THE CHAM MODEL OF HYPERARTICULATE ADAPTATION DURING HUMAN-COMPUTER ERROR RESOLUTION*

Sharon Oviatt†

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

ABSTRACT

When using interactive systems, people adapt their speech during attempts to resolve system recognition errors. This paper summarizes the two-stage **C**omputer-elicited **H**yperarticulate **A**daptation **M**odel (**CHAM**), which accounts for systematic changes in human speech during interactive error handling. According to CHAM, *Stage I* adaptation is manifest as a singular change involving the increased duration of speech and pauses. This change is associated with a moderate degree of hyperarticulation, which occurs during a low rate of system errors. In contrast, *Stage II* adaptations are associated with more extreme hyperarticulation during a high system error rate. It entails change in multiple features of speech— including duration, articulation, intonation pattern, fundamental frequency and amplitude. This paper summarizes the empirical findings and linguistic theory upon which CHAM is based, as well as the model's main predictions. Finally, the implications of CHAM are discussed for designing future interactive systems with improved error handling.

1. INTRODUCTION

When speaking to interactive systems, recent research has demonstrated that people typically adapt their language during attempts to resolve system recognition errors (Oviatt, MacEachern & Levow, 1998; Oviatt, Levow & MacEachern, in press; Oviatt, Bernard & Levow, in press). This change in speaking style toward *hyperarticulate* speech involves a stylized and clarified form of pronunciation that speakers routinely adopt when they try to communicate with "at risk" human listeners, in adverse environments (e.g., noise), or during miscommunications. 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 (Levow, 1998; Shriberg, 1992).

The goal of the present paper is to summarize CHAM— the Computer-elicited Hyperarticulate Adaptation Model. CHAM was developed as a model to account for hyperarticulate adaptations observed in users' speech during system error handling. This paper summarizes the empirical findings and linguistic theory upon which CHAM is based, as well as the model's main predictions.

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

† 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/CHCC/)

2. THE ORIGINS OF CHAM

The CHAM model derives from recent empirical findings, and has been motivated by theoretical linguistic concepts on the topic of hyperarticulation.

2.1. Linguistic Theory

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; Lindblom et al., 1992). According to Lindblom's H & H theory, this adaptation varies actively along a continuum from *hypo- to hyper-clear speech.* Hypo-clear speech is relatively relaxed, and contains phonological reductions. A hypo-clear 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, hyper-clear 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 speaking style ranges from hypo- to hyper-clear in a way that contributes substantial variability to the speech signal.

2.2. Empirical Findings

Recent research on human-computer interaction during system error resolution has analyzed the type and magnitude of speech adaptations, with a special focus on the acoustic, prosodic and phonological features of hyperarticulate speech. In these studies, a semi-automatic simulation method was used for collecting data on spoken input during system error handling (Oviatt, MacEachern and Levow, 1998; Oviatt, Levow, MacEachern and Moreton, in press; Oviatt, Bernard & Levow, in press). This technique used a random error generation capability that was adapted to simulate different recognition error rates (e.g., low, high), different types of recognition error (e.g., rejections, substitutions), different characteristics of recognition error (e.g., single error, error spirals), and so forth.

During the test procedure, for example, users input speech such as: "San Francisco airport," but received feedback from the system confirming "San Diego airport." Following this simulated substitution error, users then typically responded by repeating their initial spoken input.

TABLE I - SUMMARY OF ABSOLUTE CHANGE IN LINGUISTIC FEATURES OF STAGE I & II HYPERARTICULATION, [a] BASED ON PAST & PRESENT RESEARCH [b]

Linguistic Feature	*Stage I Change* [c]	*Stage II Change*
Duration:		
Pause interjection	+.57 pauses	**+.32** — +.38 pauses [d]
Pause elongation	+ 97 msec	**+78** — +102 msec
Speech elongation	+190 msec	**+127**— +171 msec
Articulation:		
Hyper-clear phonology	N.S.	**+6** — +9% [e]
Disfluencies	N.S.	**–.25** — –.25 [f]
Pitch:		
Intonation - final fall	N.S.	**+9** — +9% [g]
Pitch minimum	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, MacEachern & Levow (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.

This simulation method and its novel error generation capability was used to collect and compare samples of users' speech immediately before and after simulated system recognition errors. These data on matched original-repeat utterance pairs then were analyzed for the type and magnitude of linguistic adaptations following different types of recognition error.

