Modeling global and focal hyperarticulation during human–computer error resolution

Sharon Oviatt,[a] Gina-Anne Levow,[b] Elliott Moreton,[c] and Margaret MacEachern[d]
Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, Oregon 97291

(Received 18 July 1997; accepted for publication 11 August 1998)

When resolving errors with interactive systems, people sometimes *hyperarticulate*—or adopt a clarified style of speech that has been associated with increased recognition errors. The primary goals of the present study were: (1) to provide a comprehensive analysis of acoustic, prosodic, and phonological adaptations to speech during human–computer error resolution after different types of recognition error; and (2) to examine changes in speech during both *global* and *focal* utterance repairs. A semi-automatic simulation method with a novel error-generation capability was used to compare speech immediately before and after system recognition errors. Matched original-repeat utterance pairs then were analyzed for type and magnitude of linguistic adaptation during global and focal repairs. Results indicated that the primary hyperarticulate changes in speech following all error types were durational, with increases in number and length of pauses most noteworthy. Speech also was adapted toward a more deliberate and hyperclear articulatory style. During focal error repairs, large durational effects functioned together with pitch and amplitude to provide selective prominence marking of the repair region. These results corroborate and generalize the computer-elicited hyperarticulate adaptation model (CHAM). Implications are discussed for improved error handling in next-generation spoken language and multimodal systems. © *1998 Acoustical Society of America.* [S0001-4966(98)04511-1]

PACS numbers: 43.72.Kb, 43.70.Fq [JLH]

INTRODUCTION

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 *et al.*, 1995; Rhyne and Wolf, 1993). As a result, future spoken language systems will need to be designed to handle recognition errors effectively if they are to perform in a reliable manner and succeed commercially. Although ''designing for error'' has been advocated widely for conventional interfaces (Lewis and Norman, 1986), to date this concept has not been applied effectively to the design of recognition-based technology.

A. Hyperarticulation and the cycle of recognition failure

When speaking to interactive systems, recent research has demonstrated that people typically adapt their language during attempts to resolve system recognition errors (Oviatt *et al.*, 1996, 1998). This change in speaking style toward *hyperarticulate speech* involves a stylized and clarified form of pronunciation that speakers routinely use when accommodating what they perceive to be ''at risk'' listeners, adverse communication environments, or interactions involving miscommunication (Lindblom *et al.*, 1992; Oviatt *et al.*, 1998). Unfortunately, hyperarticulate speech introduces difficult sources of variability into the task of spoken language processing, which has been associated with elevated rates of system recognition failure (Shriberg *et al.*, 1992).

When people hyperarticulate to spoken language systems in an effort to correct recognition errors, recognition rates would be expected to degrade as hyperarticulated speech departs from the 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 current recognition algorithms typically model phonemes and coarticulation effects, they do not tend to model dynamic stylistic changes in the speech signal that are elicited by environmental factors, such as the hyperarticulate speech adaptations that speakers make during miscommunication, or the ''Lombard speech'' adaptations that occur in a noisy environment (Lombard, 1911). With respect to training, current speech recognizers tend to be 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, the signal variability posed by hyperarticulate speech represents a hard-to-process source of variability that threatens to degrade recognizer performance. Since hyperarticulate speech can be both a *reaction* to system recognition failure, and a potential *fuel* for precipitating a higher error rate, the net effect is that it has the potential to generate a *cycle of recognition failure*.

[a] Electronic mail: oviatt@cse.ogi.edu;http://www.cse.ogi.edu/~oviatt/
[b] Currently at Artificial Intelligence Laboratory, MIT, Boston, MA.
[c] Currently at Linguistics Department, University of Massachusetts, Amherst, MA.
[d] Currently at Linguistics Department, University of Pittsburgh, Pittsburgh, PA.

The design of recognition technology also can contribute to this cycle of recognition failure, and to *clustering* of recognition errors. For example, the design of Hidden Markov Models can propagate recognition errors, since a misrecognized word can cause others in its vicinity to be misrecognized (Rhyne and Wolf, 1993). Language models based on conditional probabilities also can propagate recognition errors, because an error can force the language model into an incorrect state and increase the likelihood of an error on subsequent words (Jelinek, 1985). In short, once a recognition error has occurred, both the properties of spoken language technology and users' reactive hyperarticulation can lead to perpetuation of the error in a way that complicates graceful recovery.

To design for both avoidance and resolution of errors, one research strategy is to analyze human–computer interaction specifically during system recognition errors. Such work could include modeling of users' hyperarticulated speech during interactive error handling, and the design of spoken language interfaces that aim to manage these strongly engrained speech patterns.

B. The CHAM model

Human speech to computers varies along a spectrum of hyperarticulation, such that its basic signal properties change dynamically and sometimes abruptly (Oviatt *et al.*, 1996, 1998). When a system makes a recognition error, the miscommunication that occurs can be a particularly forceful elicitor of hyperarticulate speech from users. Furthermore, the presence, form, and degree of hyperarticulation in users' speech to computers is a predictable phenomenon, which is transformed in principled ways during human–computer interaction. Compared with speech to a human partner during expected or actual miscommunication, users' hyperarticulate speech to a computer is in some ways unique, and the pattern of adaptation is consistent with their perception of the computer as a kind of ''at risk'' listener (Oviatt *et al.*, 1998).

During system error resolution, speech primarily shifts to become lengthier and more clearly articulated. In recent research, uniform increases in utterance duration were demonstrated during both low and high error-rate conditions (i.e., 6.5% versus 20% rate of utterances containing an error), with no significant difference in elongation between conditions. On average, a +12% relative increase was found in clongation of speech during error repair, whereas +92% more pauses were interjected, and the relative increase in pause duration was +75% (Oviatt *et al.*, 1996, 1998). That is, the most salient change in speech during error handling was alteration of pause structure.

During a high error rate, the phonological features of repeated speech also adapt toward an audibly clearer articulation pattern, with frequent changes including fortition of alveolar flaps to coronal plosives, such as eɪɾeɪt changing to eɪt⌉eɪt, and shifts to unreduced **nt** sequences, such as twɛ̃ɾi changing to twɛnti (Oviatt *et al.*, 1996, 1998). Users' speech basically becomes more deliberate and well specified in its signal cues to phonetic identity. This shift toward hyperclear speech has also been shown to correspond with a drop in spoken disfluencies during a high error rate (Oviatt *et al.*, 1996, 1998). In contrast, during a low system error rate neither of these articulatory phenomena were observed to change significantly.

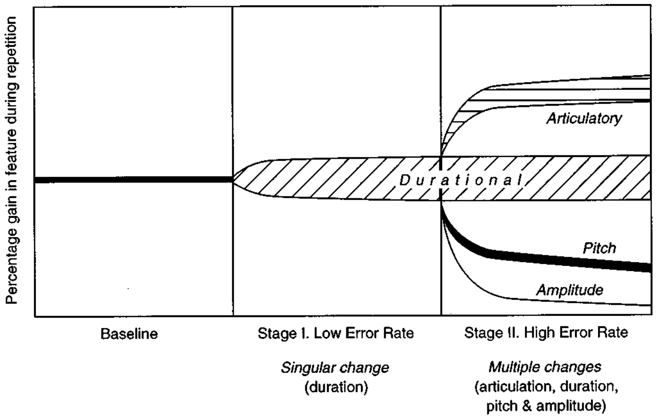


FIG. 1. Computer-elicited hyperarticulate adaptation model (CHAM).

With respect to intonation and fundamental frequency, during a high error rate, speakers also adopt a final falling intonation contour when repairing error subdialogues. This shift in intonation is also related to a slight decrease in fundamental frequency, which is reflected as a −2% average drop in minimum pitch (Oviatt *et al.*, 1998). Basically, speakers use final falling tones and a drop in pitch as cues to mark their repair subdialogues during human–computer dialogue interaction. However, neither of these changes are evident during a low system error rate, nor have reliable increases been reported in previous work in maximum pitch, pitch range, or amplitude (Oviatt *et al.*, 1996, 1998).

The two-stage branching computer-elicited hyperarticulate adaptation model (CHAM), illustrated in Fig. 1 and originally introduced in Oviatt *et al.* (1998), has been proposed as a unifying framework to account for these systematic changes in users' speech during interactive error handling. According to the empirically derived 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 rate of system errors. *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 rate of system errors. The two-stage CHAM model basically summarizes an unfolding of hyperarticulate speech adaptations that is consistent with the literature outlined above. In brief, it predicts that: (1) users' speech will adapt toward the linguistically specified hyperarticulation profile discussed above, including the type and magnitude of changes in articulatory, durational, and fundamental frequency features that has been outlined in previous empirical findings; (2) systems characterized by a low versus high error rate will elicit different types of hyperarticulate linguistic features, as illustrated in Stage I and II of the model shown in Fig. 1; and (3) abrupt rather than gradual transitions will occur in the signal profile of users' speech from one moment to the next as they begin and end episodes of error handling. Implications of the CHAM model for designing interactive systems with improved error handling have been discussed elsewhere (Oviatt *et al.*, 1998).

C. The hyperarticulation spectrum: When and why speech is adapted

Based on experimental phonetics data involving interpersonal speech, Lindblom and colleagues have argued that speakers make a moment-by-moment assessment of their listener's need for explicit signal information, and they adapt their speech production to the perceived needs of their listener in a given communicative context (Lindblom, 1990, 1996; Lindblom *et al.*, 1992). According to Lindblom's H & H theory, this adaptation varies actively along a continuum from *hypo- to hyperclear speech*. Hypoclear speech is relatively relaxed, and contains phonological reductions. A hypoclear speech style involves 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, hyperclear articulation is a clarified style that requires more speaker effort in order to achieve ideal target values for the acoustic form of vowels and consonants, thereby relying less on listener knowledge. Essentially, Lindblom and colleagues maintain that the relation between the speech signal and intended phonemes is a highly variable one, which is neither captured entirely by mapping phonemes to physical acoustic or phonetic characterizations, nor by factoring in local coarticulation effects (Lindblom, 1996). During human interaction, speaking style also can range from hypo- to hyperclear in a way that contributes substantial variability to the speech signal.

Lindblom and colleagues believe that speakers operate on the principle of supplying *sufficient discriminatory information* for a listener to comprehend their intended meaning, while at the same time striving for articulatory economy. When a speaker perceives no particular threat to their listener's ability to comprehend them, articulatory effort typically is relaxed (Lindblom, 1996). The result is hypoclear speech, which represents the default speaking style. 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 hyperclear to deliver more explicit signal information. In this sense, phonetic signals are dynamically modulated by the speaker to complement their listener's perceived speech processing ability and world knowledge. The effect of these speaker adaptations is to assist the listener in identifying a signal's intended lexical content.

In accord with these theoretical notions, there is evidence from a variety of studies that adaptation toward hyperarticulate speech does improve intelligibility by both normal and impaired listeners (Bond and Moore, 1994; Chen, 1980; Cutler and Buttertield, 1990; Gordon-Salant, 1987; Lively *et al.*, 1993; Moon, 1991; Payton *et al.*, 1994; Picheny *et al.*, 1985; Uchanski *et al.*, 1996). There is also linguistic and psychological literature indicating that people routinely adapt their speech during interpersonal exchanges when they expect or experience a comprehension failure from their listener. For example, modifications have been documented in parents' speech to infants and young children (Ferguson, 1977; Fernald *et al.*, 1989; Garnica, 1977), in speech to the hearing impaired (Picheny *et al.*, 1986), and in speech to nonnative speakers (Ferguson, 1975; Freed, 1978). Systematic changes also have been observed in speech during noise (Hanley and Steer, 1949; Junqua, 1993; Schulman, 1989; Summers *et al.*, 1988), during heavy workload or in stressful environments (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, 1991; Moon, 1991; Moon and Lindblom, 1994).

