Analyzing the interplay between spoken language and gestural cues in conversational child-machine interactions in pre/early literate age groups

Simona Montanari, Serdar Yildirim, Sonia Khurana, Marni Landes, Lewis Lawyer,

Elaine Andersen and Shrikanth Narayanan

Integrated Media Systems Center, Speech Analysis and Interpretation Lab
Department of Linguistics, Department of Electrical Engineering,
University of Southern California, Los Angeles
[montanar, yildirim, skhurana, landes, lawyer, eanderse]@usc.edu, shri@sipi.usc.edu

Abstract

This paper reports on analysis of interplay between verbal and gestural cues in conversational child-machine interactions in pre-literate age groups. In particular, this study focuses on children's responses to trouble spots (intra-utterance self-repairs, filled pauses), reference marking, and gestural cues (i.e. pointing or touching screen) that children employ in computer-directed children-computer interaction using a data from 15 children, 3 to 6 years old. Our gestural analysis indicated that with increasing age, children rely on purely verbal utterances rather than utterances accompanied by gestures. Co-analysis of gestural use and intra-utterance repairs revealed that children were less likely to produce intra-utterance repairs when they are multimodal. Also, results revealed that children were less likely to use filled pauses when they are interacting multimodally. In addition, the more children rely on purely verbal utterance, the more they use full-form, explicit noun phrases rather than pro-forms such as deictic adverbs and deictic numerals.

1. Introduction

Human-to-human communication is mostly comprised of speech, hand and head movements, facial expressions, and gaze. Information from all those modalities helps us to understand each other better. By providing input from different communication channels to human-computer interaction (HCI) systems, it would be possible to add more robustness and naturalness to machine-human interaction. Indeed, the first step in achieving this is to better understand the interplay between the cues from the different communication modalities. This paper represents an effort in that direction.

With limited fine motor skills and limited ability to read, write or type, young children are primary potential beneficiaries of computers that use conversational interfaces for interactive tutoring and computer instruction. Computer agents must be tailored to understand a child's *intent* while providing a natural, engaging experience [1, 2]. However, in order to achieve this, we need to understand how the speech and gestures of children interacting with computers can be used to gauge the child's communicative state.

Children's behavior in interacting with a computer is different from those of adults, both speech only and multimodal communication scenarios. For instance, Arunachalam et al. [3] analyzed politeness and frustration language in child-machine interactions. Their results indicated that older children (10-14 years old) use overt politeness markers and more polite information requests compared to the younger ones (6-8 years old). Also, they showed that younger children expressed frustration verbally more than the older ones. Xiao et al. [4] analyzed children's speech and pen-based multimodal integration patterns and compared to those of adults. Their results indicated that children were more often integrated speech and gesture (pen input in their case) simultaneously than adults.

In this paper, we examine speech repairs and reference marking in children's computer-directed speech along with gestural cues that they employ while interacting with a computer agent. The first study relates to examining children's responses to trouble spots and their pacing needs; in particular, this includes intra-utterance self-repairs and filled pauses, along with the gesture patterns during those disfluencies. Children's verbal responses to trouble spots, and gestural patterns during those responses, are expected to differ from those in adults. It has been shown that modification gesture patterns have a high correlation with content replacement speech repairs in adult-adult interaction [5]. It has also been shown that children interact more multimodally when they are needed to repair an utterance than when they are fluent [4].

The second study examines reference marking in children's computer-directed speech. It has been shown that young children speech is referentially implicit and lacking in overall cohesion and coherence due to inappropriate alteration of full noun phrases and pronouns to talk about new and previously mentioned referents [6]. It is hypothesized that young children will compensate their inappropriate use of referential devices by switching to multimodality.

The rest of the paper is organized as follows. The data and methods are described in Section 2. The analysis results and discussion are given in Section 3. Conclusions are provided in Section 4.

2. Methods

In this study we analyzed audio/video data from 15 children aged 3-6 interacting with both computer and human agents while resolving a series of age-appropriate cognitive challenges. A Wizard of Oz

(WOZ) design where a hidden human controls the actions of computer agent was used for data collection. The experiment contains five tasks. First, the child subject is briefed by the human interviewer and the human interviewer introduces the computer agent to the child. This is followed by a briefing by the computer agent. Then, the child is asked to resolve a series of age-appropriate cognitive challenges, i.e. pattern recognition, sorting and category membership, all directed by the computer agent. After completion of the experimental battery, the child is debriefed by the computer agent. The experiment is concluded with a debriefing by the human experimenter.

Transcriptions of synchronized speech and gestural data from audio and video recordings of each session were annotated by native English speakers. The PRAAT tool [7] was used for speech transcriptions by following a modified version of the CHILDES annotation format [8]. Gestures (hand, face positions and orientations) were aligned with the speech transcriptions and spoken discourse annotations using the ANVIL toolkit [9]. Time-synchronous visualization is an output of the system. These annotated data were used for analysis and modeling.

