Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface

Courtney Darves and Sharon Oviatt

Department of Computer Science and Engineering
Oregon Health & Science University
+1-503-748-1342; {court, oviatt}@cse.ogi.edu
http://www.cse.ogi.edu/CHCC

ABSTRACT

The design of robust new interfaces that process conversational speech is a challenging research direction largely because users' spoken language is so variable, which is especially true of children. The present research explored whether children's response latencies before initiating a conversational turn converge with those heard in the text-to-speech (TTS) of a computer partner. A study was conducted in which twenty-four 7-to-10-year-old children conversed with animated characters that responded with different types of TTS voices during an educational software application. Analyses confirmed that, while interacting with opposite TTS voices, children's average response latencies adapted 18.4% in the direction of their computer partner's speech. These adaptations were dynamic, bi-directional, and generalized across different types of users and TTS voices. The long-term goal of this research is the predictive modeling of human-computer communication patterns to guide the design of well synchronized, robust, and adaptive conversational interfaces.

1. INTRODUCTION

During the past decade, rapid advances in spoken language technology, natural language processing, dialogue modeling, multimodal interface design, and mobile applications all have stimulated interest in a new class of *conversational interfaces*. These conversational interfaces are unique in their ability to support parallel natural language processing for both input (e.g., speech recognition) and output (e.g., TTS) between human and computer. Unfortunately, past research has tended to focus either on input technologies or output technologies, but has largely ignored the overall interactive cycle between human and computer. One exception to this limitation is the recent literature on conversational interfaces with animated characters [3], which has begun to explore the impact of their output on the intelligibility and efficiency of users' spoken dialogue to a system [2, 7]. In the present research, we explore whether a conversational system's TTS output has a significant influence on users' speech to a computer.

The cognitive science literature on Communication Accommodation Theory (CAT) has identified adaptation that occurs during conversation in the linguistic and nonverbal behavior between two interlocutors. Basically, this theory asserts that interpersonal conversation is a *dynamic adaptive exchange* in which a speaker's lexical, syntactic, and speech signal features are tailored to their conversational partner in a way that fosters predictability, intelligibility, and efficiency [1, 5]. Both children and adults will adapt features of their speech to more closely match their partner, including amplitude, pitch, duration, dialogue response latency, and phonological features [1, 5, 10, 11, 12].

Despite the existence of this interpersonal literature, HCI research has yet to investigate whether users of a conversational interface likewise adapt their speech systematically to converge with a computer partner's TTS output. Lexical and syntactic accommodation previously have been reported during database query applications [15], but the possibility that users may adapt signal-level features of their speech has not yet been explored. In particular, user accommodation of the temporal aspects of their speech, including response latencies before initiating a conversational turn, could have important implications for synchronizing turn-taking behavior, and for the processability of speech by a recognizer. The present research explores whether children's response latencies before initiating a conversational turn adapt in the direction of those heard in the text-to-speech (TTS) of a computer partner. Users' adaptation of other acoustic-prosodic characteristics (i.e., amplitude) are reported in a companion paper [4].

Recent research has estimated that children's speech is subject to recognition error rates two-to-five times higher than adult speech [8, 13]. Children's speech is harder to process than that of adults because many basic features are less mature, more variable, and change developmentally from year to year [14]. During interaction with computers, for example, children's speech is considerably more disfluent and includes invented words [7]. Children also can be shy and difficult to engage in a conversational setting, such that they are reluctant to speak at all and low in volume when they do. As a result, the development of conversational interfaces for children will require specialized design strategies that engage child users, and that guide their speech to be processable and well coordinated with a spoken dialogue system. Animated characters have the potential to be an effective interface vehicle for eliciting speech from children and guiding their spoken exchange with a computer system.

If children's speech does converge with the TTS output heard from a computer partner, then further studies need to determine which speech features adapt, what magnitude of change occurs, and what the implications are for designing a

new generation of robust conversational interfaces. To the extent that convergence occurs, it may be possible to design interfaces that proactively guide children's spoken input to remain within optimal system processing bounds. It also may be possible to begin designing adaptive conversational systems that accommodate children's dialogue patterns. That is, given better cognitive science modeling of children's actual spoken interaction with computers, the long-term goal is to design more robust, well synchronized, and mutually adaptive conversational systems.

