1) conducted only on the beginning of the session, (2) subdivided by gender, and (3) on normalized F0 data5 all confirmed that no significant change was evident in maximum F0 between original and repeated input for either the high or low error-rate conditions.

Pitch Minimum. The minimum F0 did not differ in the low error-rate condition, which averaged 111.3 hz and 110.4 hz, paired t < 1. However, minimum F0 dropped between original and repeat input in the high error-rate condition, averaging 122.2 hz and 119.5 hz, paired t test on log transformed data, t = 1.96, (df = 221), p < .05, two-tailed. This drop in minimum F0 represented a decline of just –2.2%.

Pitch Range. The F0 range did not differ significantly between original and repeat speech for either the low or high error conditions, t < 1. Reanalysis subdivided by gender revealed only that, when repeating their input during error resolution, female speech was significantly more expanded in pitch range when the error rate was high rather than low, paired t = 3.89 (df = 10), p < .0015, one-tailed. Male speech showed no expansion of F0 range under any condition.

Pitch Average. Average F0 dropped significantly between original and repeat input in the high error-rate condition, paired t = 2.83 (df = 206), p < .005, two-tailed, although no difference was found in the low error-rate condition. Once again, the decline in average F0 was a small one averaging just –1.4%.

3.5 Intonation Contour

The probability of *shifting* final intonation contour from rise to fall, or vice versa, averaged only 11% between original and repeated input. Wilcoxon Signed Ranks analysis confirmed that speakers were significantly more likely to hold their intonation the same between original input and first repetition than to change it, z = 3.88 (N = 20), p < .001, one-tailed. That is, whatever intonation contour originally was applied to the utterance tended to persist during verbatim correction.

Of the cases in which a change was observed in final intonation contour during repetition, 90% of the time the shift was from rising to falling, rather than the reverse. This difference was significant by Wilcoxon test, T+ = 71 (N = 12), p < .01, two-tailed. Overall, the likelihood of a final falling contour was 47% during original input, increasing to 56% during repetitions— for a relative increase in final falling contours of +19%.

3.6 Phonological Alternations

Approximately 9% of first repetitions in this corpus contained phonological alternations. In addition, 93% of those subjects who had sufficient data for analysis purposes were observed to alter their speech phonologically during error resolution at some point. Table 1 summarizes the number and type of alternations observed for each of the subjects who had a minimum of 12 scorable utterance pairs, as well as the classification of alternations occurring during a low error rate (left side) and a high error rate (right side) by direction of shift with respect to clear speech.

LOW ERROR RATE		HIGH ERROR RATE	
Clear to Conversational	Conversational to Clear	Clear to Conversational	Conversational to Clear
0	0	0	1 (a)
0	0	0	1 (c)
0	0	0	1 (a)
0	0	0	2 (b, c)
0	0	0	1 (a)
0	0	0	2 (c, c)
0	1 (b)	0	2 (b, c)
0	0	0	2 (b, b)
0	0	0	0
0	0	0	1 (a)
0	1 (c)	0	2 (c, c)
0	0	0	2 (b, c)
0	0	0	5 (b, b, c, c, c)
1 (d)	0	0	0
Total— 1	2	Total— 0	22

Table 1. Number and type of phonological alternations involving a shift from clear ® conversational versus conversational ® clear speech during low error rate (left) and high error rate (right), listed by subject

[Key: a—unreleased t > released t; b— alveolar flap > coronal plosive; c— nasal stop or flap > nt sequence; d— nt sequence > nasal stop or flap]

The majority of subjects, or 93% who had sufficient data and showed at least one spoken adaptation, shifted from a conversational to clear speech style, rather than the reverse. This was a significant difference by Wilcoxon Signed Ranks test, T+ = 87.5 (N= 13), p < .001, one-tailed. Followup analyses revealed that clear-speech adaptations were significantly more prevalent only during the high error-rate condition, Wilcoxon Signed Ranks test, T+ = 78 (N = 12), p < .001, one-tailed. There was no evidence that they increased significantly during the low error rate condition, as is evident in Table 1 (left side). Figure 5 also illustrates that the average rate of clear-speech adaptations per 100 words increased by a substantial +163% from the low error-rate condition (0.95) to the high one (2.50).

3.7 Relation between Phonological Alternations and Duration

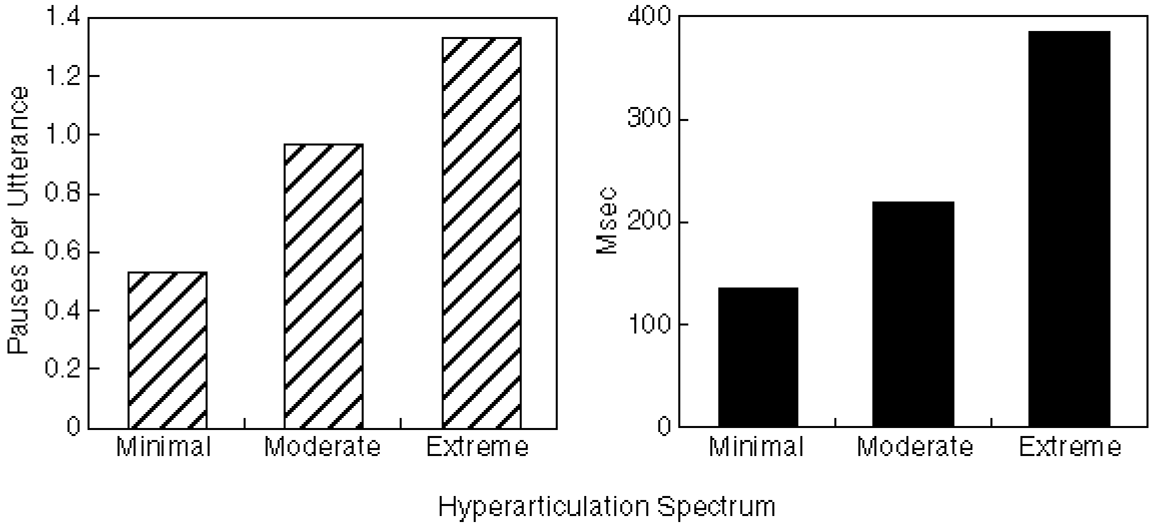


Figure 4. Total number of pauses (left) and total pause duration (right) during original input, repetitions without phonological change, and repetitions with audible phonological change.