3. Results and Discussion

3.1 The Gestural Analysis

In order to determine which modalities the children employed in the context of our experiments, we calculated the distribution of utterance type, i.e. verbal only, multimodal (verbal-gesture), and gestural only, in the data. Because most of the gestures were in the form of touching and pointing to the screen and of nods and head shakes we only counted touching, pointing, nodding and head shaking as gestures. The results of the quantitative analysis are given in Figure 1.

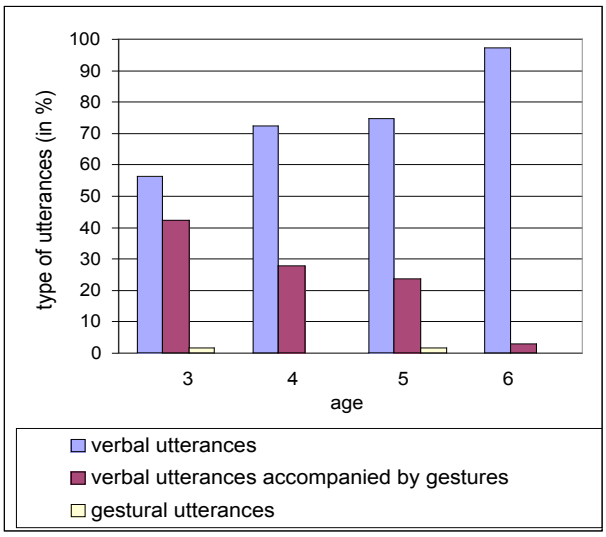

Figure 1. The mean percentage of utterance types (verbal, gestural, and verbal accompanied by gestures) employed by the 3-, 4-, 5- and 6-year-olds while interacting with the computer agent.

As can be seen, multimodal interactions consistently decrease with increasing age. For instance, the percentage of multimodal utterances for the 3-years olds is 42% as opposed to 2.2% for the 6-year-olds. Also, an ANOVA (Analysis of Variance) indicated that the effect of age is significant on the type of interaction that children prefer ($p<0.05$). Finally, it appears that for all age groups gesture-only interactions are very rare while interacting with a computer agent.

3.2 Children's Responses to Trouble Spots

3.2.1 Intra-utterance repairs

The goal of our second study was to examine the children's responses to trouble spots, i.e. intra-utterance self-repairs, while interacting with the computer agent as well as the to investigate whether there was a relation between gestural use and speech repairs. For this purpose, we analyzed 891 utterances from 11 children aged 4-6 (4 four year olds, 3 five year olds and 4 six year olds).

We focused on three types of speech repairs, i.e. retracing without correction, retracing with correction and interruption/trailing off. Example for each speech repair types and corresponding tagging is given below.

1. Retracing without correction: the child begins to say something, stops and then repeats the earlier material without changes.

Child*: One is hiding in a baseball helmet and one's hiding* ***<in> [/] in*** *a box.*

2. Retracing with correction: The child begins to say something, stops and then repeats the phrase with some sort of form or content correction.

Child*: I don't know* ***<what> [//] who*** *use the ladder.*

3. Interruption/trailing off: incomplete utterances.

Child*:* ***And [+/].***

Child*: I don't know.*

In order to get a general sense of the presence of speech repairs in the data, we first calculated the percentage of utterances that contained speech repairs both in the

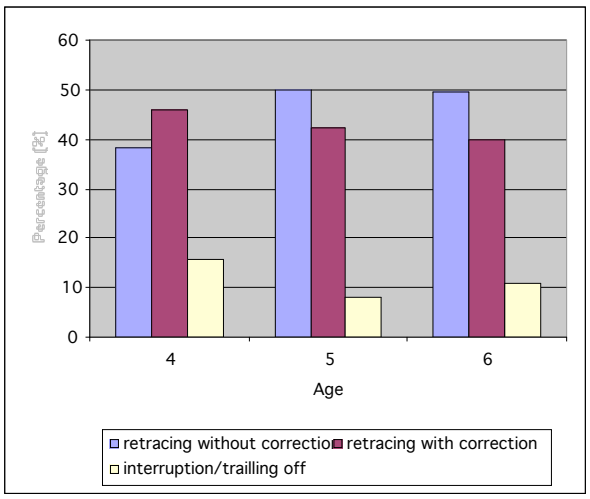

Figure 2. Percentage of speech repair types in total number of repairs employed in verbal only utterances for each age group.

verbal-only and in the multimodal (verbal + gestural) utterances. Of the utterances analyzed, 82.94% were verbal-only utterances, and 17.06% were multimodal utterances. Of the verbal-only utterances, 31.26% contained speech repairs, and of the 152 multimodal utterances, only 25.66% contained speech repairs. Also, the percentage of multimodal constructions in utterances that contain speech repairs was less than the percentage of multimodal constructions in fluent utterances, 14.44% vs. 18.20%. These results seem to indicate that the more the children interact multimodally, the less they need to repair their utterances.

