AMPLITUDE CONVERGENCE IN CHILDREN'S CONVERSATIONAL SPEECH WITH ANIMATED PERSONAS

Rachel Coulston, Sharon Oviatt and Courtney Darves

Department of Computer Science and Engineering
Oregon Health & Science University
+1-503-748-1602; {rachel|oviatt|court}@cse.ogi.edu
http://www.cse.ogi.edu/CHCC

ABSTRACT

During interpersonal conversation, both children and adults adapt the basic acoustic-prosodic features of their speech to converge with those of their conversational partner. In this study, 7-to-10-year-old children interacted with a conversational interface in which animated characters used text-to-speech output (TTS) to answer questions about marine biology. Analysis of children's speech to different animated characters revealed a 29% average change in energy when they spoke to an extroverted loud software partner (E), compared with an introverted soft-spoken one (I). The majority, or 77% of children, adapted their amplitude toward their partner's TTS voice. These adaptations were *bi-directional*, with *increases* in amplitude observed during I to E condition shifts, and *decreases* during E to I shifts. Finally, these results generalized across different user groups and TTS voices. Implications are discussed for guiding children's speech to remain within system processing bounds, and for the future development of robust and adaptive conversational interfaces.

1. INTRODUCTION

Communication Accommodation Theory (CAT) describes interpersonal conversation as a *dynamic adaptive exchange* in which a person's spoken language is tailored to their interlocutor in fundamental ways. Both children and adults will accommodate the spoken language of a conversation partner, including its basic acoustic-prosodic features such as amplitude, pitch and duration [1, 6, 14]. However, research on human computer interaction has yet to investigate whether users likewise adapt their speech during conversational interaction with a computer partner.

In the present study, we explore whether children adapt their speech amplitude while conversing with animated characters that respond with different types of TTS voices during an educational exchange. In a companion paper, we also present results on children's adaptation of speech duration--in particular dialogue response latencies--as they converge with those of a software partner [4]. One goal of this work is the modeling of children's speech, because it is well known to produce substantially higher recognition error rates than adult speech, typically by a factor of two-to-five fold [11, 15]. Children's speech currently is difficult to process because it is more disfluent, more variable in acoustic-prosodic features, and changing developmentally [9, 16]. Children also can be shy and hard to engage in conversation, such that they are reluctant to speak at all and low in volume when they do.

As a result, the development of future conversational interfaces for children will require specialized design strategies that engage child users, and that guide their speech to be audible and processable by a speech recognition system. Recent work has made progress in designing educational software that effectively engages child users [2, 5], and even increases the amount of language they direct to a conversational interface [3]. In this research, we explore whether the TTS voice of an animated character can be used to entrain children's speech to be higher or lower in amplitude, which would be expected to influence the robustness of a conversational system.

2. GOALS OF THE STUDY

Since TTS acoustic parameters can be controlled precisely, their manipulation provides a unique opportunity to study the dynamics of amplitude accommodation. For this research, introvert and extrovert TTS voices were used, in part because their acoustic-prosodic features are well defined and have been used in previous research. In contrast to an introvert voice profile, extrovert speech typically is louder and faster in rate, exhibits higher pitch and wider pitch range, and shorter dialogue response latencies [8, 13].

The specific goals of this study were to:

- Examine whether children's speech amplitude is influenced by an animated software partner's TTS output during conversational interaction
- Determine whether their amplitude readapts dynamically if a contrasting computer voice is introduced part way through an interaction
- Assess whether amplitude adaptation is bi-directional, increasing and decreasing in accord with the TTS amplitude heard
- Evaluate the magnitude of amplitude adaptation
- Establish the generality of any adaptation effects across different 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. Participants ranged in age from 7 years, 6 months to 10 years, 2 months, and were gender balanced. The study was conducted at a local elementary school.

Children participating in the study were introduced to Immersive Science Education for Elementary kids (*I SEE!*). Figure 1 illustrates the *I SEE!* interface. *I SEE!* is an application in which children could use speech, pen, or multimodal input while conversing with animated software characters to learn

about marine biology. The marine animals were available as conversational partners who answered questions about themselves using text-to-speech (TTS) output. A "Spin the dolphin" character (lower right of Figure 1) also was available to answer questions and provide help (e.g., spelling, using the system) and entertainment (e.g., telling jokes).

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. 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., lobster, as shown in Figure 1), which were organized into three task sets 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 characteristics have been summarized elsewhere [9].

