Fun and Fair: Influencing Turn-taking in a Multi-party Game with a Virtual Agent

Sean Andrist, Iolanda Leite, Jill Lehman
Disney Research, Pittsburgh, USA
{sean.andrist, iolanda.leite, jill.lehman.-nd}@disneyresearch.com

ABSTRACT
Language-based interfaces for children hold great promise in education, therapy, and entertainment. An important subset of these interfaces includes those with a virtual agent that mediates the interaction. When participants are groups of children, the agent will need to exert a certain amount of turn-taking control to ensure that all group members participate and benefit from the experience, but must do so without being so overtly directive as to undermine the children's enjoyment of and engagement in the task. We present a hierarchy of nonverbal and verbal behaviors that a virtual agent can employ flexibly when passing the conversational turn. When used effectively, these behaviors can equalize participation, and potentially decrease the amount of overlapping speech among participants, improving automatic speech recognition in turn. We evaluated the behaviors by having children play a language-based game twice, once with a flexible host and once with an inflexible host that did not have access to the behaviors. Post-game opinion cards revealed no difference between the conditions with respect to fun or likability of the host, despite the flexible agent eliciting more evenly distributed play.

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology, User-centered design*

General Terms
Design, Experimentation, Human Factors

Keywords
Turn-taking, non-verbal behaviors, virtual agent, multi-party interaction, child-agent interaction.

1. INTRODUCTION
In order to design effective language interfaces for multi-party groups, a number of key challenges must be addressed and overcome, including robust automatic speech recognition (ASR), addressee identification (AID), face and gesture tracking, and turn-taking. These challenges become more difficult when the participants are children, who are, in general, less clear in their articulation, less competent in their vocabulary, and less observant of conversational conventions.

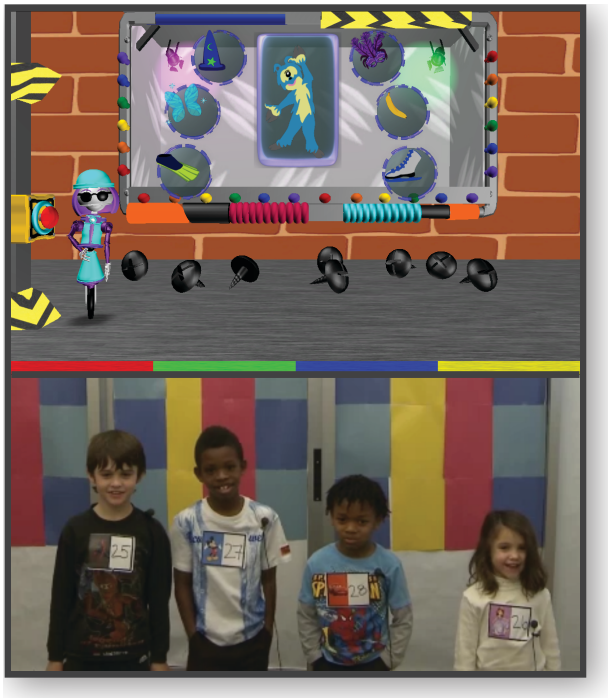

Figure 1. A screenshot of Robo Fashion World, a language-based game with a virtual agent, and four children playing the game, arranged in a line directly in front of the game screen.

An important subset of language interfaces includes those in which action occurs through communication with a virtual animated character, or agent (Figure 1). If such interfaces are to be effective, the agent must be both flexible and adaptable to different groups. Some groups of children will self-organize their turn-taking effectively, while others will be chaotic, turning ASR and AID into nearly impossible tasks. Additionally, less assertive children may be "shut out" of the interaction by more aggressive children, diminishing their overall experience. Unfortunately for the interface designer, children often find chaotic interaction to be, in and of itself, fun. The challenge, then, is how to influence the children to take turns without being so explicitly directive that the interaction loses the sense of excitement and spontaneity that might otherwise keep the children interested and engaged.

Interaction Design and Children'13, June 24–27, 2013, New York City, New York, United States.