2. GOALS OF THE STUDY

The present research examined whether the duration of children's interspeaker response latencies might be influenced by a computer partner's TTS output. Response latencies were investigated in part because they are known to converge during interpersonal conversation, even in young children, and also because of their pivotal role in synchronizing a successful dialogue exchange [10, 11]. We hypothesized that users' response latencies would adapt in the direction of those heard from their computer partner. More specifically, it was predicted that the duration of children's conversational response latencies would be longer when conversing with an animated character that embodied an introvert voice, and shorter when speaking with an extrovert TTS voice. It is well known that extrovert and introvert voices are associated with distinct speech signal characteristics. For example, in contrast to the introvert voice profile, extrovert speech typically is louder and faster in rate, exhibits higher pitch and wider pitch range, as well as shorter response latencies [6, 9]. A related goal was to investigate whether users' response latencies during human-computer interaction would adapt bi-directionally, that is, whether they would be equally likely to adapt by shortening or lengthening in response to different TTS voices. In addition, it was predicted that users' conversational response latencies would readapt if a contrasting computer voice was introduced later during their session. Finally, the generality of any results regarding speech adaptation was examined across different types of user groups and TTS voices.

3. METHODS

3.1. Participants, Task, and Procedure

Twenty-four elementary-school children participated in this study as paid volunteers. The group of participants ranged in age from 7 yrs., 6 mos. to 10 yrs., 2 mos., and was gender balanced. Participation was conducted at a local elementary school field site.

Children participating in the study were introduced to *Immersive Science Education for Elementary kids* (*I SEE!*), which is an application designed to teach children about marine biology, simple data tabulation, and graphing. The interface permitted children to use speech, pen, or multimodal input while conversing with animated software characters as they learned about marine biology. The marine animals were available as "conversational partners" who answered questions about themselves using text-to-speech (TTS) output. Figure 1 illustrates the *I SEE!* interface.

Before starting a session, each child received instructions and practice with a science teacher on how to use the *I SEE!* interface on a small hand-held computer, shown in Figure 2. Then the teacher left, and the child spent approximately one hour alone in a quiet classroom playing with the educational software. During this time, he or she conversed with 24 marine animals (e.g., octopus, shown in Figure 1), which were organized into three task groups of eight animals apiece.

During data collection, children's input was received by an informed assistant who interpreted their queries and provided system responses as part of a simulation method, although children believed they were interacting with a fully-functional system. The simulation environment ran on a PC, and received input from a Fujitsu Stylistic™ 2300 that was used by the children. Details of the simulation infrastructure and its performance have been summarized elsewhere [7].

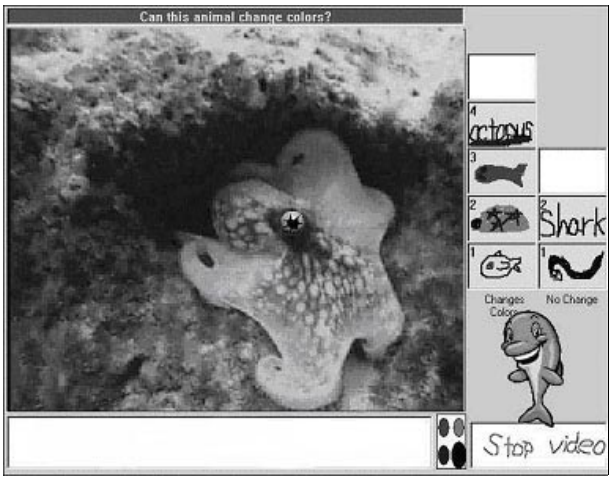

Figure 1: I SEE! application interface

Figure 2: Eight-year old boy at school as he asks an animated marine character questions about itself

3.2. Text-to-Speech Output

Text-to-speech voices from Lernout and Hauspie's TTS 3000 were used to convey the animated characters' spoken output, and were tailored for intelligibility of pronunciation. They were further tailored to represent opposite ends of the introvert-extrovert personality spectrum, as indicated by the speech signal literature [9]. In total, four TTS voices were used in this study: male extrovert (ME), male introvert (MI), female extrovert (FE), and female introvert (FI). Table 1 summarizes

the differences in global speech signal profiles between the introvert and extrovert TTS voices used in the present study, as manifest in the male and female voice samples. Due to preloading of system responses, lexical content was controlled in the different TTS voice conditions. In addition, the TTS voice conditions were counterbalanced across task sets, which controlled for the visual appearance of different animated characters that were presented during the study.

Table 1: Differences between extrovert and introvert TTS conditions for female and male prototype voices.

TTS Voice Type	Mean Amplitude (dB)	Mean Pitch Range (Hz)	Utterance Rate (syl/sec)	Dialogue Response Latency (sec)
FE	60	186	5.2	1.65
ME	58	106	5.2	1.65
FI	45	71	3.3	3.36
MI	44	58	3.3	3.36

3.3. Research Design

The research design for this study was a repeated measures factorial, and the dependent measure was children's response latency before initiating a conversational turn. The within-subject factor was (1) Type of TTS Voice embodied by the marine character (Introvert, Extrovert). This factor remained constant for the first two tasks, but switched for the third task (from I to E, or E to I). To test the generality of any TTS effects, I and E voice types also were tested using different gender instantiations. As a result, (2) TTS Voice Gender (Male, Female) constituted a separate between-subject factor. The different visual embodiments representing the 24 animated characters were combined with these TTS voices in a counterbalanced manner across tasks and subjects. Two other between-subject factors included (3) Child Gender (Male, Female) and (4) Child Age (Young, Old), with a median split dividing children into a younger group (mean age 8 yrs., 2 mos.) and an older one (mean age 9 yrs, 7 mos.).

