Comparison of Child-Human and Child-Computer Interactions based on Manual Annotations

Matthew Black, Jeannette Chang, Jonathan Chang, and Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL) – http://sail.usc.edu
University of Southern California, Los Angeles, CA, USA
matthepb@usc.edu, jeannetc@usc.edu, jonathkc@usc.edu, shri@sipi.usc.edu

ABSTRACT

Technological advancements in recent years have been accompanied by a notable increase in research related to conversational child-machine interfaces. The technology has many applications from entertainment to education. In order to integrate this technology successfully we, however, need to understand the key differences (if any exist) in how children interact with machines versus how they interact with humans. Such knowledge could inform the design of more child-appropriate interfaces as well as highlight any distinct characteristics of child-computer interactions that may be crucial for specific applications. In this paper, we analyze a subset of the Little CHIMP corpus, in which preschool aged children have a series of conversations with a human moderator and a Wizard-of-Oz controlled computer character. We first manually transcribed and annotated the data using an objective audio-visual behavior coding scheme. We next extracted features exemplifying language and social communication from these transcriptions and annotations and performed statistical hypothesis tests comparing the child-human and child-computer interactions. Finally, we discuss the differences between these two dyadic conversations.

Categories and Subject Descriptors

H.1.2 [**Information Interfaces and Presentation**]: User/Machine Systems–*Human information processing;* H.5.2 [**Information Interfaces and Presentation**]: User Interfaces–*Natural language*.

General Terms

Design, Experimentation, Human Factors, Languages

Keywords

Child-adult vs. child-machine interactions, audio-video analysis, embodied conversational agent, human annotation, Little CHIMP

1. INTRODUCTION

Research in creating conversational child-machine interfaces [13] is an area of growing interest, with many applications ranging from automated tutors [5] to applications encouraging play and creativity [4]. As technology use has become prevalent among young children, improving the quality of child-machine interactions has become a topic of increasing significance. Successfully integrating machines into useful applications requires that we understand the differences between how children interact with humans and how they interact with machines and computers.

Previous work in [15] introduced the notion of the *media equation*, which asserts that people treat computers like humans. Other studies such as in [16] disagreed, showing that people who believed they were talking to a human spent more time establishing an interpersonal relationship, compared to when they knew they were speaking to a computer. Cassell et. al assert that embodied conversational agents in social settings should ultimately be judged by their effectiveness in eliciting natural behaviors (both verbal and non-verbal) from the user, i.e. the user should behave as if he or she is talking to a real person [2,3].

In a study with children ages 6-10 years, Oviatt showed that children's speech contained three times the number of disfluencies when interacting with a human versus animated computer characters and noted that children's speech in computer interactions was clearer but hyperarticulated [14]. In another study, children's speaking style and pitch patterns were found to adapt to those of their speaking dyad. Additionally, children talking to extroverted computer agents (as defined by pitch range, volume, and rate of speech) asked more questions than those talking to introverted computer agents; no significant difference was found across the children's age or gender [6,7,8].

In this paper, we compare briefing and debriefing conversations between child-adult and child-computer interactions in preschool aged children. We manually transcribed the sessions and annotated them with a number of audio-visual cues (e.g., whispered/soft/loud voice, hand movements, head orientation). These manual transcriptions and annotations served as our basis of comparison between the child-human and child-computer sessions. Section 2 describes the corpora. Section 3 explains our annotation scheme. Section 4 describes the features we extracted from the annotations, which are analyzed in Section 5 by comparing the child-human and child-computer interactions. Section 6 summarizes our findings and introduces future research.