The specific hyperarticulate adaptations observed in these cases have differed depending on the target population and communicative context. For example, speech adaptations to infants often include elevated pitch, expanded pitch range, and stress on new vocabulary content—features that assist in gaining and maintaining infants' attention and in subserving teaching functions (Ferguson, 1977; Fernald *et al.*, 1989; Garnica, 1977). With hearing-impaired individuals, speech reportedly is higher in amplitude and fundamental frequency, longer in duration, and contains hyperclear phonological features (Picheny *et al.*, 1986). Speech adaptation in a noisy environment, characterized by the ''Lombard effect'' (Lombard, 1911), involves an increase in vocal effort that manifests itself as more than simple amplification of the speech signal. Among other features, it includes change in articulation of consonants, and increased duration and pitch of vowels (Junqua, 1993; Schulman, 1989). Lombard speech is analogous to hyperarticulate speech in the abruptness of signal change that often occurs. That is, Lombard and hyperarticulate speech both are characterized by episodic signal variability, which is a more challenging form of variability for recognizers to process than continuous signal deformation, as in the accented speech of a nonnative speaker.

To summarize, the interpersonal dynamics associated with different populations and circumstances clearly vary, even though all of them can be viewed as high risk communications. While they share features in common, the acoustic-prosodic and phonological features observed in these different cases nonetheless are defined by distinct hyperarticulation profiles. Recent research has begun to outline users' beliefs about the cause of communication failure as well as effective repair strategies when interacting with a computer (Oviatt *et al.*, 1998). Due to the error-prone nature of current recognition systems, speakers likewise may view the computer as a kind of ''at risk'' listener.

D. The concept of focal hyperarticulation

Recent research on hyperarticulate speech during human–computer error resolution has presented an analysis based on failure-to-understand errors, in which the system indicates its inability to recognize what the speaker said (Oviatt *et al.*, 1996, 1998). However, substitution errors constitute the majority of speech recognition errors (Brown and Vosburgh, 1989). During substitution errors, the system misrecognizes the user's speech and substitutes wrong lexical content. During some substitution errors, the speaker may not need to make a global repair of the entire utterance, but rather may selectively repair one focal part it—as in ''July twenty-**first**, nineteen ninety-seven.'' There currently is a lack of research on how speakers adapt their speech to a computer when making focal repairs, or whether these adaptations share hyperarticulate features in common with those observed during global repairs. If both focal and global utterance repairs involve similar hyperarticulate change to the speech signal, then focal repairs may be viewed as a brief and highly selective form of hyperarticulate adaptation, one in which signal transition is particularly abrupt.

Although speech adaptations during focal error repairs with a computer are poorly understood, in linguistic theory the concept of stress is relevant. *Stress* involves assignment of prosodic prominence to one element or part of an utterance, and it can occur during interpersonal communication when an error is repaired in part of an utterance. Stress has several known acoustic and phonological correlates, including increased pitch, increased amplitude, longer duration, and greater differentiation of vowel formant structure (de Jong, 1995; Fry, 1955, 1958). The acoustic-phonetic features of linguistic stress are believed to enhance the overall prominence and perceptual clarity of the stressed region, which in the case of an error must serve as the critical repair region. Stress sometimes has been described as involving assignment of a pitch accent (Bolinger, 1958; Fry, 1958), or as a local shift toward hyperarticulate speech with greater phonemic contrast (de Jong, 1995).

Empirical research has analyzed cases in which people were disfluent and then spontaneously self-corrected. For example, a person might say ''Her name is Sara, no... uh, **Susan** Collins.'' In the literature on spontaneous self-corrections, acoustic-prosodic changes have been reported between error and repair segments, which indicate that the self-repair tends to be accented, or rendered more prominent intonationally (Levelt and Cutler, 1983). However, prominence marking on content self-repairs occurs only intermittently—usually in less than half of the self-repairs observed (Levelt and Cutler, 1983; Howell and Young, 1991). Furthermore, self-repairs that do not involve the replacement of wrong content (e.g., disfluent repetitions) usually do not receive prominence marking, or else receive negligible marking (Howell and Young, 1991; Levelt and Cutler, 1983; O'Shaughnessy, 1992).

During human–computer interaction, there also is prominence marking when a speaker spontaneously corrects a disfluency. This marking involves longer duration, increased pitch, and increased amplitude of the repair segment (Nakatani and Hirschberg, 1994; O'Shaughnessy, 1992), although the reported increases in pitch and amplitude have been extremely small (Nakatani and Hirschberg, 1994). It currently is not known whether these changes during self-corrected disfluencies bear any similarity to hyperarticulate change elicited by system recognition errors. Among other differences, the latter type of repair occurs in the context of a highly interactive spoken exchange, and in direct response to a computer partner's failure. Another difference is that analyses of repairs following system error have compared identical lexical content before and after system failure, whereas analyses of self-corrected disfluencies have involved comparison of different lexical content before and during the repair.

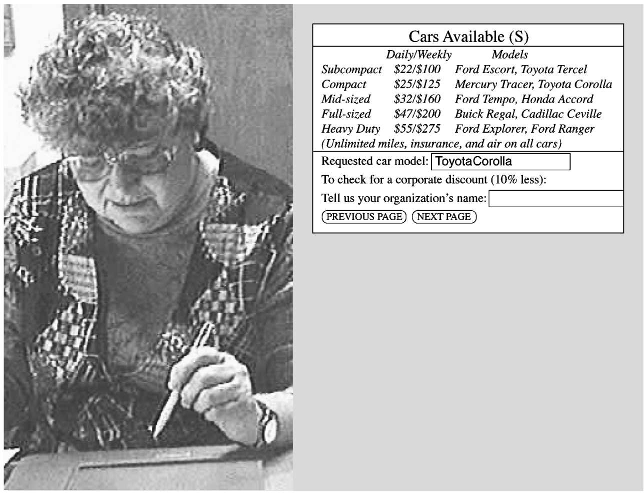
E. Goals and predictions of the study

The general goal of the present study was to examine the type and magnitude of linguistic adaptations that occur during human–computer error resolution. A further general aim was to develop a user-centered predictive model of hyperarticulate change during system error handling. The specific goals of this study were: (1) to provide a comprehensive analysis of acoustic, prosodic, amd phonological adaptations in speech during error resolution; (2) to test the generality of the CHAM model (computer-elicited hyperarticulate adaptation model) in response to qualitatively different types of system recognition error; (3) to examine changes in the speech signal during both *global* repair of an entire utterance and during *focal* repair of a syllable or word within an utterance; (4) to assess the relation between users' nonverbal reaction to system errors and change in the acoustic-prosodic features of their speech signal; and (5) to summarize implications of these findings for the development of improved error handling in next-generation spoken language and multimodal systems.

It was hypothesized that users' repetitions following system error would be adapted toward hyperclear acoustic-phonetic features, including higher amplitude, higher maximum pitch, lower minimum pitch, greater pitch range, longer duration of speech and pauses, more hyperclear 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 were analyzed following qualitatively different types of error, including failures-to-understand, related substitutions, and unrelated substitutions.[1] Results for these different error types were compared to evaluate whether the magnitude of hyperarticulate change would be greater when the computer substituted wrong lexical content, rather than simply failing to guess it, or when users responded to unintuitive system errors with visible emotional reactivity.

In addition to investigating hyperarticulation during global utterance repairs, it was hypothesized that speakers would mark focal repairs as more prominent acoustically than neighboring speech within an utterance. Increased amplitude, fundamental frequency, and durational effects were all explored as potential markers of prominence during focal repairs. Although pitch and amplitude are relatively inactive during error resolution involving global utterance repairs, it was predicted that they would exhibit more change during prominence marking in focal repairs. To calibrate durational effects and the selectivity of their placement, the magnitude of change for speech segments and pauses in the immediate focal repair region was compared with that in surrounding nonfocal areas.

A further aim of this study was to explore users' nonverbal reactivity to different types of recognition error. In particular, an assessment was made of whether users react more strongly when wrong content is introduced, especially during substitution errors perceived to be unrelated semanti-



Cars Available (S)

	Daily/Weekly	Models
Subcompact	*$22/$100*	*Ford Escort, Toyota Tercel*
Compact	*$25/$125*	*Mercury Tracer, Toyota Corolla*
Mid-sized	*$32/$160*	*Ford Tempo, Honda Accord*
Full-sized	*$47/$200*	*Buick Regal, Cadillac Ceville*
Heavy Duty	*$55/$275*	*Ford Explorer, Ford Ranger*

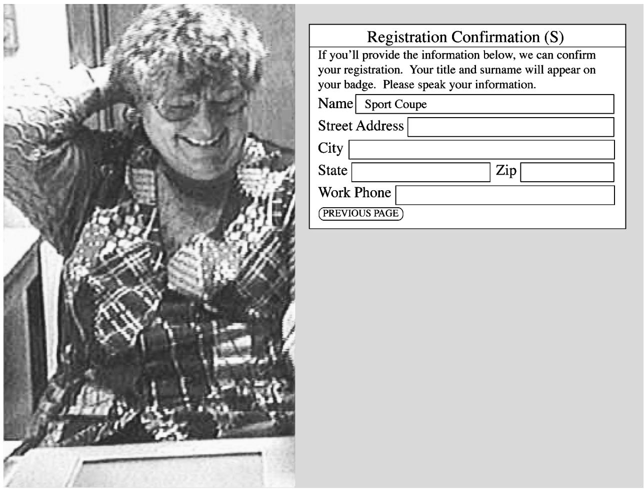
(Unlimited miles, insurance, and air on all cars)

Requested car model: ToyotaCorolla

To check for a corporate discount (10% less):

Tell us your organization's name:

FIG. 2. A user speaks her organization's name as ''National Oceanographic'' but it is misrecognized as ''International Graphics'' during a related substitution error.



Registration Confirmation (S)

If you'll provide the information below, we can confirm your registration. Your title and surname will appear on your badge. Please speak your information.

Name Sport Coupe

Street Address

City

State Zip

Work Phone

FIG. 3. A user speaks her name as ''Nancy Green'' and laughs when it is misrecognized as ''Sport Coupe'' during an unrelated substitution error.

cally and acoustically to their original input (e.g., ''Nancy Green'' misrecognized as ''Sport coupe''). If users are visibly more reactive to substitution errors, or to unrelated substitutions involving uninterpretable misrecognitions, then this greater degree of arousal may influence the signal characteristics of their repair speech. For example, volume and fundamental frequency may increase as a by-product of greater arousal.

I. METHOD

A. Subjects, tasks, and procedure

Twenty native English speakers, half male and half female, participated as paid volunteers. Their occupational backgrounds were varied, but excluded computer scientists.

A ''Service Transaction System'' was simulated that could assist users with conference registration and car rental transactions. Compared with an earlier study reported by Oviatt and colleagues (Oviatt *et al.*, 1996, 1998), in this study the corpus was designed to permit collection of a wider variety of articulated phonemes and three-fold more data than previously, in order to probe the generality of the CHAM Model. After a general orientation, people were shown how to enter information using a stylus to click-to-speak 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 *failure-to-understand* errors, the system responded with ''**????**'' feedback to indicate its failure to recognize lexical content. During these errors, the system basically informed the user of its inability to recognize what the user's input meant, so it was not necessary for the user to detect the error. In the case of *substitution* errors, illustrated in Figs. 2 and 3, the system instead responded with misrecognized or wrong content, such as ''**International Graphics**'' instead of ''**National Oceanographic**'' (i.e., acoustically and semantically *related* substitution), or with ''**Sport Coupe**'' instead of ''**Nancy Green**'' (i.e., *unrelated* substitution).

Following all errors, participants were instructed to try again by re-entering their information in the same slot until system feedback was correct. A form-based interface was used during data collection so that the locus of system errors would be clear to users. To successfully resolve a simulated error, the simulation was programmed so that the participant had to repeat their input once or twice, although analyses focused on the users' original input and first repetition.

Users were told that the system was a well-developed one with an extensive vocabulary and processing capabilities, so they could express things as they liked and not worry about what they could and could not say. They were advised that they could speak normally, work at their own pace, and just concentrate on completing their transaction. They also were told that if for any reason the computer did not understand them, they 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 had believed they were interacting with a fully functional system.

