Prosodic adaptation in human–computer interaction

Linda Bell[†], **Joakim Gustafson**[†] **and Mattias Heldner**[‡]
† Telia Research, Sweden
‡ Centre for Speech Technology, KTH, Sweden
E-mail: linda.e.bell@telia.se, joakim.k.gustafson@telia.se, mattias@speech.kth.se

ABSTRACT

State-of-the-art speech recognizers are trained on predominantly normal speech and have difficulties handling either exceedingly slow and hyperarticulated or fast and sloppy speech. Explicitly instructing users on how to speak, however, can make the human–computer interaction stilted and unnatural. If it is possible to affect users' speaking rate while maintaining the naturalness of the dialogue, this could prove useful in the development of future human–computer interfaces. Users could thus be subtly influenced to adapt their speech to better match the current capabilities of the system, so that errors can be reduced and the overall quality of the human–computer interaction is improved. At the same time, speakers are allowed to express themselves freely and naturally. In this article, we investigate whether people adapt their speech as they interact with an animated character in a simulated spoken dialogue system. A user experiment involving 16 subjects was performed to examine whether people who speak with a simulated dialogue system adapt their speaking rate to that of the system. The experiment confirmed that the users adapted to the speaking rate of the system, and no subjects afterwards seemed to be aware they had been affected in this way. Another finding was that speakers varied their speaking rate substantially in the course of the dialogue. In particular, problematic sequences where subjects had to repeat or rephrase the same utterance several times elicited slower speech.

1. INTRODUCTION

In human-human conversation, dialogue participants often adapt their manner of speaking to that of the other participants. This adaptation takes place at several linguistic levels, allowing phonetic, prosodic, semantic, and pragmatic aspects to come into play. For example, if one person speaks with a low voice, the other participants might start whispering as well. According to Lindblom's H&H theory (1990), people continuously adapt their speech along a continuum from hypo (reduced) to hyper (exaggerated). The speaker's goal is to make sure that he is making himself understood, while at the same time avoiding being overinformative. By adapting his speech from hypo to hyper, the speaker shows that he is aware of the listener's (sometimes limited) access to information [1].

People interacting with computers, e.g. with spoken dialogue systems, are also observed to adapt their manner of speaking in ways that are appropriate in human–human interaction. However, this behavior is not always suitable for human–computer interaction. A typical example is what users of dialogue systems do when they have been misunderstood by the system. Although users know that they are not interacting with another human being, they often hyperarticulate, lower the speech rate, increase the loudness, insert pauses between words, et cetera [2]. That is, they use the same means to increase intelligibility as they would in dialogue with other humans. For the speech-understanding module of a spoken dialogue system, unfortunately, this strategy appears to have an opposite effect. Hyperarticulated speech has been shown to elevate speech recognition failures in human–computer interaction [3, 4]. Efforts to model users' hyperarticulate speech during error resolution may result in the development of future systems with an improved ability to handle such input [5]. However, speech recognizers of today are almost entirely trained on natural, unaffected speech and are ill equipped to interpret speech which is hyperarticulated, emotionally colored or excessively fast or slow. Making users aware of system limitations and telling them how to speak can make the human–computer interaction seem less natural. Moreover, attempts to instruct users to 'speak naturally' to make their language correspond to that of the speech recognizer's training model have not been successful [6].

Why do people adapt their speaking style when they are interacting with a spoken dialogue system? On the one hand, users are aware that they are not speaking to another human being. This implies they should be attempting different strategies than those employed among humans. On the other hand, high expectations regarding the system's capabilities may lead speakers to believe that strategies for human-human interaction are indeed also helpful in this context. Patterns of behavior acquired through frequent repetition (albeit in another context) are difficult to lay aside. Recent studies with simulated systems have shown that children adapt both their response latencies [7] as well as their amplitude [8] to that of their conversational partner, in this case different TTS voices.

This article investigates the occurrence of prosodic adaptation in human–computer interaction. More specifically, we examine whether people who interact with an animated character in a simulated spoken dialogue system adapt their speaking rate to that of the system. Furthermore, we study the effects of errors on prosodic adaptation patterns in a spoken dialogue system.

ISBN 1-876346-48-5 © 2003 UAB

2. METHOD

Sixteen volunteer subjects, 9 men and 7 women between the ages of 17 and 59, participated in a user experiment. The study was performed in the exhibition area of the Telecom museum in Stockholm. Most of the subjects were members of the general public, and some were museum employees. They were informed about the general purpose of the study, and were told that they were being recorded. Subsequently, subjects were given a pictorial scenario and instructions on how to talk into the microphone, and were told to await further instructions. Before starting their actual interaction with the system, subjects were asked to read five sequences of digits. After each of these sequences, the system echoed the digits. The subjects then listened to a prerecorded instruction on how to proceed with the actual task. This involved helping a story tale character, Cloddy Hans, to solve a puzzle. Subjects were told that the solution to the puzzle would be revealed to them after a number of colored geometrical figures had been moved from one part of the screen to another in a certain order. Since Cloddy Hans lacked information required to perform this task, the subjects had to help him using their pictorial scenarios.

