Evaluating TTS Voices for Animated Pedagogical Characters

Courtney Darves, Sharon Oviatt, and Rachel Coulston

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

Abstract

Advances in speech recognition and text-to-speech (TTS) technologies recently have contributed to the development of conversational interfaces that incorporate animated characters. These interfaces potentially are well suited for educational software, since they can engage children as active learners and support question asking skills. In the present research, a simulation study was conducted in which twenty-four 7-to-10-year-old children used speech and pen input to converse directly with animated fish as they learned about marine biology. During these interactions, children became highly engaged, asking an average of 152 questions during a 1-hour session. The specific auditory embodiment of animated characters as text-to-speech (TTS) output also had a significant selective impact on children's engagement in asking science questions. Specifically, children asked +16% more science questions when conversing with an extrovert voice, which resembled the speech of a master teacher (e.g., higher volume and pitch, wider pitch range), rather than an introvert voice, although no differential impact was found on social questions. These findings reveal that conversational interfaces can be designed that effectively stimulate children during learning activities, thereby supporting the goals of next-generation educational software.

I. Introduction

An emerging trend in educational software is the incorporation of animated characters, which can provide an interface design vehicle for engaging children and managing the overall tutorial exchange [5, 7]. When animated characters are embedded within a conversational interface, they quite naturally can become the *central focus* of the content exchange as an interlocutor, rather than playing a subsidiary and sometimes distracting "help-agent" role. As an example, in the course of learning about science, a child could converse directly with an animated parasite or sea creature to extract information about it. The immediacy of such an interaction could be designed to facilitate children's engagement as "active learners," in which they seek answers to questions that they care about as they construct their own understanding of science [4, 8]. Consistent with a *constructivist view* of educational theory, one goal of the present research was to investigate how animated character technology can be designed to bring out the best in student's question asking skills.

While past research on animated characters has confirmed their ability to engage and motivate users [2, 3, 5], it rarely has shown any task-relevant performance enhancement as a function of their presence or specific design [3]. Likewise, most research on animated character design has focused on rendering them with high-fidelity graphics and animation, and on the impact of visual embodiment, but has ignored the question of whether *auditory embodiment* also can provide powerful cues that influence user behavior. However, in a recent study involving web-based book reviews, the TTS voice used for animated characters was found to influence users' self-reported book preferences and purchasing behavior [6]. In another web-based study, the presence of animated agents that actively monitored users' behavior as they worked was found to *decrease* users' performance and increase their anxiety level [9]. Unfortunately, there are few compelling demonstrations that animated characters significantly improve users' learning-oriented behavior in any way during a tutorial exchange.

Since conversational interfaces are social in nature, in the present research the voice characteristics of a "master" teacher were used as a design metaphor for integrating animated characters into an educational software application. The education literature indicates that students respond with increased attention and on-task behavior to dynamic and energetic speech [1, 10], or an *extroverted speech style* that is higher in volume and pitch, and more expanded in pitch range [6, 11]. As a result, we might expect that animated characters responding in an extroverted voice would be more effective in stimulating children's learning-oriented behavior, including their level of spontaneous question asking.

2. Goals of the Study

In the present study, children conversed directly with animated fish using the *I SEE!* interface as they learned about marine biology. This research was designed to:

- Explore whether conversational interaction with animated characters can be engaging for children, as measured by time spent interacting with the software, quantity of spontaneous question asking, and children's self reports
- Determine whether the TTS voice characteristics used for animated characters influence children's learning-oriented behavior (e.g., question asking), and what the implications are for designing educational software
- Assess the overall usability of the *I SEE!* conversational interface prototype

With respect to the second goal, children's queries were compared when they interacted with animated characters embodying different TTS voice profiles. In a comparison of introvert versus extrovert voices, it was predicted that an extrovert voice that shares features in common with master teachers' speech would be more effective in stimulating children to ask task-appropriate questions during learning activities. In particular, it was predicted that children would ask more biology questions when conversing with an extrovert TTS voice (compared with an introvert voice), although no differential impact would occur for general social-interaction questions. The long-term goal of this research is the design of effective conversational interfaces, in particular ones that have a desirable behavioral impact on users for the application being designed.

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 7-10 years, and was gender balanced. Participation was conducted at a local elementary school.

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 permits children to use speech, pen, or multimodal input while conversing with animated software characters as they learn about marine biology. Figure 1 illustrates the *I SEE!* Interface.