B. Semi-automatic simulation method

A flexible simulation method was devised for supporting varied studies on user responding during system recognition errors. The simulation developed for this purpose was an adapted version of a method previously outlined by Oviatt and colleagues (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 a 0.4-s 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 baserates, and different realistic properties of speech recognition errors. This error generation capability was designed to be preprogrammed and controlled automatically so that, for example, errors could be distributed randomly across all task content. For the present study, the error-generation software was adapted to deliver qualitatively different types of system recognition errors, including: (1) failures to understand; (2) related substitutions; and (3) unrelated substitutions.[2] The simulated word error rate was held constant at 15%, and approximately one error occurred every five input slots.

C. 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); (2) Type of simulated error (Failure to understand; Related substitution; Unrelated substitution). All 20 subjects completed 6 tasks. Within each task, six simulated errors were delivered—two failures to understand, two related substitutions, and two unrelated substitutions. This represents a 20% rate of utterances containing an error, which is comparable to that associated with the CHAM model's Stage II changes in previous reports. In total, data were collected on 36 simulated errors per subject, or 720 simulated errors in the study. For all matched utterance pairs in which the lexical content was the same, original input provided a baseline for quantifying change during the first repetition. In total, this included approximately 638 utterance pairs, with over 200 representing each error type.

D. 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.

1. Global linguistic adaptations

In these analyses, global spoken adaptations that occurred within the entire utterance were assessed.

a. 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 10 ms in duration. Due to difficulty locating their onset, utterance-initial voiceless stops and affricates were arbitrarily assigned a 20-ms closure, and no pauses were coded as occurring immediately before utterance-medial voiceless stops and affricates. Further details of durational scoring conventions are outlined elsewhere (Moreton, 1996).

b. 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 nonspeech sounds were excluded.

c. Fundamental frequency. Spoken input was coded for maximum $F0$, minimum $F0$, and $F0$ range. A pitch-smoothing filter was applied to the data to remove or minimize: (1) glottalized regions; (2) spurious doubling and halving; (3) points below an amplitude threshold of 400 rms; and (4) 1- to 2-point pitch value outliers (e.g., due to hissing sound in ''s''). 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 further to correct for pitch tracker errors such as spurious doubling and halving, interjected nonspeech sounds, and extreme glottalization affecting $\leqslant 5$ tracking points.

d. 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. Finally, the percentage of all original versus repeated utterances that contained a final falling contour was compared.

e. 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 speech. A focus was placed on identifying uncontroversial phonological changes with respect to the conversational-to-clear speech continuum, and those that could be coded reliably by ear without access to a spectrogram. For example, cases of glottalization and glottal stop insertion were not included due to known difficulty with reliable coding (Eisen *et al.*, 1992).

f. Disfluencies. Spoken disfluencies were totaled for each subject and condition during original spoken input as well as repeats during errors, 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).

g. Nonverbal responding. To assess users' subjective reaction to different types of recognition error, the following categories of nonverbal responding were coded from videotapes for each subject and error condition: (1) smiling—lips fully retracted upward in an unambiguous smile; (2) laughter—open-mouth smile accompanied by one or more breathy nonarticulated bursts of noise; (3) raised brows—eyebrows lifted upward, as if in surprise; and (4) knit brows—eyebrows moved together, with forehead wrinkled as muscles contract. These nonverbal facial changes, which were considered indices of emotional reactivity and heightened arousal, were assessed for possible correspondence with speech signal changes.

h. Self-reported perception of recognition errors. The percentage of subjects reporting specific beliefs about the causal basis of errors, as well as effective ways to resolve errors, was summarized from post-experimental interviews.

2. Focal linguistic adaptations

These analyses concentrated on focal error repairs involving one syllable or word within a longer multi-word utterance. In total, 96 original-repeat utterance pairs were available for analysis of focal error repairs, which constituted a subset of the related substitution errors. Examples of focal repairs during related substitution errors were *two seven seven Frill Street*→''two seven seven **Hill** Street,'' *September seven, 1996*→''September **eleven**, 1996.'' The goal of these analyses was to assess whether and to what extent the focal repair region received selective emphasis via acoustic cues during system error resolution.

a. Duration. (1) *Focal Speech Duration*—The total duration of the focal speech segment [FOC], which represented the repair region, was evaluated for original and repeat input.

(2) *Nonfocal Speech Duration*—The total duration of the surrounding nonfocal speech segments [NFOC] (i.e., total utterance duration minus focal speech duration minus total pause duration) was computed.

(3) *FOC/NFOC Speech Duration Ratio*—The ratio of focal to surrounding speech segment durations was computed to assess whether the focal region was relatively more elongated during repetition than surrounding speech.

(4) *Pause Duration Adjacent to Repair*—For all utterances with one or more pauses, total pause duration was computed both immediately before and after the repair region in original and repeated input.

(5) *Pause Duration Nonadjacent to Repair*—Total pause duration also was assessed for pauses not adjacent to a focal repair region in original and repeat utterances.

(6) *Number of Pauses Adjacent to Repair*—For all multiword utterances, the total number of pauses immediately before and after a focal repair region were scored for original and repeat utterances.

b. Amplitude. (1) *Focal Maximum Amplitude*—Maximum amplitude was computed from the loudest point during the focal repair region, and was summarized for both original and repeat utterances.

(2) *Nonfocal Maximum Amplitude*—The average maximum amplitude of spoken words not in the focal repair region also was calculated.

(3) *FOC/NFOC Amplitude Ratio*—The ratio of focal to nonfocal speech segment amplitudes was computed to assess whether the focal repair region had a relatively higher amplitude during repetition.

c. Fundamental frequency. (1) *Focal Pitch Maximum*—Maximum $F0$ during the focal repair was scored for original and repeat utterances, and analyzed separately when the repair was in sentence-final versus initial or medial position.

(2) *Nonfocal Pitch Maximum*—The average maximum $F0$ of nonfocal spoken words also was calculated, excluding words in sentence-final position.

(3) *Focal Pitch Minimum*—Minimum $F0$ during the focal repair was scored for original and repeat utterances, and analyzed separately when the repair was in sentence-final versus initial or medial position.

(4) *Nonfocal Pitch Minimum*—The average minimum $F0$ of nonfocal spoken words also was calculated, excluding words in sentence-final position.

(5) *Focal Pitch Range*—The $F0$ range ($F0$ maximum minus $F0$ minimum) was scored for focal repair segments occurring in all sentence positions, and then compared for original and repeat utterances.

(6) *Nonfocal Pitch Range*—The average $F0$ range of nonfocal spoken words also was scored, and compared for original and repeat input.

d. Reliability. For all measures reported except amplitude, 10%–100% of the data were second scored. For discrete classifications, such as number of pauses, disfluencies, phonological alternations, nonverbal responding, and intonation contour, all inter-rater reliabilities exceeded 88%. For phonological alternations, only cases agreed upon by both scorers were analyzed. For fundamental frequency, the inter-rater reliability for minimum $F0$ was an 80% match with less than 3-Hz departure, and for maximum $F0$ an 80% match with less than 9-Hz departure. For duration, pause length was an 80% match with less than 65-ms departure, and total utterance duration an 80% match with less than 59-ms departure.

II. RESULTS

Speech data were available for analysis on approximately 638 scorable utterance pairs for which the lexical content was identical during original and repeated input. Of these, over 200 utterance pairs representing each of the three error types were analysed. Spoken utterances in this corpus tended to be brief fragments averaging two to three words, and ranging from 1 to 13 words in length.

A. Overview of global linguistic adaptations

Table I presents a summary of all the significant global linguistic changes identified during human–computer error resolution. The magnitude of relative change shown for each linguistic dimension is an average across the three different error types. Specific results on each type of linguistic change are detailed in the following sections.

TABLE I. Overview of relative change in linguistic dimensions of hyperarticulation during global utterance repairs.

Type of change	Percentage change during repetition
Pause interjection	+44.0%
Pause elongation	+40.0%
Disfluencies	−38.5%
Intonation—final fall	+20.0%
Speech elongation	+8.5%
Hyperclear phonology	+6.0%
Pitch minimum	−2.0%
Amplitude	+0.5%

Table I clarifies that change in pause structure dominated hyperarticulate adaptation during error resolution, with durational increase in the speech segment also noteworthy but smaller in magnitude. Articulatory changes were a second prominent characteristic of global hyperarticulate adaptation, including both a drop in spoken disfluencies and an increase in hyperclear phonological features. With respect to prosody, speakers shifted to a final falling intonation contour during repetitions, which was associated with small decreases in fundamental frequency measures. While amplitude increases were reliably present, they were negligible. (Figure 7 illustrates that the overall profile of hyperarticulate adaptations was replicated across all three of the different error types.)

1. Duration

Total utterance duration averaged 1567 ms and 1786 ms in original and repeat input during failure-to-understand errors, 1677 ms and 1845 ms during related substitutions, and 1659 ms and 1815 ms during unrelated substitutions. The average gain in total utterance duration from original to repeated speech across all error types was +11%. A repeated measures ANOVA on log transformed data revealed that the main effect of original versus repeat speech was a significant one, $F=166.05$ (df=1, 165), $p<0.001$, although the main effect of type of recognition error was not significant, $F<1$, nor was the interaction between error type and original-repeat speech, $F=2.30$ (df=2, 330), N.S. Having ruled out significant variation in utterance duration due to type of recognition error, *a priori* paired *t* tests then were conducted on the prediction that duration would be elongated during repetition following all three types of error. These analyses confirmed a significant increase in utterance length for failure-to-understand errors, paired $t=4.58$ (df=197), $p<0.001$, one-tailed, for related substitution errors, paired $t=8.93$ (df=205), $p<0.001$, one-tailed, and for unrelated substitution errors, paired $t=6.63$ (df=219), $p<0.001$, one-tailed.

a. Speech segment duration. Analyses revealed an increase in the total speech segment from an average of 1446 ms during original input to 1591 ms during repetitions following failure-to-understand errors, 1525 ms and 1662 ms following related substitutions, and 1513 ms and 1613 ms following unrelated substitutions, as illustrated in Fig. 4. The average relative gain in speech segment duration from original to repeated speech across all error types was +8.5%. A repeated measures ANOVA on log transformed data revealed that the main effect of original versus repeat speech was significant, $F=142.46$ (df=1, 165), $p<0.001$, although error type was not, $F<1$, nor was the interaction between error type and original-repeat speech, $F=2.85$ (df=2, 330), N.S. Having ruled out significant variation due to type of recognition error, *a priori* paired *t*-tests were conducted on the prediction that repeated speech segments would be significantly elongated following all three types of recognition error. These analyses confirmed a significant increase in speech segment duration following failure-to-understand errors, paired $t=6.88$ (df=197), $p<0.001$, one-tailed, related substitutions, paired $t=8.95$ (df=205), $p<0.001$, one-tailed, and unrelated substitutions, paired $t=5.69$ (df=219), $p<0.001$, one-tailed.

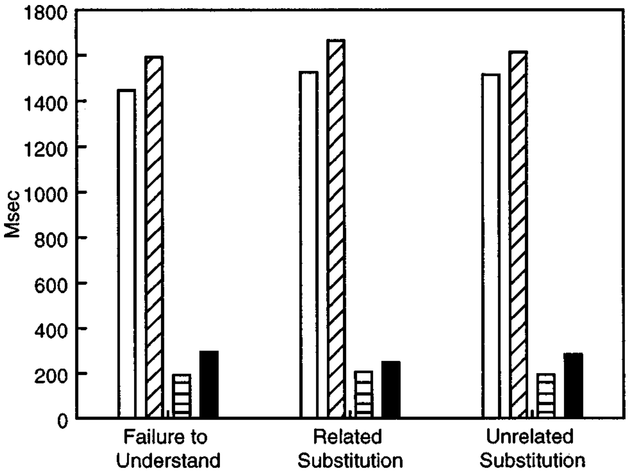


FIG. 4. Elongation of the speech segment and pauses in repeated utterances for three types of recognition error. Original speech □; repeat speech ▨; original pause ⊟; repeat pause ■.

b. Pause duration. The total pause duration of multiword utterances also increased from an average of 192–295 ms between original and repeat input after failure to understand errors, from 207 ms to 248 ms after related substitutions, and 193 ms to 283 ms after unrelated substitutions. The average gain in total pause duration from original to repeated speech across all error types was +40%. A repeated measures ANOVA on log transformed data revealed that the main effect of original versus repeat speech significantly influenced total pause duration, $F=57.68$ (df=1, 56), $p<0.001$, although type of error did not, $F<1$, nor did the interaction between error type and original-repeat speech, $F=1.93$ (df=2, 112), N.S. Having ruled out significant variation due to type of error, *a priori* paired *t*-tests were conducted on the prediction that pause duration would be elongated significantly in response to all three types of recognition error. These analyses confirmed a significant increase in pause duration following failure-to-understand errors, paired $t=5.59$ (df=77), $p<0.001$, one-tailed, related substitutions, paired $t=5.74$ (df=93), $p<0.001$, one-tailed, and unrelated substitutions, paired $t=4.59$ (df=84), $p<0.001$, one-tailed.