2. CORPUS

We used data collected as part of the Little Children's Interactive Media Project (CHIMP) [11,12]. Children, ages 4-7 years, participated in five interactive sessions in the following order: human briefing, computer briefing, computer game, computer debriefing, and human debriefing. For the human sessions, the subjects sat at a desk and spoke with an adult human moderator. For the computer sessions, the subjects spoke with an embodied

ICMI-MLMI'09 Workshop on Child, Computer and Interaction
November 5, 2009, Cambridge, MA, USA

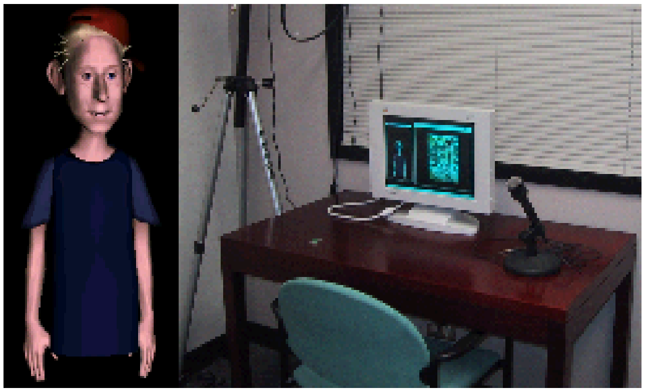
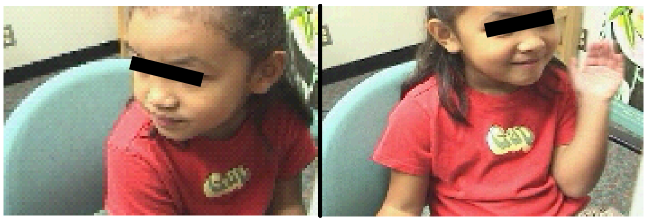


Figure 1. "Josh," the embodied computer agent (left), and a photograph of the experiment room set-up (right).



[redacted]

[redacted]

Figure 2. Screenshot of a child-adult session (left) and a child-computer session (right). Eyes covered for privacy reasons.

computer agent called "Josh," who was displayed on a computer monitor (Figure 1). Josh was controlled in a Wizard-of-Oz manner by an experimenter observing the interaction in a separate room; this was done to ensure a reasonable flow in the child-computer conversations.

During the briefing sessions, the subjects answered a series of basic questions (e.g., about their latest birthday party, their family, summer vacation, past experiences with computers). In the computer game session, Josh asked a number of age-appropriate questions (e.g., counting the number of hidden objects in a picture, ordering a sequence of pictures, identifying strange/funny occurrences in a picture scenario) and provided feedback to the child. The debriefing sessions consisted of discussions about the subjects' experiences during the computer game sessions (e.g., whether they had fun, what game they liked best, whether the games were too hard). All sessions were scripted, which helped maintain consistency between the human and computer sessions and across subjects. Many of the questions asked by the adult and computer overlapped in content, but the scripts were not identical.

Of the approximately 50 children who participated in the Little CHIMP study, we selected 9 for this study (Table 1). These children were selected due to access to word-level transcriptions from previous studies [1,11,12,17,18]. For this paper, we analyzed the briefing and debriefing sessions, which gave us comparable interactions in the child-human and child-computer categories. In total, we have 91 minutes of audio-visual data, 61 minutes from human sessions and 30 minutes from computer sessions. The data consist of a front-side view of the child (Figure 2) and a single channel of audio from a table-top microphone (Figure 1).

Table 1. Age and gender of the 9 subjects in this study

Age (years)	4.3	4.7	4.8	5.6	5.8	5.9	6.1	6.4	6.8
Gender	M	M	F	F	F	F	F	M	M

Table 2. Total number of times each annotation stream/label was marked for the nine children in the corpus

Annotation Stream (Total #)	*Annotation Label (Total #)*
Voicing Type (181)	Question (9), Reduced (103), Strong (56), Whispered (13)
Disfluencies (376)	Elongation (66), Repair (40), False start (28), Repetition (70), Filled pause (172)
Body Movement (383)	Change in orientation (174), Leaning (30), Shrugging (33), Slouching (39), Random (107)
Hand Movement (265)	Accompaniment (17), Point (7), Descriptive (27), Wave (2), Meaningless (201), Other (11)
Head Movement (284)	Shake no (34), Shake yes (137), Other (113)
Head Orientation (676)	To camera (127), To adult (36), To screen (171), Downwards (206), Other (136)
Mouth Movement (213)	Frown (25), Smile (155), Other (33)
Eyebrow Movement (144)	Furrowed (59), Raised (76), Other (9)

3. ANNOTATION SCHEME

We transcribed all four interactions for each of the 9 children at the word-level (the children's speech *and* the human's/computer's speech). In addition, we annotated the videos with high-level categorical labels (Table 2) using the Anvil software [9]; this allowed us to mark the start and end times of each observable event and track the different annotation streams in parallel.