During the study, children used the I SEE! application to view and interact with 24 different marine animals (e.g., octopus, shown in Figure 1). The marine animals were animated and available as "conversational partners" who could answer questions about themselves using text-to-speech (TTS) output. An animated "Spin the Dolphin" character, shown in the lower right corner of Figure 1, also was co-present on the screen and available as a conversational partner. 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 conversing with 24 marine animals. The child queried these animals to collect information and build a graph representing information about them (e.g., "Can this animal change colors rapidly?"). Children also were encouraged to ask any questions they wished, and to have fun learning new things about the animals. Therefore, children's spontaneous conversations with the animals primarily were self-initiated, reflecting their own curiosity and interests about these marine creatures.

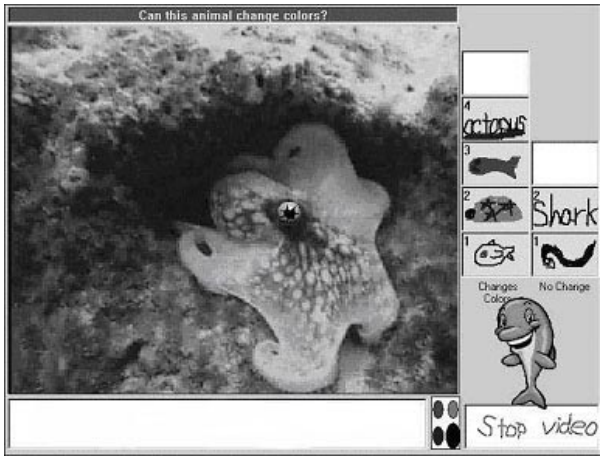

Figure 1: *I SEE!* Interface

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

After each session, the science teacher returned to interview the child about the *I SEE!* interface. Details of the *I SEE!* mobile simulation architecture, its performance, and its use in research with children have been described elsewhere [6].

3.2. Text to Speech

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 included both male and female American English prototype voices, which were further tailored to represent opposite ends of the introvert-extrovert personality spectrum as indicated by the speech signal literature [6, 11]. Introvert and extrovert voices were selected because they are relatively well understood, highly marked paralinguistically, and have been used in previous research on the design of animated characters. 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 these differences in global speech signal features between the introvert and extrovert TTS voices.

Due to pre-loading 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 presented during the study.

3.3. Research Design

The research design for this study was a completely crossed factorial, and the dependent measure was the number and type of questions asked. The main within-subject factor was (1) Type of TTS Voice (Introvert, Extrovert). This factor remained constant for the first 16 animals, but switched for the remaining 8 (from I to E, or E to I). To test the generality of any TTS effects, I and E voices were tested using both male and female voice prototypes, so (2) TTS Voice Gender (Male, Female) constituted a separate between-subject factor. Other between-subject factors included (3) Child Gender (Male, Female) and (4) Child Age (Young, Old), which was categorized using a median split to divide children into a younger (average 8 yrs., 2 mos.) and older (average 9 yrs, 7 mos.) group.

Table 1: Characteristics of the four TTS voice conditions

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

3.4. Data Coding and Analysis

Human-computer interaction was videotaped and conversational interaction transcribed. Children's conversations with the animated characters were coded for (1) time to complete activity, (2) number and type of child questions, and (3) children's self-report comments about the interface and its ease of use.

3.4.1. Time to Complete Activity

For all subjects, total time spent engaged with the *I SEE!* interface after practice was measured to the nearest second.

3.4.2. Number and Type of Self-Initiated Queries

The number and type of children's spontaneous queries to the animated characters and Spin the Dolphin were counted and coded into separate genre types. As described in Table 2, four genres were used to classify the questions into the following categories: (1) Biology, (2) Social, (3) Interface help, and (4) Other questions. In addition, the number of child requests for an animated character to repeat an utterance was counted separately to assess TTS intelligibility. Child commands, responses to system initiations, and simple acknowledgments were not coded.

Table 2: Description of query genres

Genre	Description and Examples
BIOLOGY	Questions about biology. - *What kind of marine animal are you?* - *How do you defend yourself?*
SOCIAL	Questions about social and personal issues. - *What's your name?* - *Are you married?*
HELP	Questions about how to use *I SEE!* interface. - *How do I stop the movie?* - *How do I change the ink color?*
OTHER	Questions on other miscellaneous issues.

3.4.3. Interview Self-Reports

Based on post-experimental interviews, children's appraisal of the interface and its ease of use were summarized, as were their qualitative comments about the animated characters.

3.4.4. Inter-coder Reliability

In total, 17% of child queries were second-scored by an independent coder for genre classifications, and these judgments between coders matched over 99% of the time.