3.4. Data Coding and Analysis

All child-computer interaction was videotaped and transcribed, with utterances digitized. Using Praat speech signal analysis software, interspeaker response latencies, or the duration between the end of an animated character's spoken utterance and the beginning of the child's next utterance, were hand-measured in seconds during tasks 1-3.

3.4.1. Inter-coder Reliability

In total, 15% of the data were scored by an independent coder. The mean departure between independent coders was 0.06 seconds, and 80% of measurements were accurate to within 0.08 seconds.

4. RESULTS

Data were analyzed from all 24 participants, including 1,813 utterance response latencies.

4.1. Effect of TTS Voice Type

As predicted, children's response latencies differed when they conversed with an animated character that spoke with the extrovert versus introvert voice. As shown in Table 2, their average conversational lags increased from 4.29 seconds when interacting with the extrovert voice to 5.07 seconds with the introvert one, which was a significant difference in duration (log transformed) by repeated measures ANOVA, F=6.52 (df=1), $p < .025$. This change represented a +18.4% relative increase in duration during the introvert condition.

A priori paired t-tests confirmed the prediction that this speech signal change was bi-directional, as shown in Table 2. That is, children's response latencies (log transformed) increased significantly when first exposed to the extrovert voice and then switched to the introvert (mean 4.23 and 5.04 seconds), t=2.04 (df=11), $p < .035$, one-tailed. Likewise, their latencies (log transformed) decreased significantly when exposed to the introvert voice, but later switched to the extrovert (mean 6.00 and 5.24 seconds), t=1.82 (df=11), $p < .05$, one-tailed. Table 2 summarizes the relative increase in response latency for the introvert condition for this split comparison by task order.

Table 2: Average response latencies in seconds for extrovert versus introvert TTS voice conditions, and percentage E-I increase as a function of order.

	Extrovert Voice	Introvert Voice	E-I Increase
Grand Mean	**4.29**	**5.07**	**+18.4%**
E to I Order	4.23	5.04	+19.1%
I to E Order	5.24	6.00	–12.7%

4.2. Generality of Adaptation to TTS Voice

The conversational response latencies were longer for the young group of children (mean age 8 yrs., 2 mos.) compared with the older group (mean age 9 yrs., 7 mos.), averaging 6.25 seconds and 4.23 seconds, respectively. This main effect of child age on durations (log transformed) was a significant one by repeated measures ANOVA, F=4.98 (df=1), $p < .04$. However, no significant difference was found in children's differential adaptation to the extrovert and introvert TTS voices as a function of their age or gender, or as a function of the specific TTS voice they heard (i.e., male vs. female voice). That is, the main effect of TTS voice type (E vs. I) generalized across all of these variables.

5. DISCUSSION

As during interpersonal conversation, the users of a conversational interface will alter their basic spoken dialogue patterns to converge with those of a computer partner. In this study, 7-to-10-year-old children lengthened their conversational response latencies between turns by 18% when interacting with a slow-paced introverted computer character, compared with a rapid extroverted one. These changes in their speech patterns also were bi-directional. That is, when exposed to an extrovert dialogue pattern followed by a slower introvert one, child speakers correspondingly slowed down from an average response lag of 4.23 to 5.04 seconds. Conversely,

when interacting with an introvert dialogue pattern followed by a faster extrovert, the direction of change reversed, speeding up from 6.00 to 5.24 seconds. These results on the adaptation of children's response latencies underscore the dynamic nature of their ability to adapt spoken dialogue patterns to synchronize with different computer partners.

Users' inclination to adapt their speech to accommodate their computer partner was a general phenomenon in several important respects. As explored in this study, no significant differences were observed between younger and older children in this adaptive capability, or between males and females. Likewise, no difference was detected in children's speech adaptation due to the specific gender instantiation of the TTS voice they heard. Finally, since children interacted with twenty-four visually distinct characters counterbalanced across tasks, speech adaptations were not limited to a specific visual appearance. That is, differential adaptation to the introvert and extrovert TTS voices generalized across all of these variables.