This annotation scheme was chosen to capture the dynamics of the interaction using objective observational coding. It is similar to the annotation scheme used in our previous work in [1] (please see for detailed descriptions), with two key differences: 1) we explicitly marked eyebrow and mouth movements in this paper, rather than subjective facial expressions, and 2) we incorporated some of the observational codes described in the autism diagnostic observation schedule (ADOS) manual [10], as recommended by clinical psychologists. For example, we coded descriptive hand movements (using hands to represent an object/event) and accompaniment hand movements (using hands to help convey or emphasize lexical meaning), since these gestures often are not used by children with autism spectrum disorder (ASD). We employed some of these ASD-relevant codes in anticipation of eventually collecting similar interaction data from children with ASD. The typically developing children analyzed in this paper will serve as a comparison for this future research.

4. FEATURE EXTRACTION

We extracted features from the human transcriptions and annotations at the turn-level. The human/computer turn is defined as the time they start speaking to the time they stop speaking, and

Table 3. Statistics for the four interaction session types

Turn Statistic	*Human Moderator*		*Josh (Computer)*	
	Brief	Debrief	Brief	Debrief
Avg. session length (sec)	290	116	128	72
Total number of turns	388	157	158	87
Mean # turns per session	43	17	18	10

the child turn is defined as the time in between. Back-channels (e.g., "uh huh") and other interruptions made by the child/adult/computer were marked but not labeled as separate turns. Table 3 shows statistics for each session type; the session lengths were largely determined by the length of the scripts.

To extract meaningful features from the word-level transcriptions, we had initially planned to train unigram/bigram language models and/or a bag-of-words classifier. However, since the scripts for the human moderator and Josh were not identical, a direct lexical comparison of the children's responses was not possible. As a result, we limited our analysis of the transcriptions to the following turn-level features: number of words spoken per turn, turn duration, speaking rate, speaking response time (length of the pause at the beginning of the turn), and the presence/absence of an interruption by the other speaker, including backchannels.

For the annotation streams listed in Table 2, we extracted a binary presence/absence feature for each label for all turns. Since many of the labels were only sparsely used, we also extracted this binary feature for each annotation stream. For example, rather than just extracting features for the label "false start," we also extracted features for all disfluencies.

5. COMPARISON OF CHILD-ADULT AND CHILD-COMPUTER INTERACTIONS

We compared the child-adult and child-computer interactions by analyzing differences in distributions/statistics from the turn-level features discussed in Section 4. We did not analyze the differences between the briefing and debriefing sessions in this paper. We started with the features derived from the word-level transcriptions. Figure 3 compares the histograms of the number of words spoken by the children for the child-adult and child-computer sessions. It shows that the children were more likely to say nothing during turns with an adult (relying instead on gestures to communicate). It also shows that children were more likely to give one word answers to a computer. Lastly, the longer "tails" on the histogram demonstrate that the children were more verbose for some turns when speaking to an adult, compared to a computer. See Table 4 for statistical significance values and to see the fraction of children that followed these trends.

Figure 4 compares histograms of the length of the children's turns up to 15 seconds for the two conditions (there was no significant difference in the percentage of turns lasting longer than 15 seconds for the two types of interactions). Figure 5 shows that on average the children *spoke* slower to the computer character, and Figure 6 shows that the children also *responded* slower when speaking to the computer character. As shown in Table 4, the children were nearly twice as likely to take more than 2 seconds to respond verbally to the computer character, a significant difference in proportion with $p < 0.05$. In addition, we found that the children were more than twice as likely to interrupt the human moderator (14.6% of turns), compared to the computer agent (6.46%), a significant difference with $p < 0.005$.