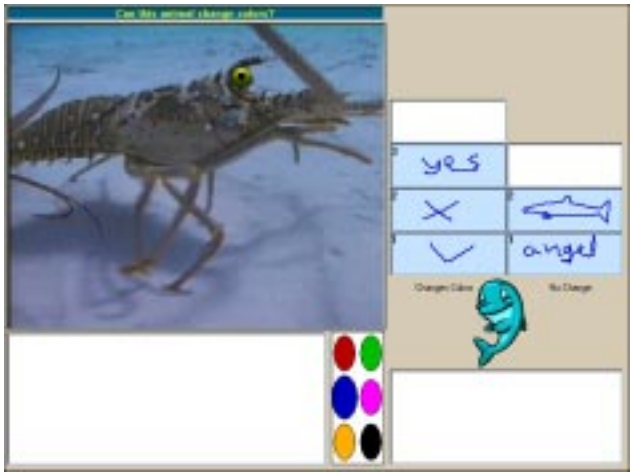

Figure 1: *I SEE!* educational software interface

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 literature [8, 13]. In total, four TTS voices were used in this study: (1) Male Extrovert (ME), (2) Male Introvert (MI), (3) Female Extrovert (FE), and (4) Female Introvert (FI). Table 1 summarizes the differences in global speech signal profiles between these TTS voices. The TTS voice conditions were counterbalanced across task sets, which controlled for the visual appearance of different animated characters presented during the study.

TTS Voice Type	Mean Amplitude (dB)	Mean Pitch Range (Hz)	Utterance Rate (syl/sec)	Dialogue Response Latency (sec)
FE	60	186	5.3	1.65
ME	58	106	5.2	1.65
FI	45	71	3.3	3.36
MI	44	58	3.3	3.36

Table 1: Acoustic differences between extrovert and introvert TTS conditions for female and male animated character voices

3.2. Research Design

The research design for this study was a repeated measures factorial, and the dependent measure was children's amplitude. The main within-subject factor was (1) the 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). During tasks 1 and 2, speech to the two co-present animated characters (marine character and Spin the dolphin) also contrasted, at which point the TTS voices differed along both the introvert-extrovert dimension and gender. This design permitted an assessment of the impact of TTS voice on children's amplitude adaptation to the marine creatures *over time* during tasks 1-3, but also their *shorter-term adaptation* while alternating brief subdialogues between co-present characters during tasks 1 and 2.

To test the generality of any TTS effects, I and E voice types also were tested using different gender instantiations. As a result, the (2) TTS Voice Gender (Male, Female) constituted a separate between-subject factor. The different visual embodiments of animated characters were combined with these TTS voices, and counterbalanced 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.).

3.3. Data Coding and Analysis

All child-computer interaction was videotaped, and conversation was transcribed, digitized, and verified to be free of audio contamination (e.g., coughing, tapping microphone). Sound files that met these criteria were analyzed using two measurement techniques, one automated and the other hand-scored. Coding was restricted to child utterances embedded within a subdialogue after first hearing their software partner's TTS voice.

3.3.1. Amplitude of Vocalic Regions

Using the hand-scoring approach, each vowel was labeled by a trained phonetician, and average amplitude was calculated using Praat speech signal analysis software [12] in those regions marked as vocalic. Vowels generally are known to carry the highest intensity in an utterance [7]. While it is not known exactly which segments or cues serve as the primary acoustic carrier of perceived loudness, it was nonetheless assumed that sounds with louder averages, broader ranges and more articulatory flexibility are better candidates for measuring amplitude effects.

3.3.2. Amplitude Over All Voiced Regions

In the automated method, children's spoken utterances were pitch-tracked, and mean intensity measurements were taken over all pitch-tracked voiced regions [12].

3.3.3. Inter-rater Reliability

Second-scoring of the hand-measured amplitude metric was completed for three subjects by an independent coder. The mean amplitude departure between coders was 0.3 dB, with 80% of amplitude measures matching to within 0.4 dB.

4. RESULTS

A total of 1,969 utterances was available for analysis of amplitude in vocalic regions. A slightly smaller subset of 1,875 utterances was analyzed automatically for amplitude across all voiced regions.

4.1. Presentation Order

The amplitude in vocalic regions averaged 57.01 dB in task 1, increasing to 57.87 dB in task 2, and 58.26 dB in task 3. A repeated measures ANOVA on task order confirmed that there was a significant increase in amplitude as tasks progressed, $F = 6.32$ (df = 2), $p < .009$. This order effect replicated for amplitude measured automatically over all voiced regions, $F = 5.39$ (df=2), $p < .015$.