Figure 4 illustrates the average increase in pause duration for all three types of error, and its relation to increases in speech segment duration. Figure 5 also shows the increasing

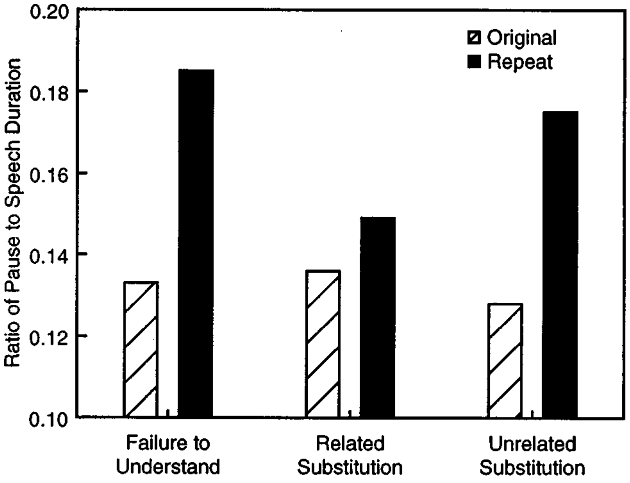


FIG. 5. Increasing ratio of pause to speech duration in repeated utterances for three types of recognition error.

ratio of pause to speech duration in repeated speech for all three types of error, which averaged +13% during original input but increased to +17% during repetitions. That is, the gain in pause duration during repetitions was relatively greater than for speech, a comparison that was statistically reliable across subjects by Wilcoxon Signed Ranks test, $z = 3.24$ ($N = 19$), $p < 0.001$, two-tailed.

To test for elongation of individual matched pauses (i.e., independent of interjecting new ones that may have been brief), 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 following all three types of errors, including failure-to-understand errors, paired $t = 2.37$ (df $=34$), $p < 0.02$, one-tailed, related substitutions, paired $t = 2.02$ (df$=$49), $p < 0.025$, one-tailed, and unrelated substitutions, paired $t = 3.60$ (df$=$45), $p < 0.001$, one-tailed.

c. Number of pauses. Approximately 63% of multi-word utterances contained one or more pauses during error resolution, even though utterances in the corpus tended to be brief. Figure 6 reveals that the average number of pauses per subject for multi-word utterances increased during repeat speech for all three error types. For failure-to-understand errors, the number of pauses increased from an average of 0.69 to 1.08 between original and repeated utterances, significant by Wilcoxon Signed Ranks test, $z = 3.32$ ($N = 17$), $p < 0.001$, one-tailed. For related substitutions, the number of pauses increased from 0.71 to 1.06 during repeat utterances, again significant by Wilcoxon, $z = 3.62$ ($N = 17$), $p < 0.001$, one-tailed. Finally, for unrelated substitutions, pauses increased from 0.83 to 1.04 during repeat utterances, significant by Wilcoxon, $z = 2.12$ ($N = 16$), $p < 0.02$, one-tailed. Overall, the net increase in average number of pauses during repeated speech was +44%.

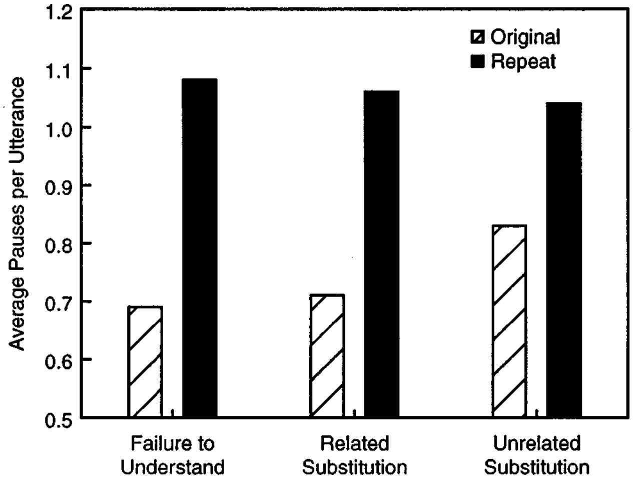


FIG. 6. Increase in number of pauses in repeated utterances for three types of recognition error.

2. Amplitude

The maximum amplitude averaged 70.3 dB and 70.6 dB in original and repeat utterances during failures to understand, 70.8 dB and 71.1 dB during related substitutions, and 70.6 dB and 71.0 dB during unrelated substitutions. A repeated measures ANOVA revealed that the main effect of original-repeat speech had a significant impact on amplitude, $F = 23.76$ (df$=$1, 163), $p < 0.001$, but there was no difference between error types, $F = 1.45$ (df$=$2, 326), N.S., and no significant interaction, $F = 1.40$ (df$=$2, 326), N.S. Having ruled out significant variation due to type of error, the prediction was tested that repeated speech would be increased in amplitude. Analyses using planned t tests confirmed a significant increase in amplitude on repeat speech for failures to understand, $t = 2.45$ (df$=$204), $p < 0.01$, one-tailed, for related substitutions, $t = 3.00$ (df$=$208), $p < 0.0015$, one-tailed, and for unrelated substitutions, $t = 3.57$ (df$=$223), $p < 0.001$, one-tailed. However, these increases were very neglible, averaging less than +0.5%.

3. Fundamental frequency

a. Pitch maximum. Maximum $F0$ averaged 190.8 and 190.2 for original and repeat speech during failures to understand, 188.8 and 189.6 for original and repeat speech during related substitutions, and 193.0 and 192.9 for original and repeat speech during unrelated substitutions. Repeated measure ANOVAs conducted on the whole sample and then reanalyzed separately by gender all revealed no significant effect of original versus repeat speech, error type, or their interaction on pitch maximum values (Fs<1).

b. Pitch minimum. Minimum $F0$ averaged 129.5 and 126.8 on original and repeat speech during failures to understand, 129.9 and 127.4 during related substitutions, and 129.1 and 127.6 during unrelated substitutions. A repeated measures ANOVA conducted on the whole sample revealed a significant main effect of original versus repeat speech, $F = 4.68$ (df$=$1, 158), $p < 0.035$, but no difference due to error type, $F < 1$, or their interaction, $F = 1.90$ (df$=$2, 316), $p > 0.15$. Since a decrease was predicted in minimum $F0$ during repetitions, *a priori* paired t-tests were conducted to assess predicted drops during different error types. Significant decreases were confirmed for failure to understand errors, $t = 2.42$ (df$=$189), $p < 0.01$, one-tailed, for related substitution errors, $t = 2.16$ (df$=$190), $p < 0.02$, one-tailed, and for unrelated substitution errors, $t = 1.76$ (df$=216$), $p < 0.04$, one-tailed. These decreases in minimum $F0$ averaged less than -2%.

c. Pitch range. $F0$ range averaged 61.9 and 63.2 for original and repeat speech during failures to understand, 62.7 and 62.8 during related substitutions, and 63.9 and 65.4 during unrelated substitutions. Repeated measures ANOVAs conducted on the whole sample and then reanalyzed separately by gender all revealed no significant main effect of original versus repeat speech, error type, or their interaction on overall pitch range values for the utterance (Fs<1).

4. Intonation contour

The probability of *shifting* final intonation contour from rise to fall, or vice versa, averaged only 11.5% between original and repeated input. More specifically, speakers maintained the same final contour 89% of the time during failure to understand errors, 87% of the time during related substitutions, and 89% during unrelated substitutions, with no significant differences apparent between error types. 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, $z = 3.88$ ($N = 20$), $p < 0.001$, one-tailed. In this sense, it appears that whatever intonation contour originally is applied to the utterance tends to persist during verbatim correction.

Of the cases in which a change was evident in final intonation contour during repetition, 88% of the time the shift was from rising to falling, rather than the reverse. This difference was significant by Wilcoxon test, $T+ = 110$ ($N = 15$), $p < 0.003$, two-tailed. Analyses of all three error types reconfirmed this pattern of significantly more final falls than rises during repetitions. Overall, the likelihood of a final falling contour was 45% during original input, increasing to 54% during repetitions—for a net relative increase in final falling contours of $+20\%$.

5. Phonological Alternations

Approximately 6% of repetitions in this corpus contained a phonological alternation that could be classified along the hyperarticulation spectrum. Table II summarizes the number and type of alternations observed for each subject by the direction of shift toward conversational versus hyperclear speech.

The majority of subjects, or 79% of those who had at least one spoken adaptation classifiable according to hyperarticulation, shifted more often from a conversational to clear speech style, rather than the reverse, a significant difference by Wilcoxon Signed Ranks test, $T+ = 80$ ($N = 13$), $p < 0.007$, one-tailed. The rate of hyperclear alternations averaged 6% of repetitions during failures to understand, 4% of repetitions during related substitutions, and 5% during unrelated substitutions, with no significant difference among error types (see Fig. 7).

TABLE II. Number and type of phonological alternations involving a shift toward clear speech (a–f) versus toward conversational speech (g–h), listed by subject.

Clear to conversational	Conversational to Clear	Phonological alternations
0	3	c, d, d
0	6	a, a, a, c, e, e
2	2	g, g / a, d
0	1	a
0	0	⋯
0	0	⋯
0	5	a, a, a, c, c
0	3	a, c, d
0	1	f
0	0	⋯
0	3	a, b, c
0	1	a
2	0	g, h
0	0	⋯
0	1	d
0	3	a, a, c
0	3	a, a, c
0	0	⋯
1	0	g
0	0	⋯
Total—5	32	

[a] Unreleased t>released t.

[b] Alveolar flap>coronal plosive.

[c] n/alveolar nasal flap>nt sequence.

[d] Segment insertion.

[e] Nasal flap>n.

[f] schwa>I altered vowel quality.

[g] Segment deletion.

[h] nt sequence>nasal flap.

When one or more clear-speech phonological changes were present during repetitions, the number of pauses correspondingly increased $+67\%$ from baseline input (i.e., from 0.90 to 1.50 pauses between original and repeat input), compared with a gain of only $+44\%$ for the whole corpus. Likewise, total pause length increased $+61\%$ from baseline input to repetitions when a phonological change was present (i.e., from 191 ms to 307 ms), although the gain only averaged $+40\%$ for the whole corpus. Total speech duration averaged 1891 ms and 2126 ms during utterances with a phonological alteration, a $+12\%$ increase over baseline input, compared with $+8.5\%$ increase for the whole corpus. In short, durational change averaged about 49% greater during repetitions involving a phonological alternation than during those without one. When original-repeat utterance pairs containing a conversational-to-clear-speech phonological change were

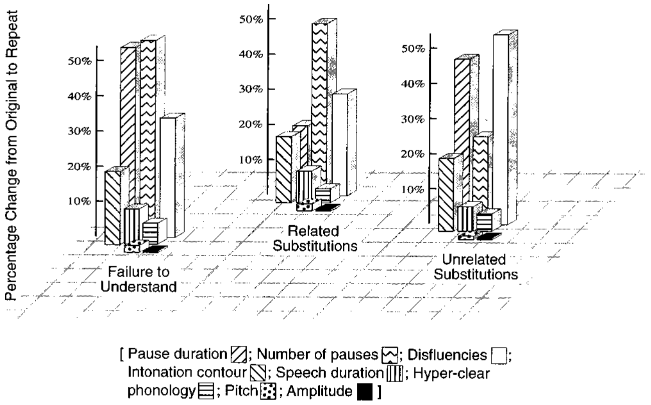


FIG. 7. Similarity of hyperarticulation profile for different error types.

compared with utterances from the corpus at large that did not contain any phonological change (i.e., that were matched on speaker and lexical content), it was confirmed that utterances with phonological change contained significantly more pauses than those in the corpus at large, $T+=55$ $(N=10)$, $p<0.001$ one-tailed, and also significantly longer pauses, $T+=79$ $(N=13)$, $p<0.01$, one-tailed. These data clarify that the degree of hyperarticulate adaptation varied along a spectrum, and also that durational and phonological changes during hyperarticulation were correlated within individual utterances.