1. RESULTS

Approximately 36 hours of videotape data and 3,643 child queries were coded for genre classifications, of which 3,340 were directed to the animated marine animals, and another 303 to Spin the Dolphin.

4.1. Engagement in Interface and Ease of Use

Even though children were alone in the classroom with no teacher present, they spontaneously asked an average of 152 queries of the animated marine animals while engaged with the interface. The total questions asked per child ranged from 71 to 309. During these interactions, children spent an average of 45.9 minutes engaged in conversation with the animated characters.

In spite of the fact that children were introduced to Spin as a character who could provide them with help using the computer, less than 0.25% of all children's queries to either the animated marine characters or to Spin involved requests for help with the interface, including help constructing graphs. In addition, children rarely

(less than 1% of the time) requested repetition of TTS feedback from the animated characters or Spin, which confirmed that the TTS was adequately intelligible for the present application.

Based on self-reports collected during post-experimental interviews, 100% of the 24 children gave a positive assessment of the interface, with 79% reporting that it was "easy to use," and 96% reporting that they wanted one to own. Typical qualitative comments included that the computer was "cool," "fun," and something they'd "like to play with all day." Children's most common spontaneous comments were that they liked "talking to the animals" (50%), "being able to write and speak to the computer" (29%), and "being able to get answers to questions and learn things" (21%).

In terms of the animated characters, 96% of children assessed them positively, with 83% describing them as being like "friends" or "teachers" (i.e., rather than parents, strangers, or other). Children's engagement with the characters was reflected in the social quality of their conversations. For example, they gave the fish compliments (e.g., "You sure are pretty"), showed empathy toward them (e.g., "I'd never eat fish"), and displayed emotional attachment ("I'll miss you, Spin!").

4.2. Distribution of Question Types

As shown in Table 3, the majority of children's queries (75%) to the animated marine characters focused on marine biology factual information. The remaining questions (24%) primarily were social in nature, with only a small percentage on miscellaneous topics.

Table 3: Distribution of total queries to sea creatures by topic

Genre	Occurrences	Percent of Corpus
BIOLOGY	2493	74.6
SOCIAL	794	23.8
OTHER	53	1.6
INTERFACE HELP	0	0

Children's questions to Spin the Dolphin reflected a markedly different question distribution, as shown in Table 4. When interacting with Spin the Dolphin, the majority (74%) of queries was social in nature, while biology questions comprised only 15% of their questions to Spin.

Table 4: Distribution of total queries to Spin the Dolphin by topic

Genre	Occurrences	Percent of Corpus
BIOLOGY	45	14.8
SOCIAL	223	73.6
OTHER	28	9.2
INTERFACE HELP	8	2.3

4.3. Impact of TTS Voice Type on Child Queries

Children asked more questions overall when interacting with animated marine characters embodying the extrovert TTS voice, compared with the introvert voice (mean 141 vs. 126 questions, respectively). Figure 3 illustrates children's differential level of question asking when interacting with the introvert and extrovert voices, broken down into the two main genre types of biology versus social questions. *A priori* paired t-tests confirmed that children asked a greater number of biology questions when conversing with the extrovert voice, rather than the introvert one (mean 108 and 93 biology queries, respectively), paired t=2.08 (df=23), $p < .025$, one-tailed. This represented a 16% overall increase in children's question asking when interacting with the extrovert TTS voice. Furthermore, the majority of children, or 19 of 24, responded in this manner. In contrast, no significant difference was found in the level of children's social queries when interacting with these two voice types, $t < 1$, N.S. Comparisons involving the young vs. old age groups and male vs. female children all confirmed that the extrovert TTS voice stimulated significantly and selectively more biology queries. Finally, these results generalized across testing with the male and female TTS voice prototypes.