4.2. TTS Voice Type

As predicted, children's amplitude differed when they conversed with an animated character that spoke with an extrovert versus introvert voice. As shown in Table 2, their average amplitude in vocalic regions increased from 57.19 dB when interacting with the introvert voice to 58.29 dB with the extrovert one, which was a significant difference by repeated measures ANOVA, $F = 15.04$ (df =1), $p < .001$. This change represented a +29% relative increase in energy between the introvert and extrovert conditions. Automatic analysis of amplitude over all voiced regions replicated this finding, $F = 7.01$ (df = 1), $p < .03$.

The repeated measures ANOVA also revealed that children's age had a significant impact on amplitude in vocalic regions, $F = 5.92$ (df = 1), $p < .03$. Children in the older age group had significantly higher amplitude, $t = 3.56$ (df = 13), $p < .006$, two-tailed (separate variances). An ANOVA based on auto-scored amplitudes replicated this result, $F = 7.47$ (df =1), $p < .03$.

Finally, when assessing amplitude in vocalic regions, the repeated measures ANOVA also revealed a significant TTS voice type by gender interaction, $F = 8.52$ (df = 1), $p < .01$. A follow-up comparison indicated that although both groups adapted their amplitude and no particular gender effects were predicted, females adapted amplitude to a greater degree than males, independent t-test (separate variances), $t = 3.17$, (df = 16.4), $p < .006$, two-tailed. For the less sensitive auto-scored metric, however, this interaction was not present.

In addition, children's accommodation of amplitude in vocalic regions was *bi-directional*. A paired t test indicated that amplitude *decreased* significantly when the TTS voice switched from extrovert to introvert (means 58.15 and 57.44, respectively), $t = 3.08$ (df = 10), $p < .006$, one-tailed, as shown in Table 2. It likewise *increased* significantly when the TTS voice switched from introvert to extrovert (means 56.95 and 58.43, respectively), $t = 2.69$ (df = 10), $p < .015$, one-tailed. Table 2 summarizes the amplitude means (vocalic regions) and percentage change in energy between conditions for I to E and E to I conditions. These bi-directional adaptation effects replicated when amplitude was assessed across all voiced regions, with the extrovert to introvert switch once again a significant decrease, paired $t = 5.62$ (df = 7), $p < .0005$, one-tailed, and the introvert to extrovert switch a significant increase, paired $t = 1.98$ (df = 7), $p < .05$, one-tailed.

	Introvert Voice (dB)	Extrovert Voice (dB)	% Change in Energy
Grand Mean	**57.19**	**58.29**	**+29%**
E to I	57.44	58.15	–15%
I to E	56.95	58.43	+41%

Table 2: Bidirectionality of TTS voice type effect

Analysis of hand-scored vocalic regions revealed no significant effect of TTS voice type on children's amplitude when alternating subdialogues between the two characters that were co-present on the interface, $F < 1$. Although no short-term adaptation was found for amplitude measured in vocalic regions, an analysis of all voiced regions did reveal a TTS voice type by visual embodiment interaction, $F = 8.63$ (df = 1), $p < .008$. Follow-up analysis revealed that the combination of an introvert TTS voice with the diminuitive visual embodiment of "Spin the dolphin" elicited significantly lower amplitude speech from children than did marine animals with an extrovert TTS voice, $t = 4.50$ (df = 8), $p < .002$, two-tailed. No other contrasts were significant.

4.3. Individual Differences

In total, 17 of 22 subjects, or 77%, accommodated their amplitude in vocalic regions toward that of their conversational partner, with individual children ranging between +293% to –9% relative change. Table 3 summarizes these individual differences as a relative percentage change in energy from the I to E condition, which is based on a linear scale rather than a logarithmic decibel scale (for conversion, see [10]). The results for amplitude accommodation of all voiced regions replicated this general pattern, with 15 of 16 subjects, or 94%, adapting amplitude toward that of their conversational partner.

Subject	I to E % Energy Change
S1	+293%
S2	+120%
S3	+60%
S4	+56%
S5	+45%
S6	+40%
S7	+34%
S8	+31%
S9	+31%
S10	+26%
S11	+25%
S12	+21%
S13	+20%
S14	+19%
S15	+14%
S16	+7%
S17	+7%
S18	-2%
S19	-4%
S20	-6%
S21	-9%
S22	-9%
Grand Mean	**+37%**

Table 3: Percentage change in energy between introvert and extrovert conditions

4.4. Generality of Effects

No significant difference was found in children's differential adaptation to the introvert and extrovert TTS voices as a function of age or the specific TTS voice they heard (i.e., male versus female prototype TTS voices). That is, the main effect of TTS voice type (E vs. I) generalized across these variables.

