MODELING HYPERARTICULATE SPEECH DURING HUMAN-COMPUTER ERROR RESOLUTION*

*Sharon Oviatt,*** *Gina-Anne Levow, Margaret MacEachern & Karen Kuhn*

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

ABSTRACT

Hyperarticulate speech to computers remains a poorly understood phenomenon, in spite of its association with elevated recognition errors. The present research analyzes the type and magnitude of linguistic adaptations that occur when people engage in error resolution with computers. A semi-automatic simulation method incorporating a novel error generation capability was used to collect speech data immediately before and after system recognition errors, and under conditions varying in error base-rates. Data on original and repeated spoken input, which were matched on speaker and lexical content, then were examined for type and magnitude of linguistic adaptations. Results indicated that speech during error resolution primarily was longer in duration, including both elongation of the speech segment and substantial relative increases in the number and duration of pauses. It also contained more clear speech phonological features and fewer spoken disfluencies. Implications of these findings are discussed for the development of more user-centered and robust error handling in next-generation systems.

1. INTRODUCTION

One aim of recent research on spoken language interfaces has been to identify hard-to-process sources of linguistic variability in spontaneous speech, and to develop predictive models accounting for these phenomena and corresponding interface techniques for reducing their occurrence (Oviatt, Cohen & Wang, 1994; Oviatt, 1995). Although hyperarticulate speech to computers has been noted informally and has been associated with significantly elevated recognition errors (Shriberg, Wade & Price, 1992), nonetheless to date it remains an ill-defined concept. In particular, it is unclear what the definition of hyperarticulation is in the context of human-computer interaction in terms of the type and magnitude of linguistic adaptations that actually occur, especially during error resolution. Furthermore, as a potentially difficult source of variability, its impact on degradation of speech recognition rates is poorly understood. If people do typically hyperarticulate while trying to resolve system errors, then recognition rates would be expected to degrade as hyperarticulated speech departs further from the original training data upon which a recognizer was developed. To our knowledge, current speech recognizers invariably are trained on original input, often collected under constrained or unnatural task conditions, but omit any training on repetitions during simulated or actual error conditions. Essentially, the widely-advocated concept of "designing for error" (Lewis & Norman, 1986) has not been applied effectively to the design of spoken language systems, even though many researchers and corporate designers regard error resolution to be the most challenging interface problem facing this technology (Rhyne & Wolf, 1993).*

Although literature on hyperarticulate speech during human-computer error resolution currently is lacking, some guidance is available from related research on how people adapt their speech during human-human exchanges when they expect or experience a comprehension failure from their listener. Systematic modifications have been documented in parents' speech to infants and children (Fernald et al., 1989) and in speech to the hearing impaired (Picheny et al., 1986), although the specific accommodations differ depending on the target population. For example, spoken adaptations to infants include elevated pitch, expanded pitch range, and stress on new vocabulary content, phenomena that subserve attentional and teaching functions. Speech to hearing-impaired individuals involves increased volume, duration, and a shift from conversational to "clear speech" articulatory patterns. At present, the profile of hyperarticulatory adaptations that may occur during human-computer interaction simply is not known.

The goal of the present research was to identify the type and magnitude of linguistic adaptations that occur during human-computer interactions involving error resolution. It was hypothesized that users' repetitions during error resolution would be delivered to *contrast* with their original spoken input along selected linguistic dimensions, and also would be adapted toward *clear speech* acoustic/phonetic features. More specifically, in the present study within-subject data were examined for possible systematic adaptation toward longer duration, greater acoustic intensity, and increased maximum frequency and frequency range during error resolution. When speakers repeated the same lexical content after a simulated recognition error, it also was predicted that clear speech phonological features would increase and disfluencies correspondingly decrease, with these effects magnified during a high error base-rate versus a low one. The long-term goal of this research is the development of a user-centered predictive model of linguistic adaptation during human-computer error resolution, as well as the development of improved error handling capabilities for advanced recognition-based interfaces.**

2. METHOD

2.1. Subjects, Tasks, and Procedure

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

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

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

** First author: Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science & Technology, P.O. Box 91000, Portland, OR, 97291 (oviatt@cse.ogi.edu; http://www.cse.ogi.edu/~oviatt/) Collaborators' respective affiliations: Laboratory for Computer Science, MIT; Linguistics Department, UCLA; Linguistics Department, Portland State University.

form displayed on a Wacom LCD tablet. As input was received, the system interactively confirmed the propositional content of requests by displaying typed feedback in the appropriate input slot.