Further analyses explored whether the phonological changes during error resolution were related to changes observed in pause phenomena at the utterance level, rather than constituting independent adaptation strategies. All original-repeat utterance pairs containing a conversational-to-clear phonological change were compared with utterances not containing a phonological change, but matched for speaker, lexical content, and error base-rate. It was revealed that repeated utterances with one or more phonological changes contained significantly more pauses than those without, T+ = 45 (N = 10), p < .05 (one-tailed), and significantly longer total pausing, T+ = 97 (N = 14), p < .002 (one-tailed).

Figure 4 also illustrates that when clear-speech phonological changes were present in repetitions, the number and length of pauses increased relatively more than when they were not. After factoring in corrections to control for speaker, lexical content, and error base-rate, Figure 4 (left panel) illustrates that the number of pauses per utterance averaged 0.53 during original input, 0.97 during repetitions without phonological alternations, and 1.33 during repetitions with audible phonological alternations. Figure 4 (right panel) also illustrates that the average pause duration in milliseconds increased incrementally from 136 during original input, to 220 during repetitions without phonological alternations, to 386 during repetitions with phonological alternations.

These data demonstrate that speakers are capable of systematically varying the degree to which they hyperarticulate along a graduated spectrum. Furthermore, the variability in their hyperarticulate speech can be adapted substantially during error handling with a computer—ranging from minimal to extreme.

These data also clarify that durational and phonological dimensions of change during hyperarticulate adaptation are not independent, but rather they co-occur and are related within individual utterances.

3.8 Disfluencies

The disfluency rate during original spoken input in this study averaged 0.78 disfluencies per 100 words, which replicates what has been reported previously in a structured human-computer interface (Oviatt, 1995). However, the disfluency rate dropped significantly to 0.37 when repeating speech during error resolution, paired t = 2.03 (df = 19), p < .03, one-tailed. Further analysis revealed that in the low error-rate condition, disfluencies averaged 0.85 per 100 words, which was similar to the overall rate of 0.78 during original input. However, the disfluency rate dropped significantly to 0.53 when the computer's error-rate was high, compared with the disfluency rate during a low base-rate of errors, paired t = 1.90 (df = 19), p < .04, one-tailed.

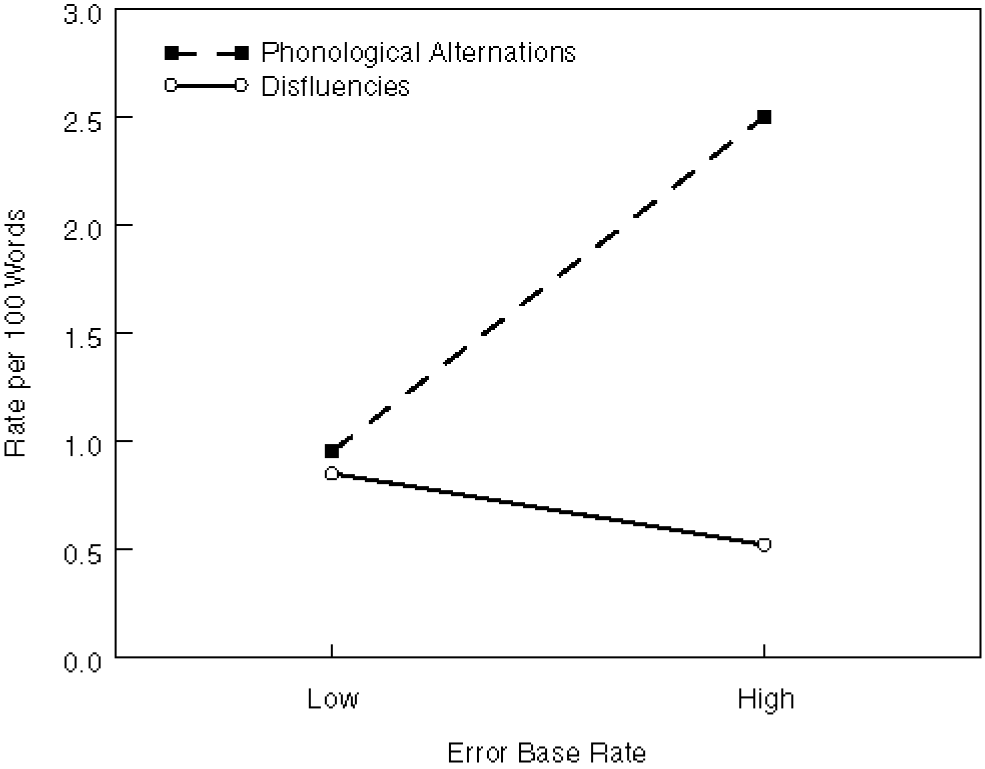


Figure 5. Rate of disfluencies and phonological alternations per 100 words as a function of error base-rate.

Followup analyses confirmed that disfluencies did not decrease significantly between original input and repetitions in the low error-rate condition, based on a Wilcoxon Signed Ranks test, T+ = 38 (N = 15), N.S. However, when the error rate was high the disfluency rate did drop significantly between original and repeated spoken input, T + = 72 (N = 13), p < .035, one-tailed. Figure 5 illustrates the inverse relation between increasing clear-speech phonological alternations and decreasing spoken disfluencies that occurred in the high error rate condition.

3.9 Self-Reported Perception of Recognition Errors

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 the delivery of errors was not contingent on their input in any way, people nonetheless believed strongly that they could influence the resolution of errors. Of the causal error theories expressed, 85% of users focused on linguistic characteristics of their own language. Another 10% said they had no idea why system errors occurred when they did, and 5% primarily gave a mechanical reason for the failure (e.g., "It needed a little time out, so I waited before entering the car preferences again" or "I think the zip code line was stuck, so I went back and reentered the state name before trying it again").

Of the large majority of participants who believed in a linguistic reason for failure, the following specific resolution strategies were most frequently stated as being effective: (a) speaking more slowly— mentioned by 53% of participants who maintained a linguistic theory, (b) pausing to separate words more— 29% of participants, (c) speaking more clearly— 24% of participants. This dominant sentiment is summarized in the following statement, "I just needed to speak more slowly and clearly, and to— you know— put more spaces between where the words were." Only a small minority of people who expressed a linguistic theory said they believed that speaking more loudly to the computer was effective in resolving errors (6%), or changing voice inflection (6%). In short, participants' self-reports regarding error repair strategies were consistent with the major changes observed in hyperarticulate speech.