It is important to take context into account when analyzing these numbers. Compared to the human moderator, on average, Josh spoke significantly fewer words per turn, spoke slower, took longer to respond to what the child said, and interrupted the child less frequently. Therefore, these significant differences in the speaking style of the children may be due to the children adapting to the speaking style of their dyad (as also suggested in [6,7]). That is, the children may be treating the computer agent like a person, but it just so happens that the computer agent is speaking differently (and perhaps unnaturally) compared to the human moderator. Another plausible explanation for these significant differences could be that the children were hesitant towards speaking to a computer agent, since doing so was a new experience for them.

We now discuss the features derived from the annotations listed in Table 2. Table 4 shows the annotation streams/labels that had a significant difference (p < 0.05) in the percentage of turns they were marked by comparing the child-adult versus child-computer sessions. It also shows the fraction of children that followed this trend. Table 4 shows that 8 out of the 9 children were more likely to speak quietly (using a reduced voice) when interacting with the human moderator compared to when speaking with the computer agent (p < 0.01). There were also more disfluencies in the children's speech during child-adult interactions compared to the child-computer interactions (p < 0.05). Children used repetitions almost four times as often when speaking to a human (9.54% of child turns) compared to when speaking to the computer (2.45% of turns), a significant difference with p < 0.001. This suggests that the children were more articulate when speaking with the computer agent, which may be a result of the children adapting to the computer agent's own clear, articulate speech.

Children also physically moved around much more when speaking to the human moderator, changing their body orientation during 17.8% of their turns (compared to 10.2% of their turns when interacting with the computer agent). They were more inclined to slouch during child-human interactions (p < 0.001). Moreover, the children's head orientation was directed away from the dyadic conversation for more than half of the child turns during child-adult interactions (55.2%), compared to less than a third in the child-computer interactions (29.0%). The children tended to spend more of the conversation looking downward when speaking with the human versus the computer (23.9% versus 10.6%). Overall, the children appeared to fidget less during interactions with the computer agent, making significantly fewer random body and hand movements (both p < 0.0005). These differences suggest that the children were more engaged and better able to concentrate when interacting with the computer agent compared to the human. However, the conversations with the computer agent may have maintained their interest due in part to the shorter duration of the computer sessions compared to the human sessions.

The children used more descriptive hand gestures when speaking to the human moderator compared to the computer (p < 0.01). Descriptive hand gestures were defined as those that were directly associated with the content of the child's speech, such as miming how a toy works while explaining with words.

Table 4: Various child turn percentages (we used a difference in proportions test to verify statistical significance)

Annotation Stream	*Annotation Label*	*% of Child Turns*		*Significance of Difference*	*Children Following this Trend*		
		Child-Adult	Child-Computer		Fraction	Gender (F:M)	Mean age
Word-Level Transcriptions	# of words in turn = 0	7.34 %	1.22 %	p < 0.001	5/9	3:2	5.7
	# of words in turn = 1	37.5 %	49.0 %	p < 0.01	9/9	5:4	5.6
	# of words in turn > 15	10.3 %	2.04 %	p < 0.001	9/9	5:4	5.6
	Response time > 2 sec	5.98 %	11.0 %	p < 0.05	6/9	4:2	5.7
	Child Interruption	14.6 %	6.46 %	p < 0.005	8/9	5:3	5.5
Voice Quality	(All Labels)	21.7 %	14.3 %	p < 0.01	7/9	4:3	5.7
	Reduced	12.7 %	6.94 %	p < 0.01	8/9	4:4	5.6
Disfluencies	(All Labels)	30.6 %	24.9 %	p < 0.05	8/9	5:3	5.5
	Repetitions	9.54 %	2.45 %	p < 0.001	8/9	5:3	5.5
Body Movement	(All Labels)	39.6 %	20.8 %	p < 0.0001	8/9	5:3	5.5
	Change in Orientation	17.8 %	10.2 %	p < 0.005	6/9	4:2	5.4
	Random Movements	15.8 %	6.94 %	p < 0.0005	6/9	4:2	5.6
	Slouching	6.42 %	0.41 %	p < 0.001	7/9	3:4	5.5
Hand Movement	(All Labels)	34.9 %	19.6 %	p < 0.0005	7/9	3:4	5.5
	Descriptive Gestures	3.49 %	0.41 %	p < 0.01	6/9	3:3	5.7
	Meaningless Gestures	30.8 %	17.6 %	p < 0.0001	7/9	3:4	5.5
Head Movement	Shake 'Yes'	14.3 %	9.39 %	p < 0.05	6/9	4:2	5.8
Head Orientation	(Not at dyad)	55.2 %	29.0 %	p < 0.0001	7/9	3:4	5.6
	Downwards	23.9 %	10.6 %	p < 0.0001	6/9	3:3	5.7
Mouth Movement	(All Labels)	22.8 %	28.6 %	p < 0.05	5/9	2:3	5.5
	Smile	16.3 %	26.5 %	p < 0.0005	6/9	3:3	5.3
Eyebrow Movement	(All Labels)	16.7 %	8.16 %	p < 0.001	7/9	4:3	5.4
	Furrowed	6.97 %	3.67 %	p < 0.05	6/9	3:3	5.9
	Raised	11.2 %	4.08 %	p < 0.001	7/9	4:3	5.4