Analyses based on these studies have indicated that hyperarticulate adaptations during human-computer miscommunication primarily include: (1) change in pause structure toward more pauses and longer pausing, (2) elongation of speech segments, (3) suppression of disfluencies, (4) increase in hyper-clear phonological features, (5) increase in final falling intonation contours. To a lesser degree, or during focal corrections in which one syllable or word is singled out for repair, the following adaptations also typically are found: (6) expansion of pitch range, and (7) small increases in amplitude and fundamental frequency (Oviatt, MacEachern and Levow, 1998; Oviatt, Levow, MacEachern and Moreton, in press; Oviatt, Bernard & Levow, in press).

With respect to the relative magnitude of change in acoustic-prosodic and phonological features during hyperarticulation, durational increases were the most prominent. In particular, adaptation in pause structure dominated the changes observed, with speech segment increases also noteworthy in magnitude. As illustrated in Table I, these durational adaptations represented the only significant change when the rate of system recognition errors was low. Figure 1 illustrates that comparable durational changes were observed following different types of system recognition error.

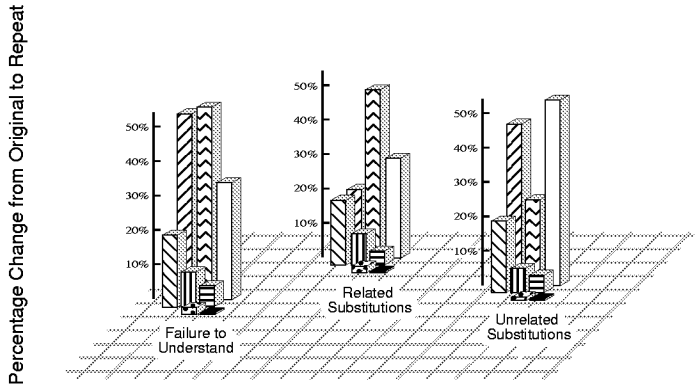

Figure 1. Similarity of Hyperarticulation Profile for Different Error Types [Pause duration ▨; Number of pauses ▨; Disfluencies □; Intonation contour ▨; Speech duration ▥; Hyper-clear phonology ▤; Pitch ▩; Amplitude ■]

Articulatory changes also were a moderately prominent characteristic of hyperarticulate adaptation, including both a drop in spoken disfluencies and an increase in hyper-clear phonological features. Examples of change toward a hyper-clear articulatory pattern include the insertion of previously deleted segments, fortition of alveolar flaps to coronal plosives, and shifts to unreduced **nt** sequences. Essentially, users' speech not only slowed down and separated words more distinctly, it also became more deliberate and better specified in its signal cues to phonetic identity. As illustrated in Table I, these changes were not evident during a low rate of errors, but emerged clearly when the error rate was high. Figure 1 shows that these adaptations were replicated across all different types of system recognition error.

With respect to prosody, speakers shifted to a final falling intonation contour during error correction, which marked the close of their repair subdialogue with the computer. This change also was associated with small decreases in minimum fundamental frequency. While amplitude increases were reliably present during corrections, they were negligible in size. The relative magnitude of change in both pitch and amplitude was small, as seen in Figure 1, although these changes were replicated across all types of system error. Table I reveals that these adaptations were evident only during a high rate of system errors.

Figure 1 summarizes the striking similarity in users' hyperarticulation profile across different types of system recognition error. The most dramatic relative change in hyperarticulate speech occurred in its pause structure, durational characteristics, and pattern of articulation and intonation. However, smaller relative change also can be seen in Figure 1 in pitch and amplitude. Finally, Table I highlights the fact that the degree of hyperarticulation in users' speech is graduated— with only durational changes observed during a low error rate (i.e., 6.5% error rate per utterance), but all of the described features changing during a high rate (i.e., 20% error rate per utterance).

Additional data on hyperarticulate change during users' persistent attempts to correct the same error are detailed elsewhere (Oviatt, Bernard & Levow, in press), as is data on patterns of hyperarticulation during focal versus global utterance repairs (Oviatt, Levow, MacEachern and Moreton, in press).

3. CHAM— MODEL & PREDICTIONS

The two-stage branching **C**omputer-elicited **H**yperarticulate **A**daptation **M**odel (**CHAM**), which is illustrated in Figure 2, has been proposed as a unifying framework to account for these systematic changes in users' speech during interactive error handling (Oviatt, MacEachern & Levow, 1998). According to this empirically-derived 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, fundamental frequency, and amplitude characteristics. This stage is associated with a more extreme degree of hyperarticulation during a high rate of system errors. The two-stage model basically summarizes a progressive unfolding of hyperarticulate speech adaptations, as illustrated in Figure 2.