6. Disfluencies

The disfluency rate during baseline speech (i.e., throughout the interaction when no errors were occurring) averaged 0.65 disfluencies per 100 words. However, this rate dropped to 0.40 during repeated input following system errors, a significant decrease by Wilcoxon Signed Ranks test, $T+=103$ $(N=15)$, $p<0.01$, one-tailed. The rate of disfluencies per 100 words averaged 0.43 during failures to understand, 0.46 during related substitutions, and 0.30 during unrelated substitutions, which did not differ significantly.

7. Nonverbal responding

Users frequently reacted emotionally to system recognition failures. They smiled in response to 9% of errors, laughed after another 6%, raised their eyebrows after 4%, and knit their brows after 3% of errors. In total, 22% of system errors elicited a nonverbal response.

Participants were significantly more likely to smile after an unrelated substitution than after a failure to understand error, $z=2.73$ $(N=11)$, $p<0.003$, one-tailed, or after a related substitution error, $z=1.69$ $(N=11)$, $p<0.05$, one-tailed. Users also were significantly more likely to laugh after unrelated substitutions than after a failure to understand error, $z=2.40$ $(N=9)$, $p<0.01$, one-tailed, or after a related substitution, $z=2.45$ $(N=9)$, $p<0.007$, one-tailed. Finally, although raised eyebrows were not expressed more often after any particular error type, users also knit their brows significantly more often after unrelated substitutions than failure to understand errors, $z=1.81$ $(N=7)$, $p<0.04$, one-tailed, and related substitutions, $z=1.62$ $(N=7)$, $p<0.053$, one-tailed. In summary, participants were most reactive to the unrelated substitution errors.

8. Self-reported perception of recognition errors

Post-experimental interviews revealed that users typically posited a cause for errors that involved self-attribution of blame and a linguistically based cause of system failure (e.g., ''I just needed to speak more slowly and clearly''). Although the delivery of simulated recognition errors was not contingent at all on users' input, 70% of interviewees stated that altering the linguistic characteristics of their own language was effective in repairing system errors successfully. Another 15% said they had no idea why system errors occurred, and the remaining 15% cited mechanical reasons for recognition failure (e.g., ''My pen wasn't inside the input box, so it didn't get the last few digits'').

TABLE III. Overview of relative change in linguistic dimensions of hyperarticulation during focal repairs.

Type of change	Percentage change during repetition
Focal repair region:	
Pause duration next to repair	+149%
Number of pauses next to repair	+113%
Duration of speech repair	+18%
Pitch range of speech repair	+11%
Pitch maximum of speech repair[a]	+3%
Pitch minimum of speech repair[b]	−3%
Amplitude of speech repair	+1%
Nonfocal region:	
Pause duration nonadjacent to repair	+9%
Duration of nonfocal speech	+9%

[a] Change for all focal repairs, except those in sentence-final position.
[b] Change for focal repairs in sentence-final position only.

With respect to linguistic repair mechanisms, the following specific ones were cited most frequently as being effective: (a) speaking more clearly—mentioned by 45% of participants who maintained a linguistic theory; and (b) speaking more slowly—40% of participants. A small minority of people said they believed that speaking more loudly to the computer was effective in resolving errors (10%), or changing voice inflection (5%). In short, participants' self-reports regarding error repair strategies were consistent with the major changes observed in hyperarticulate speech.

B. Overview of focal linguistic adaptations

Table III presents a summary of all the significant focal linguistic adaptations that were identified during human–computer error resolution. It summarizes changes that occurred when users selectively emphasized a focal repair region in a related substitution error. Specific results on each type of linguistic adaptation are detailed in the following sections.

Table III clarifies that change in pause structure still dominated focal hyperarticulate adaptation, although it was three- to four-fold greater than that observed during global utterance repair. Changes in pause interjection and elongation also were selectively placed adjacent to the focal repair region. In fact, these pause changes were twelve-to sixteen-fold more pronounced immediately before and after the repair region than in other sentence positions. The focal speech region also was substantially elongated, approximately two-fold more than speech in surrounding nonfocal regions or during global utterance repairs.

Although relatively smaller in magnitude of change, the focal repair region also was selectively marked with a moderate increase in pitch range that was derived from an increase in maximum pitch in sentence-initial and medial positions and a decrease in minimum pitch in sentence-final position. Finally, the focal repair region was selectively marked with a small increase in amplitude. These data clarify how duration, fundamental frequency, and amplitude work together in a finely tuned manner to mark a highly specific repair region as acoustically more prominent than surrounding ones during human–computer error resolution.

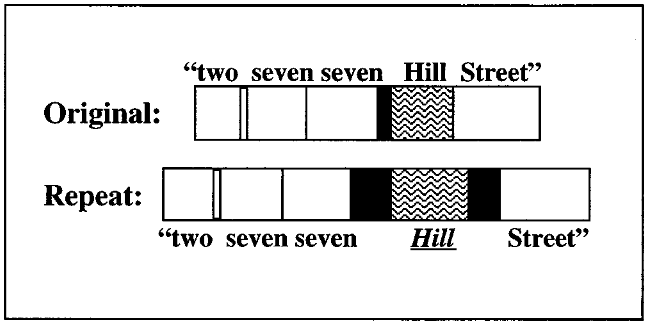


FIG. 8. During repair of a related substitution error, elongation of the focal speech region (box with wavy lines) and selective pause interjection and elongation next to the focal repair (■), compared with nonfocal utterance regions (□).

1. Duration

a. Focal speech duration. The total duration of the focal speech segment increased from an average of 400 ms during original input to 473 ms during repetition, a gain of +18%. This increase was significant by paired t test on log transformed data, $t = 6.02$ (df=95), $p < 0.001$, one-tailed.

b. Nonfocal speech duration. The total duration of the surrounding speech segments also increased from an average of 745 ms during original input to 811 ms during repetition, a gain of +9%. This increase also was significant by paired t-test on log transformed data, $t = 5.11$ (df=95), $p < 0.001$, one-tailed.

c. FOC/NFOC speech duration ratio. The ratio of focal to nonfocal speech duration increased significantly during repetition, paired $t = 2.13$ (df=95), $p < 0.02$, one-tailed. That is, the focal speech region was demonstrated to increase significantly more than other surrounding speech segments.

d. Pause duration adjacent to repair. Approximately 47% of all multi-word utterances contained one or two pauses adjacent to the focal speech repair during error resolution. The total duration of such pauses averaged 72 ms during original input, increasing to 179 ms during repetition, which was significant by paired t test on log transformed data, $t = 5.60$ (df=35), $p < 0.001$, one-tailed. That is, a substantial +149% increase was discovered in average pause duration immediately next to the focal repair region during repetitions. This increase in total pause duration was comparable in size for pauses positioned immediately before versus after the repair region (i.e., averaging 178.5 vs, 180.0 ms, respectively).

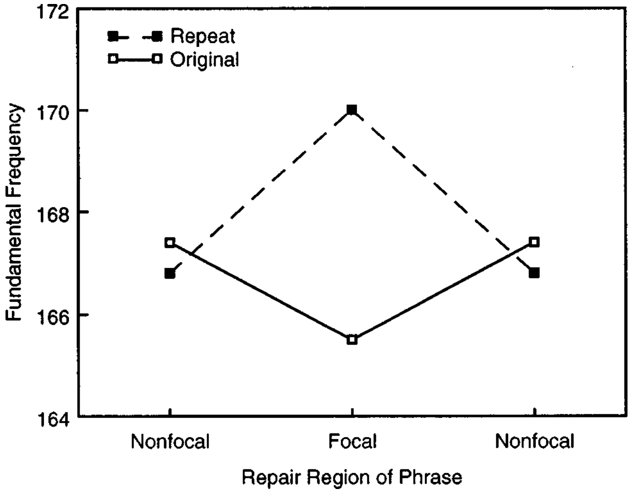


FIG. 10. Change in pitch maximum on focal repair region versus nonfocal segments, for focal repairs in sentence initial and medial position.

Further analysis confirmed that both interjection of new pauses and elongation of existing ones contributed independently to observed increases in total pause duration immediately around the focal region. In original-repeat utterance pairs for which the number of pauses was matched, pause elongation still was significant by paired t test, $t = 2.96$ (df $= 13$), $p < 0.01$, one-tailed.

e. Pause duration nonadjacent to repair. Pause duration for positions nonadjacent to the repair region averaged 128 ms during original input and 140 ms during repetitions, a +9% increase. This increase also was significant by paired t test, $t = 2.02$ (df=13), $p < 0.04$.

f. Number of pauses adjacent to repair. The number of pauses immediately adjacent to a repair region averaged 0.80 during original input, increasing to 1.70 during repetitions, a

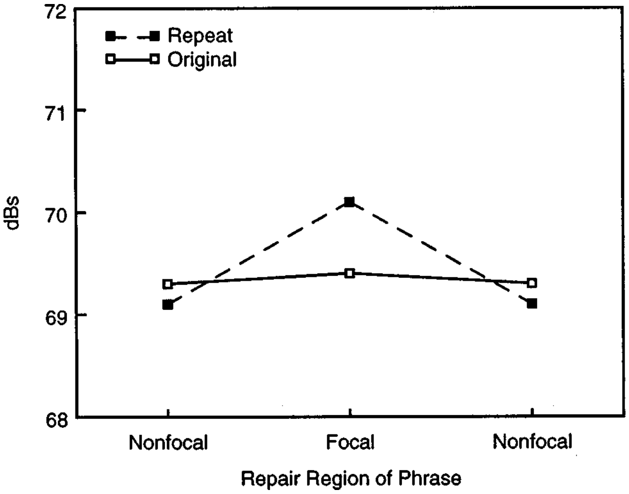


FIG. 9. Amplitude change on focal repair region versus nonfocal segments during related substitutions.

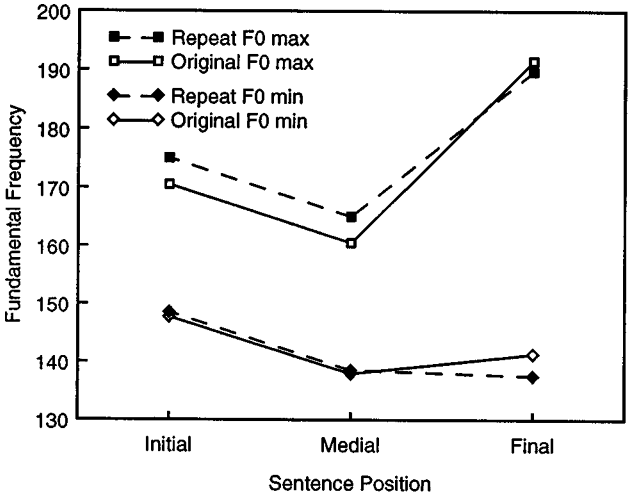


FIG. 11. Change in pitch maximum and minimum on focal repairs as a function of sentence position.