3.10 Summary of Hyperarticulation Profile

Table 2 presents summarized results regarding the hyperarticulation profile during human-computer error resolution. The magnitude of relative change shown for each linguistic dimension is an average across low and high error-rate conditions. Table 2 clarifies that change in pause structure dominated hyperarticulate adaptation during error resolution, with durational increase in the speech segment also noteworthy. Articulatory changes were a second prominent characteristic of hyperarticulate adaptation, including both a drop in spoken disfluencies and an increase in hyper-clear phonological features. Finally, speakers shifted to a final falling intonation contour during repetitions, and this appeared associated with small decreases in fundamental frequency measures.

TABLE 2 – SUMMARY OF RELATIVE CHANGE IN LINGUISTIC DIMENSIONS OF HYPERARTICULATION[†]	
Pause interjection	+92%

Pause elongation	+75%
Disfluencies	-53%
Intonation–final fall	+19%
Speech elongation	+12%
Hyper-clear phonology phonology	+9%
Pitch minimum	-2%
Pitch average	-1%

[†] All magnitudes shown represent statistically significant change during repetition.

4. DISCUSSION

Human speech to computers varies along a spectrum of hyperarticulation, such that its basic signal properties change dynamically and sometimes abruptly. The present data demonstrate that the presence, form, and degree of hyperarticulation in users' speech to computers is a predictable phenomenon. That is, the speech signal is not simply "noisy" during real interactions, but rather is transformed in principled ways during human-computer exchange. When a system makes a recognition error, the miscommunication that occurs can be a particularly forceful elicitor of hyperarticulate speech from users. These research findings raise concerns regarding traditional algorithmic approaches to recognizing spoken language, which do not tend to model dynamic stylistic changes in the speech signal that are elicited during natural interactions— such as hyperarticulation during miscommunication, or Lombard speech during noise.

4.1 Hyperarticulate Speech to Computers

During system error resolution, human speech primarily shifted to become lengthier and more clearly articulated. An increase in utterance duration was evident during both low and high error-rate conditions— including +12% average elongation of the speech segment, +73% elongation of pause duration, and interjection of +91% more pauses. A corresponding decrease was evident in speech rate during repetitions. Clearly, the single most salient relative change in repeated speech was its altered pause structure. Essentially, users' speech became more discrete, departing from the pattern of continuous speech upon which current recognizers typically are trained.

Increased pause insertion and lengthening have been found in hyper-clear speech between humans (Cutler and Butterfield, 1990 and 1991), and such changes in pause structure play an important role in assisting the listener to mark word boundaries and segment a continuous stream of speech. For example, low-intelligibility speech can be rendered more understandable by inserting pauses at word boundaries (Maasen, 1986). The relatively large durational increases documented in this research are similar to those found in hyper-clear speech to the hearing impaired (Uchanski et al., 1996). Durational changes in speech to the hearing impaired also vary for different types of consonants and vowels. For example, there is relatively less elongation of short vowels and voiced plosives than is evident in fricatives and semi-vowels (Uchanski et al., 1996). Relative durational change in different classes of vowels and consonants, and their corresponding phonological features, is a topic that merits further exploration in human speech to interactive systems.

During a high base-rate of errors, the phonological features of repeat speech also adapted toward an audibly clearer articulation pattern, with the most frequently observed changes including fortition of alveolar flaps to coronal plosives (e.g., eIReIt changing to eIt|eIt), and shifts to unreduced **nt** sequences (e.g., twER,i to twEnti). In other words, speakers reduced **t** to the flapped **d** sound during original input (e.g., saying "fordy" for **40**), but then repeated it as an unreduced **t** during error correction (e.g., "forty"). Likewise, they omitted the **t** sound in **20** during original input (e.g., "tweny"), but clearly articulated the **t** following an error (e.g., "twenty"). Users' speech basically became more deliberate and well specified in its signal cues to phonetic identity. This shift toward hyper-clear speech during error resolution also corresponded with a drop in spoken disfluencies when the error-rate was high. The present findings are consistent with the literature on hyper-clear speech between humans, which has reported change in both vowel and consonant quality (Chen, 1980; Cutler and Butterfield, 1991; Moon, 1991; Picheny et al., 1986)— including more fully released word-final stops, in which the release provides clear information about both voicing and place of articulation (Malecot, 1958; Picheny et al., 1986).

When the base-rate of errors was high, both pitch minimum and pitch average dropped significantly during error correction. In addition, female pitch became more expanded in range when resolving errors during a chronically high error-rate. The small drop observed in fundamental frequency, which averaged only –2%, appeared related to speakers' tendency to adopt a final falling intonation contour during error correction. In the present corpus, continuation rises were prevalent during original input on list-like content such as addresses, but when an error occurred the speaker initiated an error correction subdialogue which more often was closed with a final falling contour. Both the increased rate of final falling tones and the small decline in overall pitch on error subdialogues apparently were used by speakers as cues to mark the close of an error repair with their computer partner. These findings are consistent with previous research indicating that a final falling contour and reduction in pitch are the strongest cues used to produce finality judgements during human-human speech (Swerts, Bouwhuis and Collier, 1994). However, it should be emphasized that the changes observed in fundamental frequency during error correction were small ones. This finding is consistent with the observation by Bruce and colleagues (1995) that the degree of prosodic variation in human-computer interaction is more attenuated than that typical of human interactive dialogues.

Perhaps counterintuitively, hyperarticulate speech to computers did not increase in amplitude. During post-experimental interviews, users likewise did not mention speaking more loudly as a potentially effective means of resolving errors with a computer. Since analyses of speech from the session's beginning reconfirmed this lack of significant change in amplitude and maximum pitch, the inactivity of these linguistic features cannot be attributed to the possibility that speakers initially altered them but then discontinued as the session progressed. These signal- and interview-level findings on unchanged amplitude during human-computer speech are in contrast with the amplitude increases often found in hyperarticulate speech between humans— for example, in speech to the hearing impaired, and in noisy environments. However, both human-computer and human-human hyper-clear speech are similar in that their general profiles involve larger durational adaptations than occur in either pitch or amplitude (Cutler and Butterfield, 1990 and 1991). In summary, adaptation of both amplitude and intonation appear to play a relatively minor role in speech to computers— compared with that to human listeners, and also compared with durational and phonological changes.