For example, if the system prompted with **Car pickup location:__________** and a person spoke "San Francisco airport," then "**SFO**" was displayed immediately after the utterance was completed. In the case of simulated errors, the system instead responded with "**????**" feedback to indicate its failure to recognize input. In this case, subjects were instructed to re-enter their information into the same slot until system feedback was correct. Each simulated error required 1-6 repeats before error resolution was successful, thereby simulating *spiraling* in recognition-based systems. A form-based interface was used during data collection so that the locus of system errors would be clear to users. They were told that the system was a well-developed one with extensive processing capabilities, so they could speak normally, express things as they liked, work at their own pace, and just concentrate on completing their job.

2.2. Semi-automatic Simulation Method

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

2.3. Research Design

The research design was a within-subject factorial that included the following independent variables: (1) Error status of speech (Original input; Repeat input after error), (2) Base-rate of system errors (Low- 6.5% of input slots; High- 20% of slots). All 20 subjects completed 12 subtasks, half involving a low base-rate of errors and half a high one, with the order counterbalanced across subjects. In total, data were collected on 480 simulated errors, of which over 250 involved the same speaker repeating identical lexical content during the first repetition of a repair attempt. For these matched utterance pairs, original input provided a baseline for assessing and quantifying the degree of change along linguistic dimensions of interest.

2.4. 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 original and first repeat utterance pairs 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. For the present acoustic/prosodic and phonetic analyses, only the first repair was compared with original input, although disfluency rates were based on all spoken repetitions that occurred during error resolution.

Duration. The following were summarized: (1) total utterance duration, (2) total speech segment duration (i.e., total duration minus pause duration), (3) total pause duration for multi-word utterances, and (4) total number of pauses for multi-word utterances. No attempt was made to code pauses less than 10 msec in duration. Due to difficulty locating their onset, utterance-initial voiceless stops and affricates were arbitrarily assigned a 20-msec closure, and no pauses were coded as occurring immediately before utterance-medial voiceless stops and affricates.

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

Fundamental Frequency. Spoken input was coded for maximum F0, minimum F0, F0 range, and F0 average. The fundamental frequency tracking software in ESPS Waves+ was used to calculate values for voiced regions of the digitized speech signal. Pitch minima and maxima were calculated automatically by program software, and then adjusted to correct for pitch tracker errors such as spurious doubling and halving, interjected non-speech sounds, and extreme glottalization affecting ^{2} 5 tracking points. To avoid skewing due to line noise, only voiced speech within the coder-corrected F0 range were used to calculate F0 mean.

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. Phenomena considered difficult to code reliably were not included, such as glottalization and glottal stop insertion.

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

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

3. RESULTS

3.1. Duration

Spoken utterances in this corpus tended to be brief fragments averaging two words, and ranging from 1-13 words in length. When the error rate was low, total utterance duration averaged 1544 msec and 1802 msec during original and repeat input, a gain of 16.5%, a significant increase by paired t test on log transformed data, $t = 7.05$ ($df = 49$), $p < .001$, one-tailed. When the base-rate of errors was high, the total utterance duration averaged 1624 msec during original input, increasing to 1866 msec during repeat input, a gain of 15%, which again was significant by paired t test on log transformed data, $t = 10.71$ ($df = 208$), $p < .001$, one-tailed.

Speech Segment Duration. Analyses revealed an increase in the total speech segment from an average of 1463 msec during original input to 1653 msec during repeat input when the error rate was low, a 13% gain, significant by paired t test on log transformed data, t = 7.44 (df = 50), p < .001, one-tailed. During a high error-rate, it also increased from 1515 msec during original input to 1686 msec during repeat input, an 11.5% gain, significant by paired t test on log transformed data, t = 10.20 (df = 215), p < .001, one-tailed.

Pause Duration. The total pause duration of multi-word utterances increased significantly from an average of 112 to 209 msec between original and repeat input when the error rate was low, an 86.5% gain, significant by paired t test on log transformed data, t = 1.89 (df = 11), p < .05, one-tailed, and it again increased significantly from an average of 159 msec during original input to 261 msec during repeat input when the error rate was high, a 64% gain, significant by paired t test on log transformed data, t = 3.17 (df = 36), p < .002, one-tailed.

Number of Pauses. The average number of pauses per multi-word utterance also increased significantly from an average of 0.49 during original input to 1.06 during repeated speech when the error rate was low, a 116% gain, significant by Wilcoxon Signed Ranks test, z = 2.52 (N = 12), p < .006, one-tailed. During high error rates, it increased from 0.57 to 0.95 during repetitions, or 67%, significant by Wilcoxon, z = 3.03 (N = 16), p < .001, one-tailed. Figure 1 illustrates the average relative gains in pause duration (73%) and pause interjection (91%) in relation to average segment elongation (12%) in two typical utterances from the corpus. To test for elongation of individual pauses (i.e., independent of interjecting new ones), original and repeat utterance pairs matched on total number of pauses were compared for total pause length. This analysis confirmed that pauses were elongated significantly more in repeat utterances, paired t = 1.71, (df = 27), p < .05, one-tailed.