In order to find out which repair strategies the young preschoolers employed, we calculated, for each age group, the percentage of each repair strategy over the total number of repairs employed. Figure 2 and Figure 3 show the distribution of speech repair types in total number of speech repairs employed in verbal only and in multimodal utterances respectively. As can be seen from Figure 2, retracing without correction is the most common speech repair strategy among the 5- and 6-year-olds in verbal-only utterances. However, the favourite repair strategy among the 4 year-olds seems to be retracing with correction.

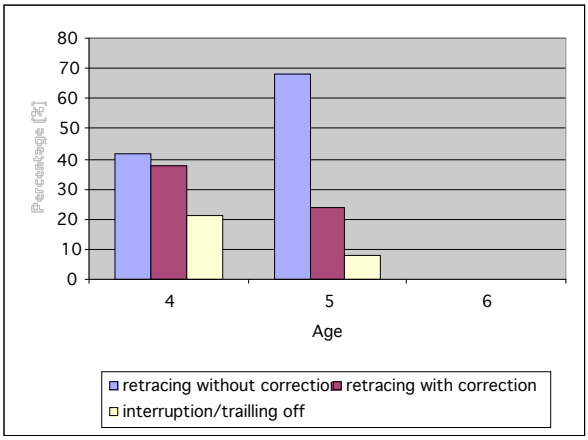

Figure 3. Percentage of speech repair types in total number of repairs employed in multimodal utterances for each age group.

In multimodal utterances, the favorite repair strategy seems to be retracing without correction especially among the 5-year-olds. Since the 6-year-olds in our experimental group preferred exclusively verbal utterances while communicating with the computer agent, it is unclear at this point which repair strategies this age group employed in multimodal constructions. For instance, of the total 318 utterances analyzed for the 6-year-olds, only 2.2% were multimodal, and no speech repairs were encountered in such utterances.

3.2.2 Filled Pauses

Filled pauses like "um" and "uh" are parts of the spontaneous speech and it is known that presence of these events degrades performance of the automatic speech recognition systems. In this section, we analyzed the presence of filled pauses in verbal only and in multimodal utterances. The percentage of utterances that contains filled pauses in total verbal only and in total multimodal utterances for each age group is shown in Figure 4. As can be seen, 4- year-olds tend to produce fewer filled pauses compared with the older ones. Also, analyses revealed that children were less likely to use filled pauses when they are interacting multimodally.

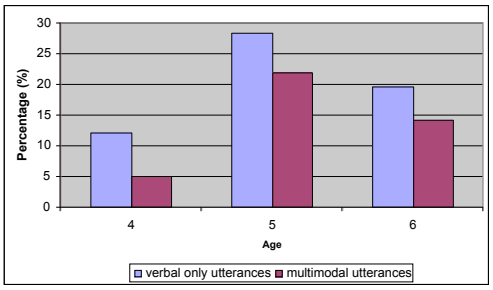

Figure 4. Percentage of utterances that contain filled pauses in total number of verbal only and in total number of multimodal utterances for each age group.

3.3 The Referential and Gestural Analysis

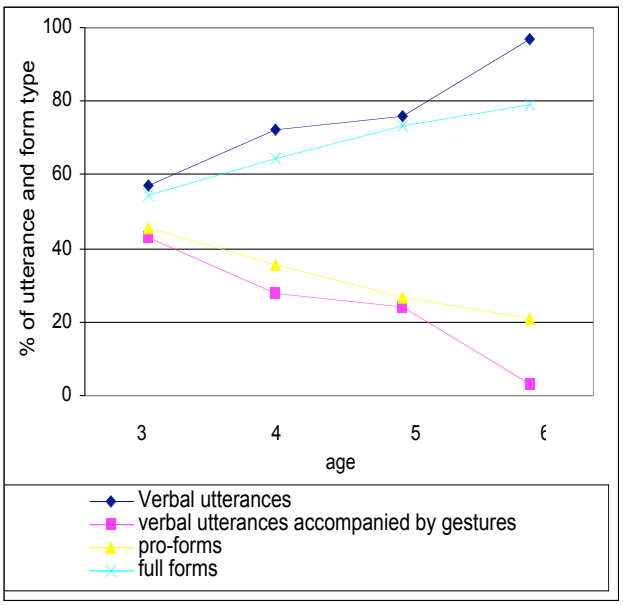

Figure 5. The mean percentage of utterance and form types employed by the 3-, 4-, 5- and 6-year-olds while interacting with the computer agent.