In reality, the system was a Wizard-of-Oz simulation. One of the co-authors, sitting behind a screen, acted as the system's speech recognition and dialogue management following a predetermined template. On three occasions in the dialogue, the system deliberately 'misunderstood' the user. During one of these error sequences, subjects had to repeat themselves three times in a row. The graphical user interface and one of the subjects interacting with the system can be seen in Figure 1.

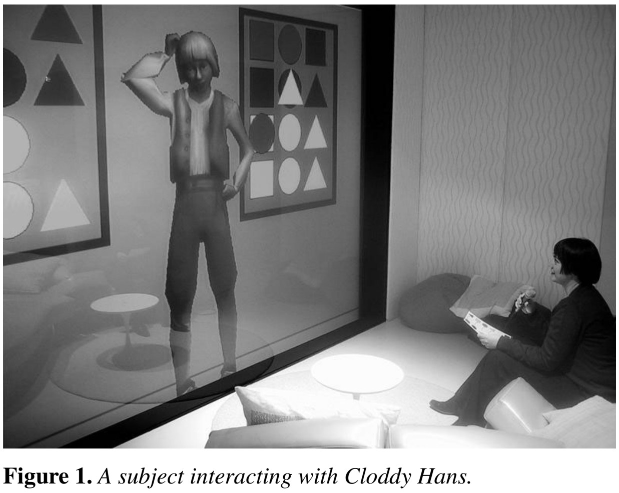

Figure 1. *A subject interacting with Cloddy Hans.*

After each user utterance, Cloddy Hans replied with an explicit feedback prompt. Subjects were randomly assigned to interact with either one of two versions of the simulated system: One version in which the speaking rate of the feedback prompts was increased, and one in which it was slowed down. All other aspects of 'fast' and 'slow' Cloddy Hans were kept identical. Cloddy Hans' spoken output was generated in the following way: One of the co-authors recorded the prompts, enacting Cloddy Hans' personality and speaking style. The speech was then manipulated to simulate vocal tract lengthening and lowering of f_0, using a TD-Psola algorithm [9]. The same algorithm was also used to produce the two test sets by increasing or decreasing the speech rate of the original utterances by 30%.

3. DATA CODING AND ANALYSIS

3.1. CORPUS STATISTICS AND TAGGING

The total number of recorded user utterances was 297. Each dialogue consisted of 6 tasks (i.e. there were 6 colored geometrical figures to be moved) and 16 to 27 user turns. The total number of words in the corpus was 2173. The entire corpus was orthographically transcribed, and all user turns were tagged with information on position in the dialogue, type of user turn, previous system output, etc. At the word level, the user utterances were also labeled for lexical content, distinguishing between 'color', 'shape' and 'other' words. However, for the purpose of this particular study, only a subset of the user turns was of interest, namely the first turns in each task, and repetitions and rephrases of these. Thus, 130 user turns consisting of short comments such as "ok" or "yes" and a few erroneous utterances were excluded from the subsequent analyses. The average length of the remaining 167 utterances was about 6 seconds.

3.2. DIALOGUE EXAMPLE

Table 1 shows an example dialogue from the corpus with labeling of position and type of user turns. For example, the utterance labeled '2.1_original' is the first turn of the second task in the dialogue, '4.2_repeat' is a (near) verbatim repetition of the first turn of the fourth task, and '4.3_rephrase' is a turn where the user repeats his previous turn but in different words. The turns labeled 'ack' are acknowledgements of the system's previous explicit question, which were excluded from the speech rate analyses.

3.3. ESTIMATE OF SPEAKING RATE

The acoustic analysis involved a quantitative estimate of user speaking rate calculated by taking the average normalized segment duration across each word. A standard z-score technique [e.g. 11] was used to normalize for inherent duration, and the means and standard deviations used in the normalization were taken from a database of all segments from all speakers in the experiment. However, a prerequisite for this estimate was a segmentation of the speech material into words and phonemes. This was achieved by means of an automatic alignment algorithm [10]. The input to this aligner is a speech file and a verbatim transcription of the speech. The output consists of two tiers marking words in standard orthography, and phonemes, respectively. The phoneme tier is supplemented with lexical prosodic features such as primary and secondary stress and word accent type (i.e. accent I or II), thus enabling separation of stressed and unstressed vowels etc. in the z-score normalization.