Since the delivery of simulated errors in this study was completely independent of users' spoken response to those errors, it is important to note that people were *not reinforced* in any way for responding as they did with certain speech adaptations. As a result, the signal changes reported can be viewed as representing a strong and persistent predilection on speakers' part, since their adaptations never directly led to error resolution. Although the delivery of errors was pre-programmed and not contingent on users' speech, users nonetheless reported believing that they had caused system errors, and also that they could resolve them by altering their speech delivery. In the future, systems may be designed that recover from errors in direct response to the altered signal characteristics reported here. Under these circumstances, the frequency and magnitude of speakers' hyperarticulate adaptations could be increased further, or become *entrained.*6 The extent to which hyperarticulate speech adaptations may be subject to entrainment has yet to be explored.

It recently has been demonstrated that the hyperarticulate adaptations reported in this study following failure-to-understand errors also occur in response to other qualitatively different recognition errors, such as substitutions (Oviatt, Levow, Moreton and MacEachern, in submission). These related findings demonstrate the replicability and generality of the hyperarticulate speech adaptations reported in the present research. To further examine the generality of the present results, it also would be interesting to explore cross-linguistic comparisons of the type and magnitude of hyperarticulate adaptations to computers. Additional research also is needed to pursue more detailed quantitative modeling of the major durational and articulatory phenomena identified in this study, as well as their interrelation.

4.2 Users' Model of the "At-Risk" Computer Listener

There is a sense in which people may view error-prone computers as a kind of "at-risk" listener. Compared with human-human speech during expected or actual miscommunication, however, the pattern of hyperarticulation to a computer is somewhat unique. For example, users did not alter their amplitude when resolving errors with the computer, and change in fundamental frequency was minimal. In this sense, the profile of adapted speech to a computer differs from that during interpersonal hyperarticulation, just as it varies in speech to distinct at-risk human populations.

Speakers not only are able to vary their speech along a spectrum of hyperarticulation to all of these "partners," they also adjust their signal characteristics to accommodate what they perceive to be specific obstacles in their listener's ability to extract lexical meaning. In speech to computers, for example, the major adaptations observed in durational and phonological features were consistent with users' self-report that they believed they needed to "speak more slowly and clearly" to the computer. Basically, users' model of the likely causes of system recognition failure in part focus on segmentation of the speech stream (i.e., related to durational effects and large changes in pause structure), as well as on clarity or "goodness" of phonemes (i.e., related to clearer articulation of particular sounds and decreased disfluencies). When interacting with a computer in a quiet office environment, basic audibility repairable with increased amplitude simply may not be perceived as a likely source of communication failure. Attentional lapse or deficit on the computer's part, which could be focused via pitch variation, may not be perceived as a likely source of computer failure either. That is, speakers may not believe that they need to work to attract and direct the computer's attention toward corrected lexical content, so pitch variation consequently may not be viewed as a useful repair strategy.

In future multimedia systems in which the system is presented as an animated partner, the characterization rendered of the system might influence the manner in which speakers' hyperarticulation is fine-tuned. For example, a talking head presented as a wizened elder might elicit amplitude gains, and a childlike waif might elicit greater pitch change. That is, if an animation can be rendered *believable* enough that it begins to influence the user's model of system fallibility, then the potential exists for manipulating the basic signal characteristics of users' speech. One topic that should be explored in future research is whether animated renderings can be used to manipulate users' speech to minimize hard-to-process forms of signal variability, or to guide the speech signal to match system processing capabilities.

4.3 The CHAM Model

When resolving errors with a computer, users actively tailored their speech along a spectrum of hyperarticulation. The graduated nature of users' speech adaptations was evident in increases in the number and length of pauses between: (1) *original input* — 0.53 pauses per utterance; 136 msec., (2) *repeated input without phonological alternations*— 0.97 pauses; 220 msec., and (3) *repeated input containing phonological alternations*— 1.33 pauses; 386 msec., as illustrated in Figure 4. These data establish that substantial durational effects can occur in the absence of any audible phonological change, or independently of such change. However, in more extreme hyperarticulate speech in which the two dimensions co-occur, durational effects are magnified further. This latter subgroup of utterance pairs provides evidence of interdependence between the durational and articulatory dimension of adaptation, although the exact nature of this relation is unclear and should be explored further. It is possible that durational changes in the speech signal may mediate and play a role in altering the expression of phonetic gestures, as in vowel formant transitions observed during "vowel undershoot" in hyper-clear speech (Moon and Lindblom, 1994).

The spectrum of hyperarticulate adaptations also can be viewed by comparing: (1) original input as a *baseline speech* value, (2) repeated speech during a *low error base-rate*, and (3) repeated speech during a *high error base-rate*. During both low and high error rates, durational changes were pervasive, including elongation of the speech segment and large relative increases in the number and duration of pauses. During a high error-rate, speech also was adapted to include more hyper-clear phonological features, reduced disfluencies, and modest changes in fundamental frequency. The generality of these empirical findings has been further corroborated by recent research that has replicated this hyperarticulation profile in connection with other types of recognition error (Oviatt, Levow, Moreton and MacEachern, in submission).

The two-stage branching **C**omputer-elicited **H**yperarticulate **A**daptation **M**odel (**CHAM**) is presented in Figure 6 to account for these systematic changes in speech during interactive error resolution. According to the CHAM model, *Stage I* adaptations entail a singular change in durational characteristics. This stage is associated with a moderate degree of hyperarticulation during a low error base-rate. *Stage II* entails multiple changes in durational, articulatory and fundamental frequency characteristics. This stage is associated with a more extreme degree of hyperarticulation during a high error base-rate.