In natural conversation, people use both verbal and nonverbal strategies to facilitate turn-taking and non-overlapping speech. The goal of this work is to show how a virtual agent can make use of similar strategies, leveraging its own ability to use speech, gesture, and proxemics to affect the group interaction. In particular, we want the character to respond flexibly to different kinds of groups in order to (1) better equalize participation, and (2) reduce the amount of overlapping speech in the interaction, and to do so without eliminating the fun.

2. RELATED WORK

Previous research has studied the potential of language-based interfaces for children in different ways. Children exhibit fewer disfluencies when speaking to an animated character versus speaking to an adult or another child, but significantly more disfluencies than an adult speaking to an animated character [1], [8]. The inherent variability in children's speech makes ASR difficult, but the creation of novel language models for children can make the problem tractable [7]. Similarly, AID is made difficult by children's frequent violation of conversational conventions. Recent work addresses this problem with sophisticated machine learning models that integrate audio and visual features with temporal group interactions [6].

To be truly effective, language-based interfaces will need to support natural conversational turn-taking. In human-human interaction, turn-taking has been shown to exhibit regular structure that ensures one-at-a-time speaking and minimal gap between turns [11]. Both verbal and nonverbal mechanisms support turn-taking, including the use of interrogatives, buffers, gaze, gesture, and proxemics [4], [12]. This structure often breaks down in conversations among young children, especially in groups of three or more [3].

Turn-taking in conversation between humans and agents has been predominantly investigated in dyadic contexts. Chao et al. investigated the timing of turn-taking in human-robot interaction, and showed how a robot can use gaze, speech, and motion to facilitate turn-taking when playing a game with a single human adult [2]. Ryokai et al. developed a virtual character that takes turns with a child telling stories in order to facilitate literacy learning [10]. Less work has investigated how virtual agents can engage with groups of children, leveraging their own embodiments to make use of verbal and nonverbal cues and positively influence the interaction. Our work seeks to address this knowledge gap by identifying a hierarchy of behaviors that a virtual agent can use in conversation with groups of children.

3. METHODOLOGY

In this section we present behaviors that could allow a virtual agent to flexibly achieve the goals of equalized participation and less overlapping speech among children. We also describe the scenario of interaction in which we contextualized our work.

3.1 Flexible Behaviors

Effective turn-taking is critical to language-based interaction, especially in groups. Through implicit and explicit signals each participant takes or cedes control of the floor so that information flows to its intended recipient. When the intended recipient is the virtual agent, it must signal that it has the floor, act in a way that reflects the intent of the utterance, and then release the floor in a way that is understood by those who may wish to take it next.

Our agent uses four behavioral strategies when yielding the floor. These strategies vary in their directness, and can be employed flexibly and deliberately to ensure equal participation. By using the more direct behaviors, the agent might also influence more chaotic groups to exhibit less overlapping speech. The four behaviors are summarized below and visualized in Figure 2.

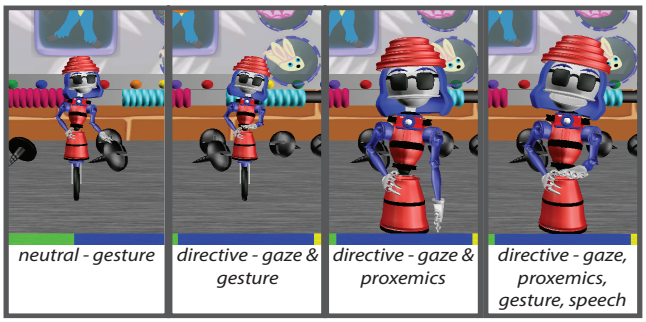

Figure 2. Charlotte demonstrating the four flexible behaviors, in order of increasing directness from left to right. In the neutral behavior, she passes the turn to the entire group. In the other three, she is passing the turn directly to the child standing on the blue line.

Neutral – gesture: The agent yields the conversational floor to the entire group, allowing any child to potentially take the next turn. In this strategy the agent stays back, maintaining distance from all children, and passes the turn with a sweeping gesture. This is the least directive behavior, and is employed when group participation is mostly equal.

Directive – gaze & gesture: The agent yields the floor to a single child for the next turn. In this strategy the agent stays back, but positions itself with gaze (head and body orientation) toward the single child it is attempting to pass the turn to. The agent also uses a pointing gesture. This is a mildly directive behavior, and is employed when group participation is becoming slightly unequal.