The goal of our third analysis was to examine the children's use of reference marking while interacting with the computer agent as well as to investigate whether there was a relationship between the use of gestures and reference marking. For this purpose, we first tagged all numerals, common, and proper names as *full-forms,* and pronouns, deictic adverbs, deictic numerals, and deictic pronouns as *pro-forms.* We then examined, both in verbal-only and in multimodal utterances, whether the children appropriately employed full forms to introduce new referents while using pro-

forms to maintain reference to previously mentioned entities.

Our analysis reveals that the older children make more extensive use of full-forms while interacting with the computer agent than the 3- and 4- year-olds. For example, they introduce new referents and describe events with full noun or prepositional phrases (*the cat* or *on top of the tree*) rather than with pronouns and deictics (*he/she, here/there, this/that*). This makes their discourse referentially clearer and hence more coherent and cohesive. On the other hand, younger children use pronouns both to introduce and maintain reference to the task entities and they use deictics rather than full prepositional phrases to describe events, producing referentially implicit, and hence unintelligible, stretches of discourse.

In order to examine the relation between reference making and gestural use, we co-analyzed deictic gestures (pointing/touching gestures) and referential devices. Our results indicate that the majority of the younger children describe the task entities or the location of the objects in the pictures with a combination of pro-forms and pointing gestures (*he*/*this one/that* + pointing/ touching gesture to describe or introduce an entity*; here/there* + pointing/touching gesture to refer to the location of an object), tacitly assuming that the computer agent can process both verbal and multimodal utterances. On the other hand, most of the older children make use exclusively of verbal utterances to make their answer explicit. This seems to suggest that while older children are more aware of the nature of the computer task and modify their speech accordingly, younger children fail to take into account the informational needs of the computer agent and produce speech that combines the verbal and gestural modalities. Notice that because their gestures are almost exclusively deictic in nature, gestural use complements the younger children's implicit verbal answers performing a crucial referential function.

The relation between the use of referential devices and gesture is given in Figure 5. As can be seen, an increase in the number of verbal utterances is accompanied by an increase in the percentage of full forms. This means that the more the children become verbal (rather than multimodal), the more they become explicit in their formulations (i.e. they use more full forms). This finding seems to suggest that plain "talking computers" might be more appropriate educational tools for early elementary school children rather than younger pre-literate children who might benefit to a larger extent from interfaces combining both verbal and non-verbal modalities. Further details on our referential and gestural analysis can be found in [10].

4. Conclusion

In this paper, we have reported the results of analysis of interplay between spoken language and gestural cues in conversational child-computer interactions in pre/early literate age groups. This age group has not been well studied in the human language technology community, yet is the age group where the most dramatic changes in language, cognitive and social skills occur. Our results indicated that preschoolers', 3- 4- year-olds, tend to communicate multimodally rather than unimodally. Also we found a direct relation between referential devices and use of deictic gestures. From our speech repairs analysis, the favorite repair strategy among all children seems to be *retracing without correction* when they are interacting multimodally. We believe that integrated analysis of information streams from different communicative cues and also identifying underlying timing and precedence relations between them can be helpful in designing computer interfaces that specifically aimed to younger children.

Acknowledgments: Work supported by NSF and a USC Provost Undergraduate Research Fund.

References

[1] Narayanan S and Potamianos A (2002). Creating Conversational Interfaces for Children, *IEEE Trans. Speech and Audio Processing*, 10(2):65-78.

[2] Oviatt S (2000). Talking to Thimble Jellies: Children's Conversational Speech with Animated Characters, *Proc. ICSLP 2000*, pp. 67-70.

[3] Arunachalam S, Gould D, Andersen E, Byrd D and Narayanan S (2002), Politeness and frustration language in child-machine interactions, in *Proc. Eurospeech*, pp. 2675-2678.

[4] Xiao B, Girand C, and Oviatt S (2002). Multimodal integration patterns in children, In *Proc. ICSLP 2002*, pp. 629-632.

[5] Chen L, Harper M and Quek F (2002), Gesture patterns during speech repairs, in Proc. of ICMI, Pittsburgh, PA.

[6] Andersen, E. S. (1996), A cross-cultural study of children's register knowledge. In D. Slobin, J. Gerhardt, A. Kyratzis and G. Jiansheng (eds.). *Social Interaction, Social Context, and Language*. Hillsdale, N.J.: Erlbaum, 125-142.

[7] Boersma P and Weenink D, Praat Speech Processing Software, Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org

[8] MacWhinney B (2000). The CHILDES Project: Tools for Analyzing Talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[9] Kipp M (2001), Anvil - A Generic Annotation Tool for Multimodal Dialogue, *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370, Aalborg. http://www.dfki.de/~kipp/anvil/

[10] 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, *submitted to* ICSLP 2004.