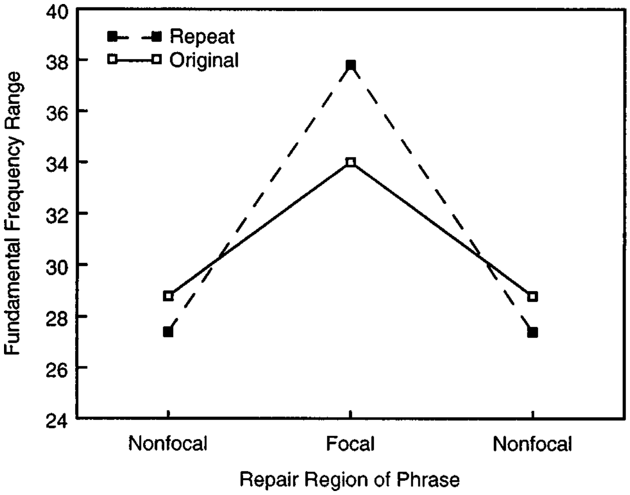


FIG. 12. Change in pitch range on focal repair region versus nonfocal segments during related substitutions.

+113% gain. This increase in average number of pauses was significant by Wilcoxon Signed Ranks test, $z = 2.49$ ($N = 12$), $p < 0.01$, one-tailed. Analysis of the position of these pauses indicated an equal split between those located immediately before versus after the repair.

Figure 8 illustrates selective pause interjection and elongation immediately around the focal repair region, as well as elongation of the spoken repair region itself, during focal repair of a typical related substitution error from the present corpus.

2. Amplitude

a. Focal maximum amplitude. Maximum amplitude of the focal region averaged 69.4 dB during original input, increasing to 70.1 dB during repetition, which represented a +1% gain. This increase on the focal segment was significant by paired t test, $t = 3.15$ (df=95), $p < 0.001$, one-tailed.

b. Nonfocal maximum amplitude. Average maximum amplitude of the nonfocal repair region was 69.3 dB during original input and 69.1 dB during repetitions, which was not a significant change, $t < 1$.

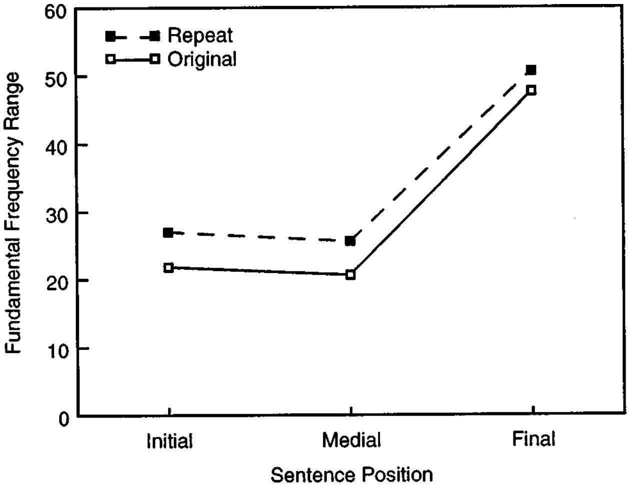


FIG. 13. Change in pitch range on focal repairs as a function of sentence position.

c. FOC/NFOC amplitude ratio. The ratio of maximum amplitude gain on focal versus nonfocal regions increased from 1.00 during original input to 1.01 during repetition, a significant relative change by paired t test, $t = 2.71$ (df=95), $p < 0.004$, one-tailed. Figure 9 illustrates the amplitude increase on focal repair regions during repeated utterances, compared with nonfocal segments.

3. Fundamental frequency

a. Focal pitch maximum. Maximum $F0$ on focal speech segments averaged 165.5 during original input and increased to 170.0 during repetition of the focal repair when it occurred in sentence initial and medial positions. This change represented a +2.7% increase in maximum $F0$ on the target repair region, which was significant by paired t test, $t = 2.86$ (df=50), $p < 0.003$, one-tailed. However, when the focal repair was in sentence-final position maximum $F0$ averaged 191.5 and 189.7 during original and repeated input, which did not represent a significant change, $t < 1$.

b. Nonfocal pitch maximum. The average maximum $F0$ on nonfocal speech segments (i.e., excluding those occurring in sentence-final position) was 167.4 and 166.8 during original and repeated input, which was not a significant increase, $t < 1$.

Figure 10 illustrates the increase in maximum $F0$ during repetition of a focal repair segment, compared with surrounding nonfocal ones. Figure 11 (top) illustrates that this increase occurred when the repair region was in any position except final.

c. Focal pitch minimum. Minimum $F0$ on focal speech segments averaged 143.6 and 143.5 on original input and repetitions when the focal repair was in sentence-initial or medial position, which did not represent a significant change, $t < 1$. However, when the focal repair was in sentence-final position, minimum $F0$ averaged 141.3 on original input and dropped to 137.5 during repetitions, which was a −2.7% decrease and significant by paired t test, $t = 1.72$ (df=41), $p < 0.05$. Figure 11 (bottom) illustrates that this decrease in minimum pitch only occurred in sentence-final position.

d. Nonfocal pitch minimum. The minimum $F0$ on nonfocal speech segments occurring in sentence-final position averaged 135.6 and 133.3 during original input and repetition, which did not represent a significant difference, $t < 1$. That is, sentence-final speech segments that were not the focus of repair showed no reliable drop in minimum $F0$ during repetitions.

e. Focal pitch range. The $F0$ range on focal repair segments occurring in all sentence positions averaged 34.0 for original input, increasing to 37.8 for repetitions. This was an +11.2% gain, and a significant expansion of pitch range, $t = 2.11$ (df=95), $p < 0.02$, one-tailed.

f. Nonfocal pitch range. The $F0$ range averaged a lower 28.8 and 27.4 for original and repeated input for speech segments throughout the sentence that were not the focus of repair, which did not represent a significant difference, $t = 1.09$ (df=92), N.S.

Figure 12 illustrates the increase in pitch range during repetition of focal repair segments, compared with surrounding nonfocal ones. Figure 13 illustrates that this pitch range expansion on focal repairs occurred uniformly in all sentence positions.

III. 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 system recognition errors can be a forceful elicitor of hyperarticulate speech from users. Furthermore, the presence, form, and degree of hyperarticulation in users' speech to computers is a highly predictable phenomenon. It has a specific constellation of linguistic features, and it occurs as a generalized response to different types of system recognition error. In addition, hyperarticulate adaptation can occur during global utterance repairs, and also during focal repairs involving one isolated region within a longer utterance. These research findings raise concerns with current algorithmic approaches to recognizing spoken language, which generally fail 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.

A. Global hyperarticulation to computers

During global utterance repairs, speech predominantly shifted to become lengthier and more clearly articulated, as summarized in Table I. Comparable durational changes were observed following all three types of system error, including +8.5% average elongation of the speech segment, +40% elongation of pause duration, and interjection of +44% more pauses. The most salient relative changes in repeated speech involved altered pause structure. Perhaps ironically, users' speech became somewhat more discrete during hyperarticulation, departing from the pattern of continuous speech upon which most current recognizers typically are trained. However, the changes observed in pause structure in no sense approached regularized discrete pausing between every word, as would be required by a discrete word recognizer. Instead, it often was highly targeted, as in selective pause interjection and elongation around focal repair regions.

The large durational increases obtained in this study are similar to those documented in hyperclear speech to the hearing impaired (Uchanski *et al.*, 1996). Previous literature on interpersonal speech also has reported increases in the number and length of pauses in hyperclear speech between people without hearing impairments (Cutler and Butterfield, 1990, 1991). In general, such changes in pause structure appear to play an important role in assisting listeners with marking word boundaries and segmenting a continuous stream of speech (Cutler and Butterfield, 1990, 1991; Maasen, 1986).

Articulatory changes also were a prominent characteristic of global hyperarticulate adaptation. The phonological features of repeat speech adapted toward an audibly clearer articulation pattern on 6% of repetitions, with frequently observed changes including the insertion of previously deleted segments (e.g., *'leven* changing to *eleven*), fortition of alveolar flaps to coronal plosives (e.g., eɪɾeɪt changing to *eIt¯eIt), and shifts to unreduced* **nt** sequences (e.g., twɛr̃i to twɛnti). This shift also corresponded with a 38.5% decrease in spoken disfluencies, which may have occurred in part because rearticulated utterances involve a reduced planning load (Oviatt, 1995). Essentially, users' speech became more deliberate and better specified in its signal cues to phonetic identity. These findings are consistent with the linguistic literature on hyperclear speech between people, which has reported change in both vowel and consonant quality including, for example, more audibly released word-final stops (Chen, 1980; Cutler and Butterfield, 1991; Moon, 1991; Picheny *et al.*, 1986). In future research, more detailed quantitative modeling will be needed on the major durational and articulatory changes observed during hyperarticulation to computers, as well as on their interrelation.

During global utterance repairs, an error correction subdialogue was initiated that also led to prosodic changes. Repeat utterances were 9% more likely to be closed with a final falling contour than were original utterances. Pitch minima also decreased significantly during global utterance repairs, although only by −2% overall. Both this increased rate of final falling tones on error correction subdialogues, and the small decline in pitch, apparently were used by speakers as cues to mark the close of a repair with their computer partner. These findings are consistent with previous research demonstrating that a final falling contour and reduction in pitch are the strongest cues used during interpersonal speech to produce finality judgements (Swerts *et al.*, 1994).

While amplitude increases were present during global utterance repairs, they nonetheless were negligible—averaging just +0.5%. In a previous study, no amplitude increases were found at all in speech during error resolution (Oviatt *et al.*, 1996, 1998). The statistically reliable amplitude effect in this study most likely was discernable because the data set was three fold larger, and the present experimental design afforded greater precision. In any event, the amplitude change observed in speech to computers was extremely small. This stands in contrast to the sizable increases often found in hyperarticulated speech between humans—for example, in speech to the hearing impaired and in a noisy environment. In summary, adaptation in both amplitude and fundamental frequency were relatively attenuated during error resolution with a computer partner, compared with the effects typically observed between humans during miscommunication.

The hyperarticulation profile described above was strikingly similar following all three types of system recognition error. Irrespective of the fact that users view substitution errors as interjecting *wrong* content, hyperarticulate change following both types of substitution error replicated the pattern found for failure to understand errors. Likewise, unrelated substitution errors were unintuitive, comical, and unique in their ability to evoke emotional reactions 22% of the time (e.g., ''Nancy Alston'' recognized as ''Dodge City''). Although one might assume that this emotional arousal would be associated with a larger magnitude of hyperarticulate change, including heightened pitch and amplitude changes, this was not the case. In spite of their evocative nature, the speech signal adapted nearly identically for unrelated substitution errors as the other two types. This striking similarity in the hyperarticulation profile for different types of system error is illustrated in Fig. 7.

Compared with interpersonal speech during expected or actual miscommunication, the overall pattern of hyperarticulation to a computer is somewhat unique. This partly was evident in users' minimal amplitude and pitch changes, which was consistent with self-reports indicating that speakers generally did not believe that volume or pitch were key factors in eliminating recognition errors. Instead, users reported that controlling rate and articulatory clarity caused computer errors to resolve—comments that corresponded with dominant changes observed in their speech at the signal level. In this sense, speakers' beliefs about rate and articulatory clarity appear to apply more broadly to resolving miscommunications with both computers and varied human listeners. The present evidence supports the view that speakers view error-prone computers as a unique kind of ''at risk'' listener—one involving communication dynamics and sources of fallibility distinct from other at risk groups such as children, the hearing impaired, or nonnative speakers.

The hyperarticulate signal changes reported in this study represent a strong and persistent predilection by speakers. They may underestimate changes during interaction with some challenging application domains that are known to have high word error rates, such as the DARPA Switchboard corpus (Martin *et al.*, 1997). The Switchboard corpus contains speech from spontaneous telephone dialogues, and the best systems currently are generating word error rates two- to three fold higher on this corpus than that in the present study. For systems or application domains known to have such high error rates, previous research indicates that speech is likely to involve a substantial intensification of hyperarticulate effects (Oviatt *et al.*, 1998).