User	User turn	System
Eeh red tr eh red square take red square eeh left	1.1_original	Now I didn't understand
Take red square and place on top of blue triangle	rephrase	Do you want me to move the red square to the blue trigon?
Yes	ack	Ok
Move the blue circle to the green square	2.1_original	Now I didn't understand
Blue circle to green square	2.2_rephrase	Do you want me to move the blue circle to the green square?
Yes	ack	Ok
Green triangle to red circle	3.1_original	Do you want me to move the green trigon to the red circle?
Yes	ack	Ok
The yellow circle to the green trigon	4.1_original	Do you want me to move the blue square to the red circle?
No the yellow circle to the green trigon	4.2_ repeat	Do you want me to move the blue circle to the green trigon?
No lightgreen circle to green trigon	4.3_ rephrase	Now I didn't understand

Table 1. *Translated excerpt of a dialogue between a user and Cloddy Hans.*

4. RESULTS

Figure 2 illustrates some of the speaking rate variations in the user utterances. As can be seen, the two groups of subjects started out with similar speech rates (cf. 1.1_original in Fig. 2). At this point in the dialogue, they had not yet interacted with Cloddy Hans's could not have been affected by his speech. As illustrated in the dialogue example above, the subjects were uncertain on how to express themselves here and many of the utterances coded as 1.1_original were low in speech rate, fragmented and disfluent. Subsequently, subjects in the 'fast' and 'slow' groups began to diverge in terms of speaking rate. The users' second utterance (2.1_original) was deliberately rejected by the system. Subjects reacted to this by either repeating their original utterance verbatim or rephrasing it. An interesting observation was that the repetitions in this dialogue context (2.2_repeat) were increased in speaking rate while the rephrases (2.2_rephrase) were decreased. In contrast, later on the dialogue the system misunderstood the users' utterance completely. In this case, both repetition (4.2_repeat) and rephrase (4.2_rephrase) were pronounced slower.

An ANOVA was used to examine the effects of system speaking rate, user turn, and lexical content. The dependent variable was the estimate of user speaking rate described above, and the independent variables were:

- system speaking rate (fast vs. slow)
- user turn (see Figure 2)
- lexical content (color vs. shape vs. other)

The ANOVA showed a significant effect of system speaking rate [*F*(1,1341)=4.1; *p*<0.05] in the expected direction (fast<slow). In addition, there were significant effects of user turn [*F*(13,1341)=2.2; *p*<0.05], and of the interaction between speaking rate and user turn [*F*(13,1341)=1.9; *p*<0.05]. That is, turns differed significantly in speaking rate, and the effect of system speaking rate differed between turns. Finally, there was a significant effect of lexical content [*F*(2,1341)=33.7; *p*<0.05]. A Bonferroni pairwise comparisons test on the effect of lexical content showed that 'color' words were pronounced significantly slower than 'shape' and 'other' words, while there was no significant difference between 'shape' and 'other' words. None of the other effects were significant.

Another ANOVA was used to examine the effects on silent pauses within user utterances. The dependent variable here was the absolute duration of the silent pauses, and the independent variables were:

- system speaking rate
- user turn

There was a significant effect of user turn [*F*(13,838)=3.6; *p*<0.05], but neither speaking rate, nor the interaction of speaking rate and user turn was significant. However, the power to detect statistical significances was fairly low (0.4 for speaking rate, and 0.6 for the interaction).

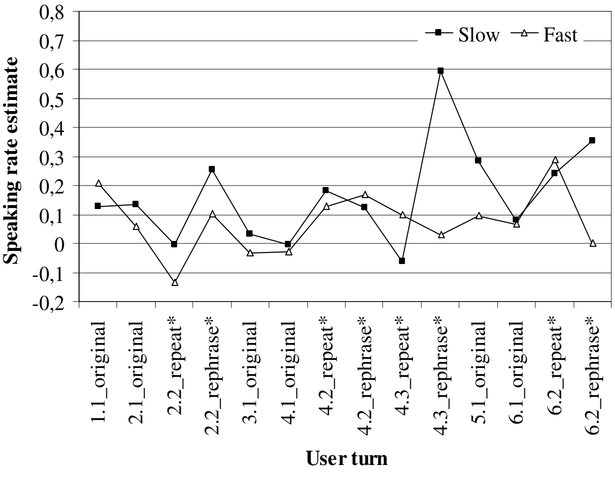

Figure 2. *Effects of system speaking rate (fast vs. slow) on user speaking rate (measured as the average z-score normalized segment duration across words in std. devs.) in the different user turns. Users either rephrased or repeated turns but not both, these turns are marked with *.*