5. DISCUSSION

In a conversational interface, 7-to-10-year-old children actively accommodated the amplitude of their software partner, and then readapted their amplitude dynamically when a new TTS voice was introduced. They increased their amplitude when conversing with the louder extroverted character, and dropped it when speaking with the quiet introverted one. The average increase in energy between the introvert and extrovert switch conditions was a substantial +29%, and these amplitude changes were bi-directional, as shown in Table 2. Furthermore, a high degree of consistency was observed in children's amplitude adaptation, with 77-94% of children adapting their amplitude toward that of their conversational partner.

Children's accommodation of their software partner's amplitude was a general phenomenon in several important ways. It occurred independent of the gender of the TTS voice that was used, or the children's specific age or gender group. Although both male and female children adapted their amplitude, one qualification is that the magnitude observed for females was significantly larger for the more sensitive hand-scored metric. Finally, since children interacted with 24 visually distinct characters counter-balanced across tasks, speech adaptations were not limited to a specific visual appearance. In brief, adaptation to the introvert and extrovert TTS voices generalized across all of these variables.

From a methodological standpoint, the hand-labeled vocalic amplitude metric was superior in its measurement sensitivity. Hand-measurement also resulted in preserving more data, since small regions of audio contamination could be excised without discarding data from individual utterances or whole subjects. In contrast, the automated amplitude measure of voiced regions was more easily scored, would be easier to replicate, and averaged data across more of the speech signal (e.g., nasals, voiced fricatives, etc., in addition to vowels). For the present exploratory research purposes, using both metrics together provided a highly convergent pattern of results.

This research and related work on dialogue response latencies [4] suggests that future conversational interfaces that include TTS output could be designed to actively manage hard-to-process features of children's spoken language. For example, low amplitude speech was a significantly greater issue for the younger children in this study, and would be problematic for a speech recognizer. To the extent that animated character design can exploit children's natural inclination to converge with their partner's speech patterns, it may provide an effective tool for transparently guiding their speech to be more processable by recognition technology without explicit instruction, training, or error messages. Future research should examine amplitude and other speech adaptations in adults, as well as prototyping next-generation conversational interfaces that are mutually adaptive.

6. ACKNOWLEDGMENTS

This research was supported by Grants IRI-9530666 and IIS-0117868 from the NSF, Special Extension for Creativity (SEC) Grant IIS-9530666 from NSF, and a gift from Intel Research Council. Thanks to Matt Wesson for programming assistance, Jason Wiles for acting as the science teacher, Cynthia Girand for second scoring, and the students for volunteering.

7. REFERENCES

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

[2] Cassell, J., Sullivan, J., Prevost, S. & Churchill, E. (eds.), 2000, *Embodied Conversational Agents*, MIT Press, Cambr., MA.

[3] Darves, C., Oviatt, S. & Coulston, R., Designing effective conversational interfaces for next-generation educational software, in submission.

[4] Darves, C. & Oviatt, S., Adaptation of users' spoken dialogue patterns in a conversational interface, in submission.

[5] Dehn, D.M. & van Mulken, S., 2000, The impact of animated interface agents: A review of empirical research, *Internat. Jour. of Human-Comp. Studies*, 52: 1-22.

[6] 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 Pub., London, UK, 13-48.

[7] Ladefoged, P., 1993, *A course in phonetics*, Harcourt Brace Jovanovich, Ft. Worth, TX.

[8] Nass, C. & Lee, K.L., 2000, Does computer-generated speech manifest personality? An experimental test of similarity-attraction, *Proc. of CHI '2000*, ACM Press, NY, 329-336.

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

[10] Plack, C. & Carlyon, R., 1995, Loudness perception and intensity coding, in *Hearing: Handbook of Perception and Cognition*, ed. by B. Moore, Acad. Press, San Diego, CA, 123-160.

[11] Potamianos, A., Narayanan, S. & Lee, S., 1997, Automatic speech recognition for children, *Europ. Conf. on Speech Com. & Tech.*, 5: 2371-2374.

[12] Praat speech signal analysis software (URL: www.praat.org)

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

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

[15] Wilpon, J. & Jacobsen, C., 1996, A study of speech recognition for children and the elderly, *Proc. Of ICASSP*, IEEE Press, Atlanta, GA, 349-352.

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