B. Focal hyperarticulation to computers

Since the majority of speech recognition errors are substitutions, sometimes cases arise in which the user selectively repairs one focal part of an utterance, as in ''July twenty-**first** nineteen ninety-seven.'' There is a sense in which these focal repairs may be viewed as a highly targeted, brief, and fine-tuned form of hyperarticulate adaptation in which durational, fundamental frequency, and amplitude cues function together to demarcate and highlight the repair region. Results from the present study clarify the nature and orchestration of hyperarticulate change during error resolution involving focal repairs.

Changes in pause structure still were dominant during focal hyperarticulate adaptation, as summarized in Table III—with a +149% increase in pause duration, and a +113% increase in pause interjection next to the repair region. However, the magnitude of these changes was three- to fourfold larger than during global utterance repair. Changes in pause interjection and elongation also were highly selective in their placement immediately before and after the focal repair region. In fact, these pause changes were twelve- to sixteen-fold greater next to the repair region than in other sentence positions. The function of this selective interjection and lengthening of pauses was most plausibly to demarcate the repair region clearly. However, there was no evidence that such pauses were placed in advance of the repair region more often than after it, for example as a way to signal upcoming repair. The focal speech region also was elongated by 18%, which was twofold more than speech elongation in surrounding nonfocal regions or speech elongation during global utterance repairs.

Although relatively smaller in magnitude of change, the focal repair region also was selectively marked with an +11% increase in pitch range, which derived from increases in maximum pitch in sentence-initial and medial positions, and decreases in minimum pitch in sentence-final position. Variation in absolute pitch levels were revealed to be highly sensitive to the location of a repair in the sentence. However, the net effect of this orchestration of maximum and minimum pitch changes was a uniform expansion of pitch range on focal repairs occurring anywhere in a sentence. As in the case of durational effects, pitch changes observed during focal repairs were highly targeted at the repair region. On average, there was a +38% greater expansion of pitch range on the focal speech repair than on surrounding nonfocal speech segments.

Expanded pitch range is known to mark linguistic segments as salient (Pierrehumbert, 1980), or as content that the listener should pay particular attention to in the moment-by-moment delivery of spoken information. Pitch range also is known to play an important role in conveying the hierarchical segmentation of discourse, generally being expanded at the beginning of new topics (Brown, 1983; Hirschberg and Grosz, 1992; Lehiste, 1975). In spontaneous conversations, pitch range expansion generally has been shown to mark the start of a new unit, whether a new topic, a new speaker turn, or a self-correction of disfluencies or content errors (Ayers, 1994; French and Local, 1986). During focal error repairs with a computer partner, both elevated pitch and expanded pitch range provided cues for identifying the precise boundaries of the correction region within a longer continuous utterance, which could facilitate linguistic processing of its lexical content. As a tool for demarcating focal repair regions, pitch clearly functioned more actively than during global utterance repairs.

The focal repair region also was selectively marked with small increases in amplitude, averaging less than a 1% gain. Although change in amplitude co-occurred with durational effects, increases in duration far exceeded the relative gains for amplitude. This finding is consistent with Turk and Sawusch's (1996) demonstration that, while duration and amplitude generally interact to yield judgements of prominence (Fry, 1955), the primary factor that gives rise to perceived prominence is increased duration. Their research demonstrates that the impact of durational and amplitude increases on the perceived salience of a speech segment are not equivalent or symmetric. At some level, speakers may be aware of this greater impact of durational increase on the intelligibility of speech, which may account for their similarly strong reliance on durational cues when resolving system recognition errors.

TABLE IV. Summary of absolute change in linguistic features of Stage I and II hyperarticulation,[a] based on past and present research.[b]

Linguistic feature	Stage I change[c]	Stage II change
Duration:		
Pause interjection	+0.57 pauses	+0.32 — +0.38 pauses[d]
Pause elongation	+97 ms	+78 — +102 ms
Speech elongation	+190 ms	+127 — +171 ms
Articulation:		
Hyper-clear phonology	N.S.	+6 — +9%[e]
Disfluencies	N.S.	−0.25 — −0.25[f]
Pitch:		
Intonation—final fall	N.S.	+9 — +9%[g]
Pitch minmum	N.S.	−2.2 — −2.7 Hz
Amplitude:		
Amplitude maximum	N.S.	N.S./+0.3 dB

[a] Values listed represent absolute change from original to repeat input for statistically significant changes (N.S.=not significant).

[b] Cumulative data included from past and present research are indicated in regular and bold font, respectively. Values based on the present research are averages across all error types. Values based on past findings are taken from Oviatt *et al.* (1998).

[c] Stage I changes were associated with a 6.5% overall error rate per utterances input, and Stage II changes with a 20% rate (upper bounds of the Stage II range based on spiral errors that repeated 1–6 times).

[d] Data represent change in average number of pauses per utterance in multiword utterances.

[e] Data represent change in percent of utterances with a phonological alternation involving a hyperarticulate shift.

[f] Data represent change in rate of disfluencies per 100 words.

[g] Data represent change in percent of utterances with a final falling intonation contour.

C. The CHAM model

These results corroborate and generalize the computer-elicited hyperarticulate adaptation model (CHAM), which is summarized schematically in Fig. 1 and elaborated quantitatively in the accompanying Table IV. The CHAM model predicts that specific features in users' speech will adapt during human–computer error resolution, and that the type and magnitude of adaptation will depend on a system's overall error rate (Oviatt *et al.*, 1998). In the present study, the hyperarticulate changes that were replicated across all three error types would be considered Stage II adaptations, and in fact the predicted multiple effects involving durational, articulatory, fundamental frequency, and amplitude changes all were evident (i.e., see Table IV values in bold font). Although no change in amplitude was reported in earlier findings by Oviatt *et al.* (1998), in the present study which was threefold larger and more carefully controlled, a significant but very small amplitude effect did emerge. As clarified by Table IV, the magnitude of adaptations for specific linguistic features in the present study was extremely close to previous reports. In addition to the above, the CHAM model predicts abrupt transitions in the signal profile from one moment to the next, which was observed continually in this study when brief episodes of hyperarticulation punctuated repetitions in juxtaposed original-repeat utterances.

Table IV summarizes the type and magnitude of absolute hyperarticulate changes during Stage I and II based on cumulative evidence from past research reported in Oviatt *et al.* (1998) (i.e., shown in plain font) and from the present findings (i.e., shown in bold font). Results from the earlier study by Oviatt *et al.* (1998) included data on Stage I and II hyperarticulation elicited by rejections errors. In contrast, the larger and more extensive present study included data on three common types of system recognition error, as well as on focal and global utterance repairs, although these latter comparative data all assessed Stage II hyperarticulation. The Stage I hyperarticulation data listed in Table IV were precipitated by a low error rate (i.e., 6.5%), whereas Stage II data were associated with a high error rate (i.e., 20%). The hyperarticulation values from the present study that are listed in Table IV tend to mark the lower bound on Stage II estimates, with Stage II values based on previous research ranging slightly but consistently higher because they involved spiral errors that could recur between one and six times. These spiral errors effectively would have compounded the error rate, which could account for the correspondingly greater changes in hyperarticulate features and would be consistent with the CHAM model.

With respect to hyperarticulate change during focal repairs, the acoustic dimensions that were examined—including duration, pitch, and amplitude—all adapted as predicted by the CHAM model. Furthermore, the relative degree of change in these three dimensions (i.e., large changes in duration, moderate ones in pitch, and minimal ones in amplitude) are similar to those observed during global error repairs. However, the absolute magnitude of durational and pitch range changes during focal repairs was larger than that found during global repairs. In addition, shifting to and from a hyperarticulate speech style was more abrupt and highly targeted than that during global utterance repairs. While consistent with the CHAM model, these characteristics of hyperarticulation during focal repairs may prove more difficult to accommodate in the design of future systems, as will be discussed further in the next section.

In brief, the present results confirm and further generalize the two-stage CHAM model, which was motivated by linguistic theory (Lindblom, 1990) and the specifics of which were derived from recent empirical research (Oviatt *et al.*, 1998). From cumulative research conducted to date, it is clear that Stage I and II of the CHAM model accurately predict the type and magnitude of hyperarticulate adaptations for a variety of linguistic features during human–computer error resolution, which vary according to a system's overall error rate. As demonstrated in the present study, the CHAM model's basic predictions apply to qualitatively different types of recognition error, and to both global and focal utterance repairs.

D. 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 and complicate their ability to resolve errors gracefully. One question raised by viewing the CHAM model in Fig. 1 is whether an utterance spoken during baseline conditions can be recognized as identical to its counterpart during Stage II conditions. Like Lombard speech, hyperarticulate speech involves episodic and often abrupt signal variability that may pose a more substantial challenge to current recognition technology than chronic forms of variability, such as accented speech. The relatively static algorithmic 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 provide a stimulus for developing fundamentally more dynamic, adaptive, and user-centered approaches to speech recognition.

There are several possible avenues for improving the performance of current spoken language systems on hyperarticulate speech. One is to train recognizers on more natural samples of users' interactive speech to systems, including error resolution with the type and baserate of errors expected in the target system. However, this alternative may be associated with trade-offs in accuracy, and it does not address the problematic issue of abrupt signal transitions in hyperarticulate speech.

Another 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. Such an alternative would be viable within a form-based interface with input slots, as was used in the present simulation, since in such an arrangement it is reasonable to assume that reentry into the same slot involves a correction. This approach would require data collection and recognizer training on a corpus of hyperarticulate speech. One advantage of this approach is that it is capable of handling abrupt shifts in hyperarticulation. However, not all applications may be amenable to identifying the start and end of error correction, which would be necessary to swap in the appropriate recognizer reliably.

Although hyperarticulate changes during focal error repairs were similar to those during global repair, in some respects they may be more difficult for systems to accommodate. For example, the durational and pitch range changes during focal repairs were more pronounced in magnitude, and shifts to and from hyperarticulate speech were more abrupt than during global repairs. One difficult problem raised by these data on focal repairs is how to identify their precise boundaries in a continuous utterance. This problem complicates the prospect of designing systems with specialized recognizers, as suggested above. In particular, it may be implausible in future systems to mark focal repair regions clearly via simple interface design techniques, for example, using a form-based interface to swap in a specialized recognizer at appropriate times. However, since strong acoustic cues naturally demarcate focal repairs, in the future it may be possible to develop methods for identifying focal repair regions automatically as an aid to advanced interface design.

The development of more adaptive systems likewise may improve current recognizer's performance, and is an option that has been advocated for processing Lombard speech (Applebaum and Hanson, 1990; Junqua, 1993). Since signal adaptations occur abruptly when users enter an error correction subdialogue, such a system should *not* be designed to adapt continuously to users' speech throughout an interaction. Rather, system adaptation specifically should avoid adapting 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 recognizer performance on a user's hyperarticulation during future correction episodes. To better assess the prospects and benefits of an adaptive approach, future research should explore individual differences in hyperarticulate speech, especially for durational effects (for discussion, see Oviatt *et al.*, 1998).

Perhaps the most promising long-term solution to improving current recognizers' performance is to avoid hyperarticulate speech by designing a multimodal rather than unimodal interface. This option has been discussed in detail elsewhere (Oviatt and vanGent, 1996; Oviatt *et al.*, in press), so will only be summarized here. First, when people are free to interact multimodally and can switch to an alternate input mode, the likelihood of both avoiding and rapidly resolving errors is facilitated. In part, this is because users have good intuitions about when to deploy a given input mode such that they avoid errors (Oviatt and Olsen, 1994). In addition, users naturally increase their alternation of input modes after a recognition error occurs. Since input modes such as speech and pen have different confusion matrices associated with the same propositional content, this switching of input modes in a multimodal interface can 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, which then reduces the overall system's error rate (Johnston *et al.*, 1997; Oviatt, in press; Oviatt, in submission).

In the near future, it will become increasingly important to model speech 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, acoustic-phonetic variability in the speech signal may be different and substantially magnified under such conditions. Rates of miscommunication also are likely to be elevated, in some cases beyond those currently reported for spontaneous telephone dialogues. Unlike the laboratory, speech in these settings can be expected to include a combination of hyperarticulate, Lombard, and other difficult forms of abrupt signal variation. The present research on user-centered modeling of speech adaptations during error begins to provide an empirical foundation for the design of these more challenging next-generation systems.