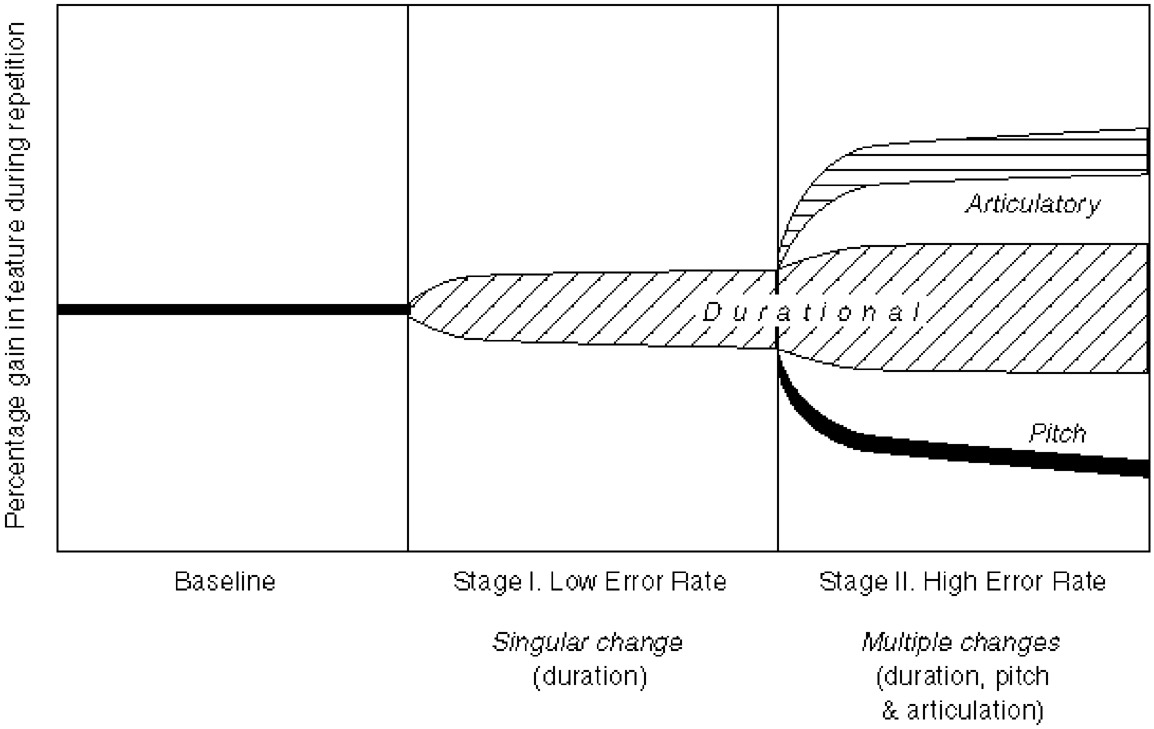


Figure 6: Computer-elicited hyperarticulate adaptation model (CHAM)

With respect to predictions, the CHAM model specifies that users' speech will adapt toward the linguistically-specified hyperarticulation profile outlined in this research. It also predicts that systems characterized by different error base-rates will elicit different degrees of hyperarticulation, as summarized in Stage I and Stage II of the model. Given that Baseline, Stage I, and Stage II speech occur in juxtaposition with one another during typical error resolution episodes, the CHAM model also predicts abrupt transitions in the signal profile of a given user's speech from one moment to the next during error handling. The implications of the CHAM model for designing interactive systems with improved error handling capabilities are summarized in the next section.

4.4 Designing Interactive Systems to Handle Hyperarticulation

The hyperarticulate speech documented in this research presents a potentially difficult source of variability that can degrade the performance of current speech recognizers, in particular complicating recognizers' ability to resolve errors gracefully. One question raised by viewing the CHAM model is whether an utterance delivered during baseline can be recognized as identical to its counterpart under Stage II conditions. Like Lombard speech elicited in variable noise conditions, this type of episodic and often abrupt signal variability may pose a more substantial challenge to current recognition technology than more chronic forms of variability, such as accented speech. The relatively static approaches that currently dominate the field of speech recognition, including techniques like Hidden Markov modeling, appear particularly ill suited to processing the dynamic stylistic variability typical of hyperarticulate speech. The present research therefore should be viewed as providing a stimulus for working toward the development of fundamentally more dynamic, adaptive, and user-centered approaches to speech recognition.

From a pragmatic viewpoint, there are several possible ways to improve the performance of current spoken language systems on hyperarticulate speech. The first is to train recognizers on more natural samples of users' interactive speech to systems, including error resolution with the type and base-rate of errors expected in the target system. This also could entail multistyle training to develop better reference models for words (Junqua, 1993; Lippmann et al., 1987). Such an approach would entail collecting a more heterogeneous training corpus as a basis for recognition. For low error-rate systems, a second alternative approach implied by the CHAM Model is adjustment of durational thresholds or models of phones (Mirghafori, Fosler & Morgan, 1996), since durational adaptation is the primary change that occurs during moderate hyperarticulation. These alternatives represent data collection and model adjustment efforts geared toward accommodating a wider range of signal variability, which may only function well during modest hyperarticulation. Such approaches may be associated with trade-offs in the accuracy of processing. Another drawback with these alternatives is that they are not specifically designed to handle abrupt transitions in signal characteristics.

A third approach is to design a recognizer specialized for error handling, which could function as part of a coordinated suite of multiple recognizers that are swapped in and out at appropriate points during system interaction. In the case of failure to understand errors, the system could swap in a specialized error correction recognizer after its prompt for user repetition. However, this would not be sufficient for handling substitution errors of which only the user is aware. A more general approach for handling all system recognition errors would be use of a form-based interface with content-specific input slots, as was designed for the present simulation study. In such an interface, it can reasonably be inferred that user re-entry into the same slot involves a correction, which then could be used to swap in the specialized recognizer. Building a successful recognizer of this kind would depend on the collection of hyperarticulate speech data and on quantitative modeling of hyperarticulation, especially of the durational and articulatory phenomena that constitute its central landmarks. The advantage of this approach is that it is capable of handling abrupt shifts in hyperarticulation, and in a manner tailored to a particular application. However, not all applications may be amenable to clear-cut identification of the start and end of error correction, such that swapping to the appropriate recognizer could be easily and reliably triggered. In such cases, a more computationally intensive fourth alternative simply could involve parallel processing with multiple recognizers representing the different durational models, with selection of the best match based on probability estimates at any given point during the human-computer dialogue.

A fifth approach to improving current recognizer performance is to develop adaptive systems that are designed to accommodate differences in a system's base-rate of errors as summarized in the CHAM model, as well as individual differences in users' hyperarticulation profile. Since signal adaptations occur abruptly when users enter an error resolution subdialogue, any system should *not be designed to adapt continuously* to users' speech throughout a human-computer interaction. Rather, system adaptation specifically should *avoid adaptation across sharp* *boundaries* that divide original input from error correction speech— instead adapting within error-correction subdialogues to the specific form and magnitude of a given user's hyperarticulation. The goal of such an approach would be to improve the recognizer's performance on lexical items encountered in future error correction episodes.