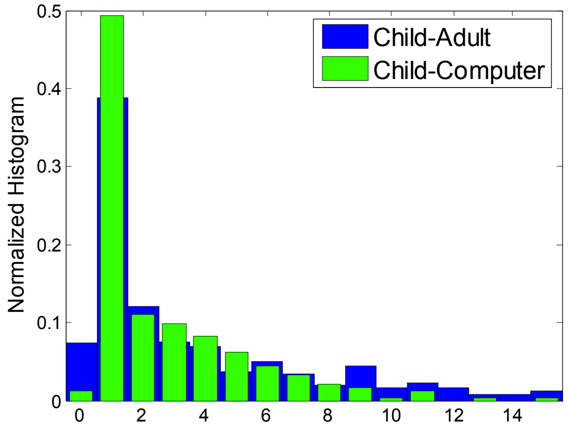


Figure 3. Normalized histograms of the number of words spoken by the children for each turn.

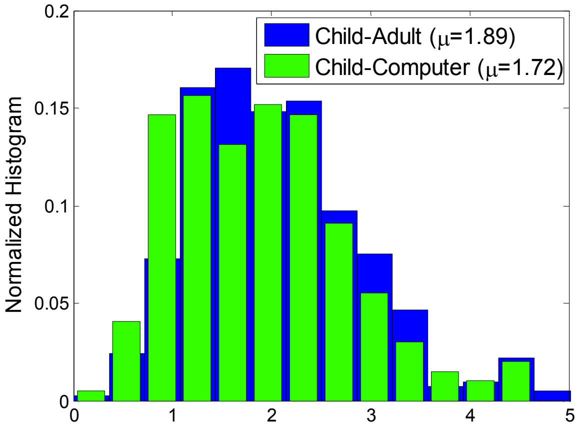


Figure 5. Histograms of children's rate of speech (words/sec). The mean rate was slower for the computer sessions (p < 0.05).

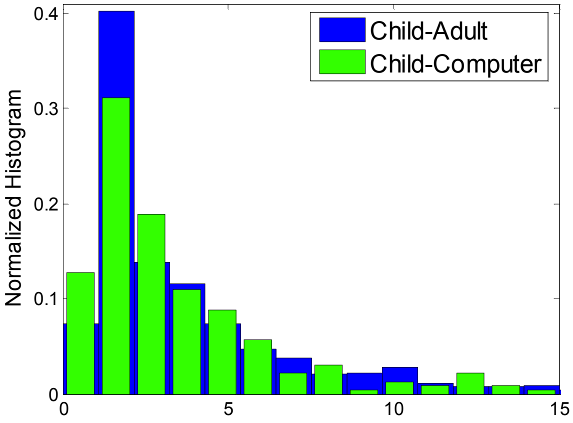


Figure 4. Histograms of child turn lengths (seconds).