ACKNOWLEDGMENTS

This research was supported by Grant No. IRI-9530666 from the National Science Foundation, and by grants, contracts, and equipment donations from Apple, GTE Labs, Intel, Microsoft, NTT Data, Southwestern Bell, and US West. We especially thank Robert VanGent, Jon Lindsay, and Eric Iverson for adapting the simulation software to support this study, and Eric Iverson for acting as simulation assistant during data collection. Thanks to Karen Kuhn for preparing detailed hard-copy transcripts, and to Karen Kuhn and David Fencsik for scoring assistance. Finally, we are grateful to the people who generously volunteered their time to participate in this research.

[1] The distinction between related and unrelated substitution errors refers to whether the system's substitution is related acoustically and semantically to the user's original input.

[2] In current systems, the percentage of substitution errors that would be classified as related versus unrelated varies considerably. As recognition technology develops algorithms that more closely model human perceptual and linguistic capabilities, an increasingly greater percentage should become related acoustically and semantically to users' input.

Applebaum, T. H., and Hanson, B. A. (**1990**). ''Robust speaker-independent word recognition using spectral smoothing and temporal derivatives,'' Proc. of EUSIPCO-90, 1183–1186.

Ayers, G. (**1994**). ''Discourse functions of pitch range in spontaneous and read speech,'' in *Working Papers in Linguistics* (Ohio State University, Columbus, OH), Vol. 44, pp. 1–49.

Bolinger, D. (**1958**). ''A theory of pitch accent in English,'' Word **7**, 199–210.

Bond, Z. S., and Moore, T. J. (**1994**). ''A note on the acoustic-phonetic characteristics of inadvertently clear speech,'' Speech Commun. **14**, 325–337.

Brenner, M., Shipp, T., Doherty, E., and Morrissey, P. (**1985**). ''Voice measures of psychological stress: Laboratory and field data,'' in *Vocal Fold Physiology, Biomechanics, Acoustics, and Phonatory Control*, edited by I. Titze and R. Scherer (Denver Center for the Performing Arts, Denver, CO), pp. 239–248.

Brown, G. (**1983**). ''Prosodic structure and the given/new distinction,'' in *Prosody: Models and Measurements*, edited by D. R. Ladd and A. Cutler (Springer-Verlag, Berlin), pp. 67–68.

Brown, N. R., and Vosburgh, A. M. (**1989**). ''Evaluating the accuracy of a large vocabulary speech recognition system,'' Proc. of the 33rd Annual Meeting of the Human Factors Society, pp. 296–300.

Chen, F. R. (**1980**). ''Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level,'' Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Cutler, A., and Butterfield, S. (**1990**). ''Durational cues to word boundaries in clear speech,'' Speech Commun. **9**, 485–495.

Cutler, A., and Butterfield, S. (**1991**). ''Word boundary cues in clear speech: A supplementary report,'' Speech Commun. **10**, 335–353.

de Jong, K. (**1995**). ''The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation,'' J. Acoust. Soc. Am. **97**, 491–504.

Eisen, B., Tillmann, H. G., and Draxler, C. (**1992**). ''Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases,'' Proc. of the Int. Conf. on Spoken Language Processing **2**, 871–874.

Ferguson, C. A. (**1975**). ''Toward a characterization of English foreigner talk,'' Anthropological Linguistics **17**, 1–14.

Ferguson, C. A. (**1977**). ''Baby talk as a simplified register,'' in *Talking to Children: Language Input and Acquisition*, edited by C. E. Snow and C. A. Ferguson (Cambridge U.P., Cambridge, MA), pp. 219–36.

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

Frankish, C., Hull R., and Morgan, P. (**1995**). ''Recognition accuracy and user acceptance of pen interfaces'' Proc. of the Conf. on Human Factors in Computing Systems—CHI'95, pp. 503–510.

Freed, B. F. (**1978**). ''Foreign talk: A study of speech adjustments made by native speakers of english in conversation with nonnative speakers,'' Doctoral dissertation, Linguistics Department, University of Pennsylvania.

French, P., and Local, J. (**1986**). ''Prosodic features and the management of interruptions,'' in *Intonation in Discourse*, edited by C. Johns-Lewis (College-Hill, San Diego, CA), pp. 157–180.

Fry, D. B. (**1955**). ''Duration and intensity as physical correlates of linguistic stress,'' J. Acoust. Soc. Am. **27**, 765–769.

Fry, D. B. (**1958**). ''Experiments in the perception of stress,'' Language and Speech **1**, 126–152.

Garnica, O. K. (**1977**). ''Some prosodic and paralinguistic features of speech to young children,'' in *Talking to Children*, edited by C. E. Snow and C. A. Ferguson (Cambridge U.P., Cambridge, MA), pp. 63–88.

Gordon-Salant, S. (**1987**). ''Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects,'' J. Acoust. Soc. Am. **81**, 1199–1202.

Hanley, T. D., and Steer, M. D. (**1949**). ''Effect of level of distracting noise upon speaking rate, duration and intensity,'' J. Speech Hear. Disord. **14**, 363–368.

Hirschberg, J., and Grosz, B. (**1992**). ''Intonational features of local and global discourse structure,'' Proc. of the 5th DARPA Speech and Natural Language Processing Workshop, 441–446.

Howell, P., and Young, K. (**1991**). ''The use of prosody in highlighting alterations in repairs from unrestricted speech,'' The Quarterly J. of Experimental Psychology **43A**, 733–758.

Jelinek, F. (**1985**). ''The development of an experimental discrete dictation recognizer,'' Proc. IEEE **73**, 1616–1624.

Johnston, M., Cohen, P., McGee, D., Oviatt, S., Pittman, J., and Smith, I. (**1997**). ''Unification-based multimodal integration,'' Proc. of the 35th Annual Meeting of the Association for Computational Linguistics, 281–288.

Junqua, J. C. (**1993**). ''The lombard reflex and its role on human listeners and automatic speech recognizers,'' J. Acoust. Soc. Am. **93**, 510–524.

Kamm, C. A. (**1994**). ''User interfaces for voice applications,'' in *Voice Communication Between Humans and Machines*, edited by D. B. Roe and J. Wilpon (National Academy, Washington, DC), pp. 422–442.

Lehiste, I. (**1975**). ''The phonetic structure of paragraphs,'' in *Structure and Processes in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer-Verlag, Berlin), pp. 195–203.

Levelt, W. J., and Cutler, A. (**1983**). ''Prosodic marking in speech repair,'' J. Semantics. **2**, 205–217.

Lewis, C., and Norman, D. A. (**1986**). ''Designing for error,'' in *User-Centered System Design*, edited by D. A. Norman and S. W. Draper (Erlbaum, Hillsdale, NJ), pp. 411–432.

Lindblom, B. (**1990**). ''Explaining phonetic variation: A sketch of the H and H theory,'' in *Speech Production and Speech Modeling*, edited by W. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–439.

Lindblom, B. (**1996**). ''Role of articulation in speech perception: Clues from production,'' J. Acoust. Soc. Am. **99**, 1683–1692.

Lindblom, B., Brownlee, S., Davis, B., and Moon, S. J. (**1992**). ''Speech transforms,'' Speech Commun. **11**, 357–368.

Lively, E., Pisoni, D. B., Van Summers, W., and Bernacki, R. (**1993**). ''Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences,'' J. Acoust. Soc. Am. **93**, 2962–2973.

Lombard, E. (**1911**). ''Le signe de l'elevation de la voix,'' Annals Maladiers Oreille, Larynx, Nez, Pharynx **37**, 101–119.

Maasen, B. (**1986**). ''Marking word boundaries to improve the intelligibility of the speech of the deaf,'' J. Speech Hear. Res. **29**, 227–230.

Martin, A., Fiscus, J., Fisher, B., Pallett, D., and Przybocki, M. (**1997**). ''System descriptions and performance summary,'' Proc. of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation.

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

Moon, S. J., and Lindblom, B. (**1994**). ''Interaction between duration, context, and speaking style in English stressed vowels,'' J. Acoust. Soc. Am. **96**, 40–55.

Moreton, E. (**1996**). ''Scoring procedure for duration,'' Oregon Graduate Institute of Science & Technology, unpublished manuscript.

Nakatani, C. H., and Hirschberg, J. (**1994**). ''A corpus-based study of repair cues in spontaneous speech,'' J. Acoust. Soc. Am. **95**, 1603–1616.

O'Shaughnessy, D. (**1992**). ''Analysis of false starts in spontaneous speech,'' Proc. of the Int. Conf. on Spoken Language Processing **1**, 931–934.

Oviatt, S. L. (**1995**). ''Predicting spoken disfluencies during human-computer interaction,'' Comput. Speech Lang. **9**, 19–35.

Oviatt, S. L. (**in press**). ''Ten myths of multimodal interaction,'' in *Communications of the ACM.*

Oviatt, S. L. (**in submission**). ''Mutual disambiguation of recognition errors in a multimodal architecture.''

Oviatt, S. L., and Olsen, E. (**1994**). ''Integration themes in multimodal human–computer interaction,'' Proc. of the Int. Conf. on Spoken Language Processing **2**, 551–554.

Oviatt, S. L., and VanGent, R. (**1996**). ''Error resolution during multimodal human–computer interaction,'' Proc. of the Int. Conf. on Spoken Language Processing **1**, 204–207.

Oviatt, S. L., Bernard, J. and Levow, G. (**in press**). ''Linguistic adaptations during error resolution with spoken and multimodal systems,'' Language and Speech (in press).

Oviatt, S. L., Cohen, P. R., Fong, M. W., and Frank, M. P. (**1992**). ''A rapid semi-automatic simulation technique for investigating interactive speech and handwriting,'' Proc. of the Int. Conf. on Spoken Language Processing **2**, 1351–1354.

Oviatt, S. L., Levow, G., MacEachern, M., and Kuhn, K. (**1996**). ''Modeling hyperarticulate speech during human-computer error resolution,'' Proc. of the Int. Conf. on Spoken Language Processing, **2**, 801–804.

Oviatt, S. L., MacEachern, M., and Levow, G. (**1998**). ''Predicting hyperarticulate speech during human-computer error resolution,'' Speech Commun. **24**, 1–23.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). ''Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,'' J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). ''Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech,'' J. Speech Hear. Res. **28**, 96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). ''Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech,'' J. Speech Hear. Res. **29**, 434–446.

Pierrehumbert, J. (**1980**). ''The phonology and phonetics of English intonation,'' Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Rhyne, J. R., and Wolf, C. G. (**1993**). ''Recognition-based user interfaces,'' in *Advances in Human–Computer Interaction*, Vol. 4, edited by H. R. Hartson and D. Hix (Ablex Publishing, Norwood, NJ), pp. 191–250.

Schulman, R. (**1989**). ''Articulatory dynamics of loud and normal speech,'' J. Acoust. Soc. Am. **85**, 295–312.

Shriberg, E., Wade, E. and Price, P. (**1992**). ''Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction,'' Proc. of the DARPA Speech and Natural Language Workshop, 49–54.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). ''Effects of noise on speech production: Acoustic and perceptual analyses,'' J. Acoust. Soc. Am. **84**, 917–28.

Swerts, M., Bouwhuis, D. G., and Collier, R. (**1994**). ''Melodic cues to the perceived finality of utterances,'' J. Acoust. Soc. Am. **96**, 2064–2075.

Tolkmitt, E. J., and Scherer, K. R. (**1986**). ''Effect of experimentally induced stress on vocal parameters,'' J. Exp. Psychol. **12**, 302–312.

Turk, A. E., and Sawusch, J. R. (**1996**). ''The processing of duration and intensity cues to prominence,'' J. Acoust. Soc. Am. **99**, 3782–3790.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (**1996**). ''Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate,'' J. Speech Hear. Res. **39**, 494–509.

Williams, C. E., and Stevens, K. N. (**1969**). ''On determining the emotional state of pilots during flight: An exploratory study,'' Aerosp. Med. **40**, 1369–1372.