In a conversational system, the optimal system response time, or in this case delay before the animated character's reply, is not necessarily the fastest possible time supportable by the software. In the present study, children's response latencies averaged 4.7 seconds, and ranged between 1.6 and 10.1 seconds for all twenty-four participants. In addition, the response latencies for the younger children averaged 6.2 seconds, which was significantly longer than 4.2 seconds in the older group. During interpersonal conversation, similar response latency durations have been observed in our lab. For example, response latencies for 7-to-10-year-old children engaged in a similar query-answer activity with an adult averaged 3.9 seconds, and ranged from 2.5-5.6 seconds. In the future, additional research will be needed to model what the optimal human-computer response delays should be for new conversational interfaces, in part depending on the user group and conversational domain. In addition, future research needs to develop adaptive conversational interfaces that can both: (1) adjust to the substantial individual differences among users, and (2) track and adapt to changes in key features of users' spoken language during dynamic conversation. Such systems then would be capable of supporting the *mutual* speech adaptation that characterizes interpersonal conversation, which could improve the synchrony and quality of the human-computer conversational exchange.

This research and related work on amplitude convergence [4] suggests that future conversational interfaces that include TTS output could be designed to actively manage hard-to-process aspects of children's spoken language (e.g., low volume speech), and also to facilitate a well-synchronized human-computer exchange. To the extent that animated character design can exploit children's natural inclination to converge with their partner's speech patterns, it may provide a very effective tool for transparently guiding their spoken dialogue to be more processable and coordinated with spoken dialogue recognition. The long-term goal of this research is the development of new predictive models of human-computer communication, which can be used to guide the design of a new class of effective conversational interfaces.

6. ACKNOWLEDGEMENTS

This research was supported in part by Grants IRI-9530666 and IIS-0117868 from the National Science Foundation, Special Extension for Creativity (SEC) Grant IIS-9530666 from NSF, and a gift from the Intel Research Council. Thanks to Matt Wesson for implementing simulation, transcription, and data analysis tools, and for assisting during testing. Thanks also to Rachel Coulston for assisting with linguistic transcription, and to the students who participated in this research.

7. REFERENCES

[1] Burgoon, J., Stern, L. & Dillman, L., 1995, *Interpersonal Adaptation: Dyadic Interaction Patterns.* Cambridge Univ. Press, Cambridge, UK.

[2] Cassell, J. & Thorisson, K.R., 1999, "The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents," *Applied Artificial Intelligence Journal,* 13 (4–5): 519-538.

[3] Cassell, J., Sullivan, J., Prevost, S., & Churchill, E., 2000, *Embodied Conversational Agents,* MIT Press, Cambridge MA.

[4] Coulston, R., Oviatt, S. & Darves, C., "Amplitude convergence in children's conversational speech with animated personas," manuscript in submission.

[5] Giles, H., Mulac, A., Bradac, J. & Johnson, P., 1987, "Speech accommodation theory: The first decade and beyond," *Communication Yearbook 10,* ed. by M. L. McLaughlin, Sage Publ., London, UK, 13-48.

[6] Nass, C. & Lee, K.L., 2000, "Does computer-generated speech manifest personality? An experimental test of similarity-attraction," *Proceedings of the Conference on Human Factors in Computing Systems,* ACM Press, NY, 329-336.

[7] Oviatt, S. & Adams, B., 2000, "Designing and evaluating conversational interfaces with animated characters," in *Embodied Conversational Agents,* ed. by J. Cassell, J. Sullivan, S. Prevost,, and E. Churchill, MIT Press, Cambridge, MA, 319-343.

[8] Potamianos, A., Narayanan, S. & Lee, S., 1997, "Automatic speech recognition for children," *European Confer. on Speech Commun. & Technology,* 5*:* 2371-2374.

[9] Scherer, K.R., 1979, "Personality markers in speech," *Social Markers in Speech,* ed. by K. Scherer & H. Giles, Cambridge Univ. Press, Cambridge, UK, 147-209.

[10] Street, R., Street, N. & VanKleeck, A., 1983, "Speech convergence among talkative and reticent three-year-olds." *Language Sciences,* 5: 79-96.

[11] Welkowitz, J., Cariffe, G. & Feldstien, S., 1976, "Conversational congruence as a criterion of socialization in children," *Child Development,* 47: 269-272.

[12] Welkowitz, J., Feldstein, S., Finklestein, M., & Aylesworth, L., 1972, "Changes in vocal intensity as a function of interspeaker influence," *Perceptual and Motor Skills,* 35: 715-18.

[13] Wilpon, J. & Jacobsen, C., 1996, "A study of speech recognition for children and the elderly," *Proceedings of the International Conference on Acoustics, Speech & Signal Processing,* IEEE Press, Atlanta, GA, 349–352.

[14] Yeni-Komshian, G., Kavanaugh, J. & Ferguson, C. (eds.), 1980, *Child Phonology, Volume 1: Production*. Academic Press, NY.

[15] Zoltan-Ford, E., 1991, "How to get people to say and type what computers can understand," *International Journal of Man-Machine Studies,* 34: 527-47.