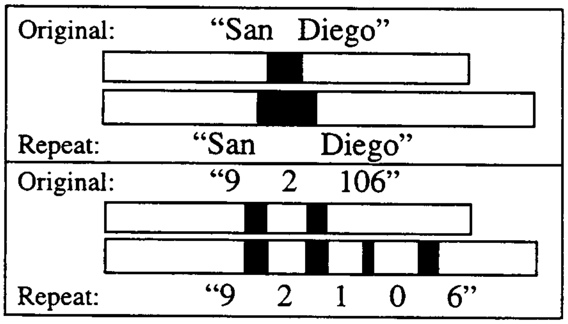

Figure 1: Pause elongation (top) and pause interjection (bottom) in matched original-repeat utterance pairs

3.2. Intensity/Amplitude

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

3.3. Fundamental Frequency

Pitch Maximum. The maximum F0 did not differ significantly between original and repeat input in either the low or high error conditions, by paired t test on log transformed data, t = 1.58 (NS, one-tailed) and t = 1.35 (NS), respectively. Reanalysis subdivided by gender confirmed this lack of a significant change in maximum F0 in both females and males.

Pitch Minimum. The minimum F0 did not differ in the low error-rate condition, which averaged 111.3 hz and 110.4 hz, paired t < 1. However, minimum F0 dropped between original and repeat input in the high error-rate condition, averaging 122.2 hz and 119.5 hz, paired t test on log transformed data, t = 1.96, (df = 221), p < .05, two-tailed. Although no significant differences were revealed in minimum F0 for female speech, or for male speech when errors were low, minimum F0 dropped marginally in male speech from 94.4 hz to 91.3 hz when the error rate was high, t = 1.86, (df = 123), p < .065, two-tailed.

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

Pitch Average. Average F0 dropped significantly between original and repeat input in the high error-rate condition, averaging 164.7 hz and 162.4 hz, paired t = 2.83 (df = 206), p < .005, two-tailed, although no difference was found in the low error-rate condition. No differences were found in females' speech, or in males' mean F0 when the error rate was low, although it dropped significantly in male speech from 128.9 hz to 126.6 hz during repeat input when the error rate was high, t = 2.47 (df = 108), p < .015, two-tailed.

3.4. Phonological Alternations

Approximately 9% of first repetitions in this corpus contained phonological alternations, and 93% of subjects altered their speech phonologically during error resolution at some point. Table 1 summarizes the number and type of alternations observed for each subject who had a minimum of 12 scorable utterance pairs, as well as their classification by direction of shift with respect to clear speech.

Clear to Conversational	Conversational to Clear	Phonological Alternations
0	1	a
0	1	c
0	1	a
0	2	b, c
0	1	a
0	2	c, c
0	3	b, b, c
0	2	b, b
0	0	
0	1	a
0	3	c, c, c
0	2	b, c
0	5	b, b, c, c, c
1	0	d

Table 1: Number and type of phonological alternations involving a shift toward conversational versus clear speech, listed by subject (**a**- unreleased t > t; **b**- alveolar flap > coronal plosive; **c**- nasal stop or flap > nt sequence; **d**- nt sequence > nasal stop or flap)

The majority of subjects, or 86% who adapted their speech, shifted from a conversational to clear speech style rather than the reverse, a significant difference by Wilcoxon Signed Ranks test, T+ = 87.5 (N= 13), p < .001, one-tailed. After correcting for the difference in total errors sampled, analysis of clear-speech adaptations revealed that they were significantly more prevalent in the high error-rate condition than the low one, as indicated by Wilcoxon Signed Ranks test, T+ = 67 (N = 12), $p < .015$, one-tailed. As shown in Figure 2, a comparison of the average rate of clear-speech adaptations per 100 words revealed a 163% increase, from 0.95 in the low error-rate condition to 2.50 in the high one.