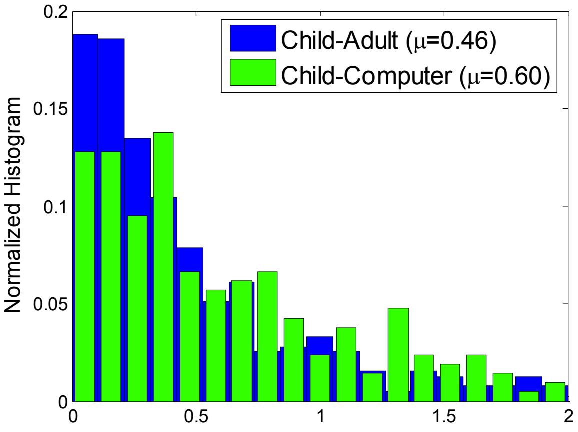


Figure 6. Histograms of children's response times (seconds).

In addition, children were 50 percent more likely to nod their head to indicate a 'yes' (14.3% versus 9.39%) and twice as likely to use expressive eyebrow movements (16.7% versus 8.16%) while speaking with humans. Again, it is possible that these differences are a result of the children adapting to their conversational partner's interaction style. For example, the computer agent's eyebrow movements and hand gestures were very limited compared to those of the human moderator. In response, the children seemed to adopt a similar interaction style, relying more heavily on verbal streams when talking to the computer agent. Alternatively, the children may have perceived that the computer agent was unable to interpret their gestures and motions and therefore did not make an effort to incorporate gestures into their conversation. The one non-verbal cue used significantly more often by children when speaking to the computer was smiling ($p < 0.05$).

We found no significant differences between the gender of the children, but the older children tended to use furrowed eyebrow movements and head shakes more often than the younger children. We also found that younger children smiled more often than older ones, suggesting that they were most amused by the computer character. Since only 9 children were analyzed in this study, these findings may not generalize well.

6. SUMMARY & FUTURE WORK

In this study, we annotated children's interactions based on audio-visual cues during briefing and debriefing sessions with a human moderator and an embodied computer agent. We compared the speaking style and the patterning of these audio-visual cues at the turn-level. We found that the children responded slower, spoke louder and less verbosely, and used fewer disfluencies during child-computer interactions. They were also more inclined to smile, sit up, and stay oriented toward their dyad when speaking with the computer character versus the adult moderator. On the other hand, children used more head, hand, and eyebrow gestures when speaking to the human moderator.

As mentioned earlier, one of the limitations of the study is the small number of children we analyzed. Furthermore, we only used one instantiation of an embodied computer agent; we did not vary the appearance, voice, or movements for this study. It would be interesting to collect another database using a computer agent with greater expressiveness and flexibility. Perhaps the children's response to a computer agent with human-like expressiveness would be more comparable to their interactions with human moderators. Future studies will need to record the gestures of the human moderator (not just the child), so that a comparison of non-verbal cues between the human moderator and computer character can be made; this is especially important in social contexts where gestures play an important role in conversation flow and turn-taking.

There are two immediate design implications of this study: identifying areas in which computer conversational agents currently excel and developing more appropriate computer conversational agents for different interaction domains. Despite the limitations of the computer agent in this study, it appeared to be better at capturing and maintaining children's interest and seems particularly suited for educational and therapeutic applications.

Our findings support the hypothesis that children tend to adapt to and emulate the speaking style of their conversational partner, as suggested by [6,7]. Further investigations can examine this entrainment phenomenon to develop computer agents that dynamically adapt to their speaking dyad to emulate the flow of human conversation. In addition, future research can look further into how to design computer agents that are tailored to a specific application. For example, computer agents that move less and speak clearly may be more appropriate for literacy tutors, while more expressive extroverted computer agents may be more appropriate for computer games encouraging play and creativity.

Another study of interest might be to explore the impact of the differences in the perceived age of these embodied conversational agents. For example, would children respond differently to an agent that looked like a young child versus one that looked like a middle-aged adult or an elderly person? Would children be quicker to befriend the computer agents that appeared younger or closer in age to the children themselves? Learning how children react to different types of embodied conversational agents may help us create systems that interact efficiently and naturally with child users.

We are particularly interested to see if our findings in this study would also be true for children diagnosed with ASD. We have planned preliminary experiments to see how higher-functioning children with ASD communicate with the computer agent used in this study, compared to a clinical psychologist. Future experiments will attempt to use computer agents as a tool that clinical psychologists can use to help teach autistic children social-communicative skills they might not naturally learn.