ISBN 1-876346-48-5 © 2003 UAB

5. DISCUSSION

Our results support the hypothesis that users interacting with the 'slow' version of the system speak slower than those who interact with the 'fast' one. The results also reveal substantial variations in speaking rate that can be attributed to the dialogue context. A tendency in our data is that when the dialogue is successful, an increase in speech rate is elicited. Local effects on speech rate, such as a slow user utterance after a system misunderstanding, appear to be transient and quickly passing. Once the system seems to understand their input again, users return to their normal manner of speaking. Among the local convergence effects, we found that repeats and rephrases tended to be slower after a system turn that echoed the referents of the user's previous input in a completely erroneous way. In these cases, users probably spoke slower as a result of an increase in cognitive load caused by the simulated error. Moreover, our analyses showed that the lowering of speech rate mainly affected the 'color' words in the dialogues. Users often modified the 'shape' words lexically by exchanging a word for a near-synonym. In the course of the dialogue, there was a lexical convergence effect where users often conformed to Cloddy Hans's vocabulary. It was more difficult for the users to come up with alternatives to the 'color' words, and only two such instances were found in the corpus. Furthermore, color words constituted contrastive material in the sense that all misunderstood turns concerned color and shape or color alone, and never shape alone. Finally, the analyses revealed that the lowered speech rate did not affect of the within-utterance silent pauses significantly.

One possible interpretation of the results is that users modify their speech according to their current model of the system's input understanding capabilities. Two factors seem to influence the users' model: The system's output speech rate and the system's ability to handle the previous utterance(s). While successful turns elicit an increase in speech rate for both groups of users, subjects who interact with the 'slow' version of the system are affected to a lesser extent. On the other hand, while both groups of users react to simulated system errors by speaking slower, users interacting with the 'fast' Cloddy Hans tend to do this to a lower degree.

Post-experimental discussions confirmed that subjects were aware of the fact that they were adapting their language at the lexical level. Several subjects spontaneously mentioned the fact that they deliberately modified their vocabulary to match that of the system. However, adaptation of speaking rate was not mentioned as a strategy consciously used by the subjects. The tendency in our data that subjects decreased their speaking rate can partly be attributed to the general impression of the animated agent. The fact that Cloddy Hans's speaking style and general appearance implied a certain limitation in his intellectual abilities was probably relevant. Several of the subjects gave Cloddy Hans less flattering nicknames (such as "pucko"), and one of subjects afterwards reported, "I thought he seemed a bit dunce".

ACKNOWLEDGEMENTS

The work described in this paper was supported by the EU / HLT funded project NICE (IST-2001-35293).

REFERENCES

[1] B. Lindblom (1990) Explaining phonetic variation: A sketch of the H&H theory. In *Speech Production and Speech Modelling* (A. Marchal, editor). Dordrecht.: Kluwer Academic Publishers.

[2] S. Oviatt, J. Bernard & G.-A. Levow (1998) Linguistic adaptations during spoken and multimodal error resolution, *Language and Speech,* 41(3-4), 419-442.

[3] N. Yankelovitch, G. Levow & M. Marx (1995) Designing Speech Acts: Issues in Speech User Interfaces. In *Proceedings of the Conference on Human Factors in Computer Systems*, pp. 369-376. Denver, CO: ACM Press.

[4] J. R. Rhyne & C. G. Wolf (1993) Recognition-based user interfaces. In *Advances in Human–computer Interaction* (D. Hix, editor), pp. 191-250. Norwood, N.J.: Ablex Publishing.

[5] S. Oviatt, M. MacEachern & G.-A. Levow (1998) Predicting Hyperarticulate Speech during Human–computer Error Resolution, *Speech Communication,* 24(2), 1-23.

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

[7] C. Darves & S. Oviatt (2002) Adaptation of users' spoken dialogue patterns in a conversational interface. In *Proceedings of ICSLP'02*. Denver, CO.

[8] R. Coulston, S. Oviatt & C. Darves (2002) Amplitude convergence in children's conversational speech with animated personas. In *Proc. of ICSLP'02*. Denver, CO.

[9] J. Gustafson & K. Sjölander (2002) Voice Transformations for Improving Children's Speech Recognition in a Publically Available Dialogue System. In *Proceedings of ICSLP'02*, pp. 297-300.

[10] K. Sjölander (2001) Automatic alignment of phonetic segments. In *Working Papers 49: Papers from Fonetik 2001*, pp. 140-143. Lund: Lund University, Dept. of Linguistics.

[11] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf & P. J. Price (1992) Segmental durations in the vicinity of prosodic phrase boundaries, *Journal of the Acoustical Society of America,* 91(3), 1707-1717.