3.5. Disfluencies

The disfluency rate during original spoken input averaged 0.78 disfluencies per 100 words, the same as previously reported for structured interfaces (Oviatt, 1995). However, it dropped significantly to 0.37 when repeating during error resolution, paired t = 2.03 (df = 19), $p < .03$, one-tailed. In the low error-rate condition, disfluencies averaged 0.85 per 100 words, similar to 0.78 previously reported with low errors (Oviatt, 1995). However, it dropped significantly to 0.53 when the error-rate was high, paired t = 1.90 (df = 19), $p < .04$, one-tailed. Figure 2 illustrates the inverse relation between clear-speech phonological alternations and spoken disfluencies as a function of error base-rate. Since the disfluency rate is influenced by both original versus repeat input and by the overall base-rate of errors, and since the high error-rate condition contained 3-fold more errors than the low one, a comparison also was evaluated between the high and low conditions with all errors removed (i.e., comparing only baseline original input). In this more controlled comparison, the disfluency rate was confirmed to decrease significantly when the error-rate became high, paired t = 2.38 (df = 19), $p < .03$, one-tailed.

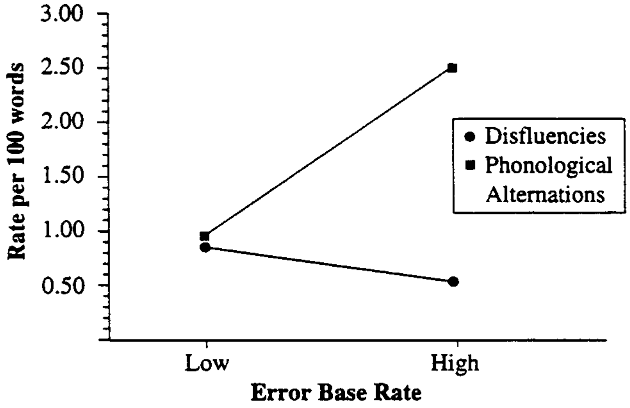

Figure 2: Rate of spoken disfluencies and phonological alternations per 100 words as a function of error base-rate

4. CONCLUSIONS

During error resolution with computers, human speech shifts to become lengthier and more clearly articulated. An increase in total utterance duration was documented during repetitions, including an average 12% elongation in the speech segment, a 73% elongation in pause duration, and interjection of 91% more utterance-internal pauses. Clearly, pause characteristics constituted the most salient relative change during repetitions. In addition, the phonological features of repeat speech adapted toward an audibly clearer articulation pattern. In the present corpus, the most frequently observed changes included fortition of alveolar flaps to coronal plosives (e.g., eɪɾeɪt changing to eɪt˺eɪt), and shifts to unreduced nt sequences (e.g., twɛɾ̃i to twɛnti). This shift to clear speech during error resolution also corresponded with a drop in spoken disfluencies. When the error rate was high, which required correcting 1 in 5 slots rather than 1 in 15, both the increase in clear-speech adaptations and the decrease in disfluencies were accentuated further. Female pitch also was found to expand in range when resolving errors during a chronically high error-rate, an adaptation also found in female speech to children. However, unlike some error resolution between humans, speakers did not alter their volume when resolving errors with the computer.

The hyperarticulate speech documented in this research presents a potentially difficult source of variability that may degrade the performance of current speech recognizers, in particular complicating recognizers' ability to resolve errors gracefully. This research has implications for the development of more user-centered recognition algorithms, the collection of more realistic speech data with interactive systems varying in error base-rate, and the design of recognizers specialized for error handling that may become part of a coordinated system of multiple recognizers. Further work is needed on quantitative modeling of the durational and articulatory phenomena identified in this study, which could contribute to establishing more user-centered and robust next-generation spoken language systems. Additional research also is needed on spoken adaptations following other types of recognition errors, such as substitutions, to assess the generality of findings identified in this work. Finally, the fuller spectrum of error handling options and potential benefits needs to be explored when speech is incorporated as one of several input modes in a multimodal interface (see Oviatt & VanGent, 1996, this volume).

5. REFERENCES

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

2. Lewis, C. & Norman, D. A. Designing for error, in *User-Centered System Design* (ed. by D. A. Norman & S. W. Draper), Lawrence Erlbaum: Hillsdale, N. J., 1986, 411-432.

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

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

5. Oviatt, S. L., Cohen, P. R. & Wang, M. Q. Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity, *Speech Communication*, European Speech Communication Association, 1994, vol. 15, nos. 3-4, 283-300.

6. Oviatt, S. L. & VanGent, R. Error resolution during multimodal human-computer interaction, *Proceedings of the International Conference on Spoken Language Processing*, 1996, in press.

7. Picheny, M. A., Durlach, N. I. & Braida, L. D. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech, *Journal of Speech and Hearing Research, 29*, 1986, 434-446.

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

9. Shriberg, E., Wade, E. & Price, P. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction, *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers: San Mateo, Ca., 1992, 49-54.