Lastly, the coding completed as part of this paper can also be used for an oracle study in automatic annotation. The observational events manually labeled in this study could be automatically detected using speech/image/video processing algorithms in future work. While the videos in this study are not ideally suited for this purpose, research with other corpora can build upon our proposed annotation scheme.

7. ACKNOWLEDGMENTS

This work was supported in part by the NSF. Special thanks to Serdar Yildirim, Sonia Khurana, Marni Landes, and Jason Maynard for collecting and transcribing the data.

8. REFERENCES

[1] Black, M., Chang, J., and Narayanan, S. 2008. An empirical analysis of user uncertainty in problem-solving child-machine interactions. In Proc. of the Workshop on Child, Computer and Interaction (Chania, Greece, Oct. 2008).

[2] Cassell, J., and Tartaro, A. 2007. Intersubjectivity in human-agent interaction. Interaction Studies 8-3:391-410.

[3] Cassell, J. 2001. Embodied conversational agents: representation and intelligence in user interfaces. AI Magazine, vol. 22-4:67-83 (Winter 2001).

[4] Cassell, J. and Ryokai, K. 2001. Making space for voice: technologies to support children's fantasy and storytelling. Personal Technologies 5(3): 203-224.

[5] Cosi, P., Delmonte, R., Biscetti, S., Cole, R. A., Pellom, B., Vuren, S. van. 2004. Italian literacy tutor - tools and

technologies for individuals with cognitive disabilities. In Proc. of ICALL (Venice, Italy, June 2004).

[6] Coulston, R., Oviatt, S. L., and Darves, C. 2002. Amplitude convergence in children's conversational speech with animated personas. In Proc. of ICSLP (Denver, CO, USA, Sept. 2002).

[7] Darves, C. and Oviatt, S. L. 2002. Adaptation of users' spoken dialogue patterns in a conversational interface. In Proc. of ICSLP (Denver, CO, USA, Sept. 2002).

[8] Darves, C. and Oviatt, S. L. 2004. Talking to digital fish: Designing effective conversational interfaces for educational software. From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer: Dordrecht, 2004, 271-292.

[9] Kipp, M. 2004. Gesture generation by imitation - from human behavior to computer character animation. dissertation.com (Boca Raton, Florida, Dec. 2004).

[10] Lord, C., Rutter, M., DiLavore, P., and Risi, S. 1999. Autism diagnostic observation schedule: manual. Los Angeles: Western Psychological Services.

[11] Montanari, S., Yildirim, S., Andersen, E., and Narayanan, S. 2004. Reference marking in children's computer-directed speech: An integrated analysis of discourse and gesture. In Proc. of ICSLP (Jeju, Korea, Oct. 2004).

[12] Montanari, S., Yildirim, S., Khurana, S., Landes, M., Lawyer, L., Andersen, E., and Narayanan, S. 2004. Analyzing the interplay between spoken language and gestural cues in conversational child-machine interactions in pre/early literate age group. In Proc. of InStil (Venice, Italy, July 2004).

[13] Narayanan, S. and Potamianos, A. 2002. Creating conversational interfaces for children. IEEE Transactions on Speech and Audio Processing, vol. 10-2:65-78 (Feb. 2002).

[14] Oviatt, S. L. 2000. Talking to thimble jellies: children's conversational speech with animated characters. In Proc. of ICSLP (Beijing, China, Oct. 2000).

[15] Reeves B. and Nass C. 1996. The media equation: how people treat computers, television, and new media like real people and places, Cambridge University Press, 1996.

[16] Shechtman, N., Horowitz, L. M. 2003. Media inequality in conversation: how people behave differently when interacting with computers and people. In Proc. of SIGCHI (Ft. Lauderdale, Florida, April 2003).

[17] Yildirim, S., Lee, C. M., Lee, S., Potamianos, A., and Narayanan, S. 2005. Detecting politeness and frustration state of a child in a conversational computer game. In Proc. of Eurospeech (Lisbon, Portugal, 2005).

[18] Yildirim, S. and Narayanan, S. 2009. Automatic detection of disfluency boundaries in spontaneous speech of children using audiovisual information. IEEE Transactions on Speech Audio and Language Processing 17:2-12 (Jan. 2009).