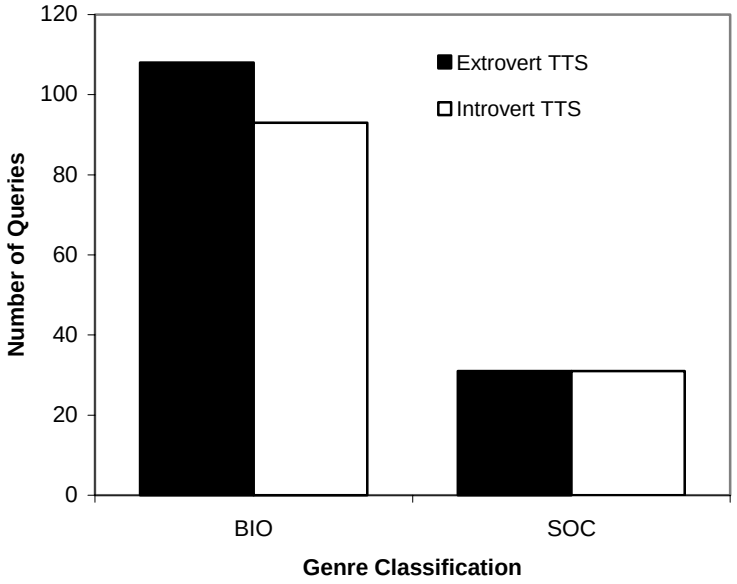

Figure 3: Number of biology and social queries asked by children when interacting with characters using extrovert versus introvert TTS voices

5. Discussion

Auditory embodiment alone, independent of an animated character's visual appearance, can be highly influential in stimulating users' behavior in task-appropriate ways. In the present conversational interface, children's question asking was substantially affected by the acoustic-prosodic features of the TTS output they heard, independent of the lexical content. When interacting with the extrovert voice in the *I SEE!* educational application, which in many ways represented the rhetorical style of a master teacher [1, 10], children were stimulated to ask 16% more marine biology questions. In contrast, children's general social questions were not differentially affected by the same introvert and extrovert voices. In other words, using an extrovert TTS voice that was louder, faster, higher in pitch and wider in pitch range had a *selective impact on children's educationally-relevant question asking behavior*. The extrovert voice essentially was more successful in motivating and managing a tutorial exchange. This finding underscores the important role of TTS design in the success of future conversational interfaces. It also suggests the importance of matching an appropriate TTS voice to an application domain in order to ensure a desirable impact on user behavior.

The two types of animated characters present in the interface (Spin versus the marine characters) elicited distinct question profiles from the children. Whereas the marine creatures, represented through high-fidelity National Geographic video footage, elicited primarily biology and factual questions, children asked predominantly social questions of Spin the Dolphin, an animated cartoon. It is unclear whether this difference was caused by Spin's cartoon graphical rendering or to other factors unique to Spin the Dolphin. As the interface vehicle, Spin the Dolphin appeared throughout the activity, perhaps fostering a social relationship, and also had a small repertoire of jokes. The unexpected difference in question profiles to Spin and the marine characters suggests further research into the design of effective visual embodiment, duration of interaction, and dialogue style to promote educationally-relevant question asking.

As a conversational interface prototype, *I SEE!* was highly intuitive and easy to use, as indicated by the extremely low rate of child requests for Spin's help (i.e., 0.25% of all queries). After only brief exposure, young children were able to converse with the fish to extract large amounts of information about marine biology, and to construct graphs tabulating data about them. This interface was highly effective at *directly engaging* young children and stimulating learning-oriented behavior. When left alone, children spontaneously asked an average of 152 questions of the digital fish, and in some cases over 300 questions. The largest percentage, or 75%, were marine biology questions. Children's most common positive comment about the computer was that they liked "talking to the animals," which may in part reflect the "immediacy characteristics" of this interface [7], as well as

the self-reinforcing nature of conversation itself. The long-term goal of this research is the design of effective conversational interfaces, in particular ones that have a task-appropriate behavioral impact on users for the application being designed.

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. Finally, we are grateful to the students who participated in this research.

7. References

[1] Bettencourt, E., Gillett, M., Gall, M. & Hull, R., 1983, "Effects of teacher enthusiasm training on student on-task behavior and achievement," *Amer. Educ. Res. Jour.*, 20 (4): 435-450.

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

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

[4] "Everyday Classroom Tools Project," Harvard Grad. School of Educ. (http://hea-www.harvard.edu/ECT/Inquiry/inquiry1.html)

[5] Lester, J.C., Converse, S.A., Stone, B.A., Kahler, S., Barlow, T., 1997, "Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation," *Proc. of the Eighth World Conf. on A. I. in Educ.*, 23-30, Kobe, Japan.

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

[7] 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, Cambridge, MA, 319-343.

[8] Richmond, V., Gorham, J. & McCroskey, J., 1987, "The relationship between selected immediacy behaviors and cognitive learning," *Commun. Yearbook*, 10: 574-590.

[9] Rickenberg, R. & Reeves, B., 2000, "The effects of animated characters on anxiety, task performance, and evaluations of user interfaces," *Proc. of the Conf. on Human Factors in Comp. Systems: CHI '2000*, ACM Press, NY, 49-56.

[10] Sallinen-Kuparinen, A., "Teacher communicator style," *Commun. Educ.*, 41: 153-166.

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