In this research, consistency was found in a given user's rate of pause interjection during error handling, which was predictable based on their pause rate during baseline spoken input. Since change in speakers' pause structure dominated other hyperarticulate adaptations, a system capable of adapting rapidly to a given user's level of pause interjection— and then predicting that user's pause interjections during repetitions— potentially could be more effective in recognizing the lexical content in their hyperarticulate speech during error resolution. That is, an adaptive approach may well be useful for successfully processing this type of pause change in hyperarticulated speech during system error handling. To better assess the benefits of an adaptive approach, future research should explore other individual differences in hyperarticulate speech, and also devise implementations for adjusting to them effectively.

In the case of Lombard speech, individual differences also can be substantial, especially for certain linguistic features associated with gender effects (Junqua, 1993). Perhaps counterintuitively, recognition rates on Lombard speech are worse for speaker-dependent mode than for speaker-independent recognition algorithms. This failure occurs because the pattern-matching algorithm for speaker-dependent recognition tries to match spectral characteristics too closely, essentially relying on the assumption of minimal intraspeaker variability (Junqua, 1993). As a result, speaker-dependent algorithms fail to model the speech variability typical of Lombard effects. Compared with the use of conventional discrete density hidden Markov model recognizers, application of the learning and classification properties of neural networks, combined with a focus on recognizing relational speech features (e.g., ratio of consonant to vowel duration, vowel formant transitions), has been demonstrated to result in improved recognition accuracy in the case of Lombard speech (Applebaum and Hanson, 1990).

A sixth solution to improving current recognizer performance is to *avoid* hyperarticulate speech by designing a multimodal rather than unimodal interface, an option that has been discussed in detail elsewhere (Oviatt and vanGent, 1996; Oviatt, Laniran, Bernard & Levow, in submission). When people are free to switch to an alternate input mode, the likelihood of both avoiding and rapidly resolving errors can be facilitated when interacting multimodally. This is partly because users have good intuitions about when to use a given input mode (Oviatt and Olsen, 1994). In addition, users actively alternate input modes after a recognition error occurs. Since different input modes (e.g., speech vs. pen input) have different confusion matrices associated with the same propositional content, switching input modes in a multimodal interface could eliminate stubborn spiral errors effectively. In addition, multimodal system architectures that unify the propositional content carried in parallel input modes can result in mutual disambiguation during semantic interpretation, and therefore a reduced error rate (Johnston et al., 1997; Oviatt, in submission). Finally, users report less frustration with system errors when they are able to function multimodally. This could increase their overall tolerance for errors, as well as their satisfaction when interacting with inherently error-prone recognition systems.

In the future, it will be important to collect samples of speech not only during realistic interactive exchanges in the lab, but also in natural field environments and while users are mobile. Due to variable noise levels, movement, collaborating groups of users, interruptions, multi-tasking, stress, and other factors, it is anticipated that acoustic-phonetic variability in the speech signal may be different and substantially magnified under such conditions. The speech encountered in these circumstances can be expected to include a combination of hyperarticulate, Lombard, and other challenging forms of abrupt signal variation. To prepare for designing next-generation field and mobile systems, both fundamental speech algorithms and spoken language interfaces will need to be capable of handling signal variation typical of the different environmental conditions and speaker-listener dynamics in these field settings— as well as the higher rates of miscommunication associated with them. The present research on user-centered modeling of speech adaptations during error begins to establish an empirical foundation for the successful design of this more challenging generation of spoken language and multimodal systems.

5. ACKNOWLEDGMENTS

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, and Karen Kuhn for transcription and scoring assistance. Thanks also to Karen Ward, Terri Lander, and Pieter Vermeulen of CSLU for helpful discussions and assistance with OGI’s Speech Tools. Finally, we are grateful to the people who generously volunteered their time to participate in this research.

6. REFERENCES

- T. H. Applebaum and B. A. Hanson (1990), "Robust speaker-independent word recognition using spectral smoothing and temporal derivatives", in: *Proc. of EUSIPCO-90*, pp. 1183-1186.

- *ARPA Workshop Proc. on Human Language Technology*, C. Weinstein ed., (1994), 8-11 March 1994, (Morgan Kaufmann Publishers, San Mateo).

- Z. S. Bond, and T. J. Moore (1994), "A note on the acoustic-phonetic characteristics of inadvertently clear speech", *Speech Communication*, Vol. 14, pp. 325-337.

- M. Brenner, T. Shipp, E. Doherty, and P. Morrissey (1985), "Voice measures of psychological stress: Laboratory and field data", in: I. Titze and R. Scherer, eds., *Vocal Fold Physiology, Biomechanics, Acoustics, and Phonatory Control*, (Denver Center for the Performing Arts, Denver), pp. 239-248.

- G. Bruce, B. Granstrom, K. Gustafson, M. Horne, D. House, and P. Touati (1995), "Towards an enhanced prosodic model adapted to dialogue applications", in: *Proc. of the ESCA Workshop on Spoken Dialogue Systems, Vigso, Denmark*, pp. 201-204.

- A. Caramazza, and A. E. Hillis (1991), "Lexical organization of nouns and verbs in the brain", *Nature*, Vol. 349, pp. 788-790.

- F. R. Chen (1980), "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level", Master’s thesis, MIT.

- J. R. Cohen (1996), "The summers of our discontent", in: T. Bunnell and W. Idsardi, eds., *Proc. of the International Conference on Spoken Language Processing, Philadelphia, PA.*, 3-6 October 1996 (University of Delaware and A.I. duPont Instit.), Addendum, pp. 9-10.

- A. Cutler and S. Butterfield (1990), "Durational cues to word boundaries in clear speech", *Speech Communication,* Vol. 9, pp. 485-495.

- A. Cutler and S. Butterfield (1991) "Word boundary cues in clear speech: A supplementary report", *Speech Communication,* Vol. 10, pp. 335-353.

- C. Danis (1989), "Developing successful speakers for an automatic speech recognition system", in: *Proc. of the Human Factors Society 33rd Annual Meeting,* pp. 301-304.

- B. Eisen, H. G. Tillmann, and C. Draxler (1992), "Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases", in: J. Ohala et al., eds., *Proc. of the International Conference on Spoken Language Processing, Banff, Alberta,* 12-16 October 1992 (University of Alberta, Alberta), Vol. 2, pp. 871-874.

- C. A. Ferguson (1975), "Toward a characterization of English foreigner talk", *Anthropological Linguistics,* Vol. 17, No. 1, pp. 1-14.