CHAM predicts that:

- users' speech will adapt toward the linguistically-specified hyperarticulation profile outlined above, with the type and magnitude of change in articulatory, durational, prosodic, fundamental frequency, and amplitude features specified in Table I
- systems characterized by different error rates will elicit different types of hyperarticulate linguistic features, with low errors associated with durational change and high errors with the full range of feature adaptations specified in Table I

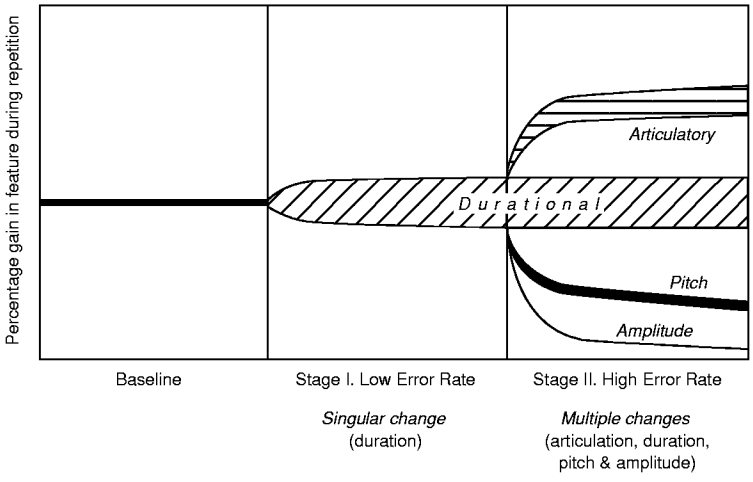

Figure 2. Computer-elicited Hyperarticulate Adaptation Model (CHAM)

- users' speech will adapt similarly to different types of system recognition error
- hyperarticulate adaptations will occur during global repairs to an entire utterance, and also during focal corrections involving an individual syllable or word within a larger utterance
- hyperarticulate adaptations will be evident immediately during a first repair attempt, and will persist during repeated efforts to repair the same error
- hyperarticulate speech adaptations during system error handling will be abrupt moment-by-moment transitions rather than gradual ones

4. IMPLICATIONS

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 model in Figure 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 more continuous 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 technology.

This research also has implications for the collection of more realistic speech data with interactive systems— ones that do in fact err, and that vary in their error base-rates. It is clear that alternative approaches to present error handling methods will need to be explored if improved robustness is to be achieved for spoken language systems. For example, the design of a recognizer specialized for error handling is one option. This approach would require data collection and recognizer training on a corpus of hyperarticulate speech. Depending on the spoken language application's interface design, the special purpose recognizer could either process speech in parallel with the main recognizer or be swapped in during correction episodes when the user is hyperarticulating. Another promising long-term solution would be to avoid eliciting hyperarticulate speech at all by designing a multimodal rather than unimodal interface. This would permit the user to switch to an alternate input mode when he or she expects or actually encounters a system error. This option and its advantages have been discussed in detail elsewhere (Oviatt and vanGent, 1996).

5. CONCLUSIONS

The Computer-elicited Hyperarticulate Adaptation Model (CHAM) has been summarized, including its theoretical origins and empirical documentation on its primary features. Predictions based on CHAM also have been outlined. This model, and the data upon which it is based, provide detailed information about hyperarticulate speech changes during system error resolution. The hyperarticulate adaptations described represent a substantial dynamic source of speech signal variability, which poses a serious challenge to current approaches to speech recognition technology. The implications of CHAM have been discussed for designing future systems with substantially improved error handling.

6. REFERENCES

1. Levow, G. Characterizing and recognizing spoken corrections in human-computer dialogue, Proceedings of the ACL'98 Conference, Association for Computational Linguistics, Montreal, Quebec: Morgan Kaufmann, 1998, 736-742.

2. Lindblom, B. Explaining phonetic variation: A sketch of the H and H theory. In Hardcastle, W. & Marchal, A. (Eds.), Speech Production and Speech Modeling, Kluwer Academic Publishers: Dordrecht, 1990, 403-439.

3. Lindblom, B., Brownlee, S., Davis, B., Moon, S.-J. Speech transforms, *Speech Communication*, 1992, 11 (5), 357-368.

4. Oviatt, S. L., Bernard, J. & Levow, G. Linguistic adaptation during error resolution with spoken and multimodal systems, *Language and Speech*, in press (special issue on ``Prosody and Conversation").

5. Oviatt, S. L., Levow, G., Moreton, E. & MacEachern, M. Modeling global and focal hyperarticulation during human-computer error resolution, *Journal of the Acoustical Society of America,* in press.

6. Oviatt, S. L., MacEachern, M. & Levow, G. Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication*, 1998, vol. 24, 2, 1-23.

7. Oviatt, S. L. & VanGent, R. Error resolution during multimodal human-computer interaction (T. Bunnell & W. Idsardi, eds.), 1996, University of Delaware and A.I. duPont Instit., vol. 1, 204-207.

8. 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, 1992, 49-54.