- C. A. Ferguson (1977), "Baby talk as a simplified register", in: C. E. Snow and C. A. Ferguson, eds., *Talking to Children: Language Input and Acquisition,* (Cambridge University Press, Cambridge), pp. 219-36.

- A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. De Boysson-Bardies, and I. Fukui (1989), "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", *J. of Child Language,* Vol. 16, pp. 477-501.

- C. Frankish, R. Hull, and P. Morgan (1995), "Recognition accuracy and user acceptance of pen interfaces", in: *Proc. of the Conference on Human Factors in Computing Systems, Denver (CHI'95)* , 7-11 May 1995 (ACM Press, New York), pp. 503-510.

- B. F. Freed (1978), "Foreign talk: A study of speech adjustments made by native speakers of english in conversation with non-native speakers", Doctoral Dissertation, Linguistics Department, University of Pennsylvania.

- C. Gagnoulet (1989), "Voice replaces dial in new public phones", *International Voice Systems Review,* Vol. 1, No. 1.

- O. K. Garnica (1977), "Some prosodic and paralinguistic features of speech to young children", in: C. E. Snow and C. A. Ferguson, eds., *Talking to children,* (Cambridge University Press, Cambridge), pp. 63-88.

- S. Gordon-Salant (1987), "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects", *J. of the Acoustical Society of America,* Vol. 81, pp. 1199-1202.

- T. D. Hanley and M. D. Steer (1949), "Effect of level of distracting noise upon speaking rate, duration and intensity", *J. of Speech and Hearing Disorders,* Vol. 14, pp. 363-368.

- F. Jelinek (1985), "The development of an experimental discrete dictation recognizer", *Proc. of the IEEE,* Vol. 73, No. 11, pp. 1616-1624.

- M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman and I. Smith (1997), "Unification-based multimodal integration", in: *Proc. of the Association for Computational Linguistics Conference, Madrid, Spain,* 8-10 July 1997 (ACL Press, Madrid), pp. 281-288.

- J. C. Junqua (1993), "The lombard reflex and its role on human listeners and automatic speech recognizers", *J. of the Acoustical Society of America,* Vol. 93, No. 1, pp. 510-524.

- C. A. Kamm (1994), "User interfaces for voice applications", in: D. B. Roe and J. Wilpon, eds., *Voice Communication Between Humans and Machines* (National Academy Press, Washington, D.C.), pp. 422-442.

- D. Karis and K. M. Dobroth (1991), "Automating services with speech recognition over the public switched telephone network: Human factors considerations", *IEEE J. of Selected Areas in Communications,* Vol. 9, No. 4, pp. 574-585.

- C. Lewis and D. A. Norman (1986), "Designing for error", in: D. A. Norman and S. W. Draper, eds., *User-Centered System Design* (Lawrence Erlbaum, Hillsdale, N. J.), pp. 411-432.

- B. Lindblom (1990), "Explaining phonetic variation: A sketch of the H and H theory", in: W. Hardcastle and A. Marchal, eds., *Speech Production and Speech Modeling* (Kluwe, Dordrecht), pp. 403-439.

- B. Lindblom (1996), "Role of articulation in speech perception: Clues from production", *J. of the Acoustic Society of America,* Vol. 99, No. 3, pp. 1683-1692.

- B. Lindblom, S. Brownlee, B. Davis and S. J. Moon (1992), "Speech transforms", *Speech Communication*, Vol. 11, Nos. 4-5, pp. 357-368.

- R. P. Lippmann, E. A. Martin and D. B. Paul (1987), "Multi-style training for robust isolated-word speech recognition", in: *Proc. of ICASSP-87*, pp. 705-708.

- S. E. Lively, D. B. Pisoni, W. Van Summers and R. Bernacki (1993), "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences", *J. of the Acoustical Society of America,* Vol. 93, No. 5, pp. 2962-2973.

- E. Lombard (1911), "Le signe de l’elevation de la voix", *Annals Maladiers Oreille, Larynx, Nez, Pharynx*, Vol. 37, pp. 101-119.

- B. Maasen (1986), "Marking word boundaries to improve the intelligibility of the speech of the deaf", *J. of Speech and Hearing Research,* Vol. 29, pp. 227-230.

- A. Malecot (1958), "The role of releases in the identification of released final stops", *Language,* Vol. 34, pp. 370-380.

- N. Mirghafori, E. Fosler and N. Morgan (1996), "Towards robustness to fast speech in ASR", in: *Proc. of ICASSP-96*, vol. 1, 335-338.

- S. J. Moon (1991), "An acoustic and perceptual study of undershoot in clear and citation-form speech", Doctoral Dissertation, Linguistics Department, University of Texas at Austin.

- S. J. Moon and B. Lindblom (1994), "Interaction between duration, context, and speaking style in English stressed vowels", *J. of the Acoustical Society of America*, Vol. 96, No. 1, pp. 40-55.

- C. Nass, J. Steuer, and E. R. Tauber (1994), "Computers are social actors", in: *Proc. of the Conference on Human Factors in Computing Systems, Boston, MA (CHI '94*), 24-28 April 1994 (ACM Press, Boston), pp. 72-78.

- S. L. Oviatt (1995), "Predicting spoken disfluencies during human-computer interaction", *Computer Speech and Language*, Vol. 9, No. 1, pp. 19-35.

- S. L. Oviatt (in submission), "Ten myths of multimodal interaction".

- S. L. Oviatt, P. R. Cohen, M. W. Fong and M. P. Frank (1992), "A rapid semi-automatic simulation technique for investigating interactive speech and handwriting", in: J. Ohala et al., eds., *Proc. of the International Conference on Spoken Language Processing, Banff, Alberta,* 12-16 October 1992 (University of Alberta, Banff, Alberta), Vol. 2, pp. 1351-1354.

- S. L. Oviatt, P. R. Cohen and M. Q. Wang (1994), "Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity", *Speech Communication*, Vol. 15, Nos. 3-4, pp. 283-300.

- S. L. Oviatt, Y. Laniran, J. Bernard & G. Levow (in submission), "Linguistic adaptations during spoken and multimodal error resolution".

- S. L. Oviatt, G. Levow, E. Moreton and M. MacEachern (in submission), "Modeling global and focal hyperarticulation during human-computer error resolution".

- S. L. Oviatt and E. Olsen (1994), "Integration themes in multimodal human-computer interaction", in Shirai, Furui, and Kakehi, eds., *Proceedings of the International Conference on Spoken Language Processing*, Kobe, Japan, 18-22 September 1994 (Acoustical Society of Japan,) Vol. 2, pp. 551-554.

- S. L. Oviatt and R. VanGent (1996), "Error resolution during multimodal human-computer interaction", in: T. Bunnell and W. Idsardi, eds., *Proc. of the International Conference on Spoken Language Processing*, Philadelphia, PA., 3-6 October 1996 (University of Delaware and A.I. duPont Instit.), Vol. 1, pp. 204-207.

- K. L Payton, R. M. Uchanski and L. D. Braida (1994), "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing", *J. of the Acoustical Society of America*, Vol. 95, pp. 1581- 1592.

- M. A. Picheny, N. I. Durlach and L. D. Braida (1985), "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech", *J. of Speech and Hearing Research,* Vol. 28, pp. 96-103.

- M. A. Picheny, N. I. Durlach and L. D. Braida (1986), "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech", *J. of Speech and Hearing Research,* Vol. 29, pp. 434-446.

- J. M. Pickett (1956), "Effects of vocal force on the intelligibility of speech sounds", *J. of the Acoustical Society of America,* Vol. 28, pp. 902-905.

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

- D. B. Roe and J. Wilpon, eds. (1994), *Voice communication between humans and machines,* (National Academy Press, Washington, D. C.).

- R. Schulman (1989), "Articulatory dynamics of loud and normal speech", *J. of the Acoustical Society of America,* Vol. 85, pp. 295-312.

- M. Shatz and R. Gelman (1973), "The development of communication skill modifications in the speech of young children as a function of listener", *Monographs of the Society of Research on Child Development,* Vol. 38, pp. 1-37.

- E. Shriberg, E. Wade and P. Price (1992), "Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction", in: *Proc. of the DARPA Speech and Natural Language Workshop, San Mateo, CA,* 23-26 February 1992 (Morgan Kaufmann Publishers, San Mateo), pp. 49-54.

- J. Spitz (1991), "Collection and analysis of data from real users: Implications for speech recognition/understanding systems", in: *Proc. of the 4th Darpa Workshop on Speech and Natural Language, Asilomar, CA,* 19-22 February 1991 (Morgan Kaufmann Publishers, San Mateo).

- W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow and M. A. Stokes (1988), "Effects of noise on speech production: Acoustic and perceptual analyses", *J. of the Acoustical Society of America,* Vol. 84, pp. 917-28.

- M. Swerts, D. G. Bouwhuis and R. Collier (1994), "Melodic cues to the perceived finality of utterances", *J. of the Acoustical Sciety of America,* Vol. 96, No. 4, pp. 2064-2075.

- E. J. Tolkmitt and K. R. Scherer (1986), "Effect of experimentally induced stress on vocal parameters", *J. of Experimental Psychology,* Vol. 12, pp. 302-312.

- R.M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, N. I. Durlach (1996), "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate", *J. of Speech and Hearing Research,* Vol. 39, pp. 494-509.

- C. E. Williams and K. N. Stevens (1969), "On determining the emotional state of pilots during flight: An exploratory study", *Aerospace Medicine*, Vol. 40, pp. 1369-1372.

- C. G. Wolf (1990), "Understanding handwriting recognition from the user's perspective", in: *Proc. of the Human Factors Society 34th Annual Meeting*, pp. 249-253.

- N. Yankelovitch, G. Levow and M. Marx (1995), "Designing SpeechActs: Issues in speech user interfaces", in: *Proc. of the Conference on Human Factors in Computing Systems (CHI'95), Denver, Colorado,* 7-11 May 1995 (ACM Press, New York), pp. 369-376.

FOOTNOTES

(1) This research was supported by Grant No. IRI-9530666 from the National Science Foundation, and

by and grants, contracts, and equipment donations from Apple, GTE Labs, Intel, Microsoft, NTT Data,

Southwestern Bell, and US West.

(2) First author: Center for Human-Computer Communication, Department of Computer Science,

Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, OR, 97291

(oviatt@cse.ogi.edu; http://www.cse.ogi.edu/~oviatt/) Collaborators' respective affiliations: Linguistics

Dept., UCLA (now at Linguistics Dept., University of Pittsburgh); Laboratory for Computer Science,

MIT.

(3) The remaining simulated errors involved either written input or a spoken correction that differed in lexical content from the user's original input, neither of which were eligible for inclusion in the present analysis.

(4) In common terms, speakers may reduce **t** to the flapped **d** sound during conversational speech (e.g., saying "fordy" for **40**), but speak it as an unreduced **t** during clear speech (e.g., "forty"). Another example may involve omitting the **t** sound in **20** during conversational speech in a relaxed style (e.g., "tweny"), but articulating the **t** during clear speech (e.g., "twenty").

(5) All F0 data in this study routinely were reanalyzed by gender. Fundamental frequency and

amplitude measures also were subjected to data normalization, although this did not yield any difference in the experimental results.

(6) Note that entrainment refers to further change in some behavior that a person already varies naturally

in a given context, and that therefore falls within the limits of constraints on learnability.

[1] This research was supported by Grant No. IRI-9530666 from the National Science Foundation, and by and grants, contracts, and equipment donations from Apple, GTE Labs, Intel, Microsoft, NTT Data, Southwestern Bell, and US West.

[2] First author: Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, OR, 97291 (oviatt@cse.ogi.edu; http://www.cse.ogi.edu/~oviatt/) Collaborators' respective affiliations: Linguistics Dept., UCLA (now at Linguistics Dept., University of Pittsburgh); Artificial Intelligence Laboratory, MIT.

[3] The remaining simulated errors involved either written input or a spoken correction that differed in lexical content from the user's original input, neither of which were eligible for inclusion in the present analysis.

[4] In common terms, speakers may reduce **t** to the flapped **d** sound during conversational speech (e.g., saying "fordy" for **40**), but speak it as an unreduced **t** during clear speech (e.g., "forty"). Another example may involve omitting the **t** sound in **20** during conversational speech in a relaxed style (e.g., "tweny"), but articulating the **t** during clear speech (e.g., "twenty").

[5] All F0 data in this study routinely were reanalyzed by gender. Fundamental frequency and amplitude measures also were subjected to data normalization, although this did not yield any difference in the experimental results.

[6]

Note that entrainment refers to further change in some behavior that a person already varies naturally in a given context, and that therefore falls within the limits of constraints on learnability.