Directive – gaze & proxemics: The virtual agent yields the floor to a specific child. Here the agent not only gazes at the intended turn recipient, but also makes use of proxemics—positioning itself closer to the user in the virtual scene by becoming larger in screen space, and giving the impression that it is moving closer to a single participant. This is a directive behavior, and is employed when group participation is moderately unequal.

Directive – gaze, proxemics, gesture & speech: In the final behavior the agent once again yields the conversational floor to a single child. This strategy makes use of gaze and proxemics in the same way as the previous behavior, but adds a gesture and an explicit verbal interrogative (e.g., "Would you like to go next?"). This is the most directive behavior, and should be employed when user participation is very unequal, i.e., when the child in question is participating far less than others in the group.

3.2 Scenario of Interaction

We have contextualized this work in an interactive language-based game called Robo Fashion World. In this game, groups of up to four children change the appearance of a model by calling out the names of items on the game board (see Figure 1). The host of the game, Edith, is an animated character responsible for making the changes and mediating the interaction. At the beginning of the game, Edith explains the two main game actions: children can either ask Edith to dress up the model with one of the items currently displayed on the board, or ask her to take a picture of the model as it is. At the end of the game, each child selects one of these pictures to take home. Children do not receive any further instructions on how to play the game. Because they use

language to refer to the items on the board, requests can be unclear or occur simultaneously with the requests of others. The game ends after 20 board changes.

During the game, Edith is controlled in a Wizard-of-Oz manner. The wizard acts primarily as the speech recognizer, indicating to Edith what was said (e.g., a fashion item was selected or a picture was requested) and who said it. Edith then autonomously selects a sequence of behavioral actions to perform, e.g., to acknowledge a child's choice, push the button that dresses up the character with the selected clothing, and then pass the turn back to the group.

4. EVALUATION

To test the effectiveness of the flexible behaviors presented in the previous section, we designed a within-group study in which children played Robo Fashion World twice, once with a character that could make use of all four flexible behaviors, and once with an inflexible character that would use only the neutral behavior. In order to create these two conditions, we designed a second character to host the game—Edith's "sister," Charlotte (Figure 2). The assignment of behaviors to host was counter-balanced along with order so that each sister hosted first or second using flexible or inflexible behaviors about the same number of times.

4.1 Participants

Thirty-three children (17 females and 16 males) were recruited through postings in physical and online community bulletin boards and compensated for their participation. Ages ranged from 4 to 10 (M = 7 years, SD = 2 years). The children participated in 10 groups, with 2 to 4 children in each group. Within-group age ranges varied from two to five years, as may occur in families or groups of strangers at public events. Most groups (8/10) contained at least two children who were siblings or otherwise knew each other, and most groups (7/10) also contained children who did not know each other prior to participating in the study.

4.2 Setup

The game was presented on a 52-inch LCD TV screen about five feet away from the line of children. Each child was positioned in front of a colored rectangle on the floor that corresponded with a colored rectangle on the screen (Figure 1). The agent would position herself behind one of these colored rectangles when she wished to indicate that she was addressing a specific participant. The wizard sat at a computer off to the side, as did the children's parents if they wished to be in the room. The experimenter stood behind the children and out of view.

Data capture for this study included one front- and two side-view HD cameras as well as individual close-talk microphones. Log files of each game were recorded automatically, and indicated the timing of participant choices, which choices were made, and the behaviors the agent elected to use during the game.

4.3 Procedure

Following informed consent, the children were brought into the study room and briefly introduced to the two agents they would be interacting with, Edith and Charlotte. They then played the game once with one of the two agents in either the flexible or inflexible condition. Following the game, the participants were asked to answer some short questions about their experience. Next the participants played the game a second time with the other host and condition. Afterward, they answered the same set of subjective questions about the second game, followed by questions comparing the two hosts. Next, the participants were asked to repeat a series of words and phrases said by the experimenter. This activity served the dual purpose of obtaining audio data of the children's voices for ASR work, as well as being a distractor task so that children could better reflect on their impressions of both games, elicited in an open-ended interview at the end of the study. In this final interview, children were asked to say what they liked and did not like about each host and each game, and were encouraged to elaborate on these responses if they could.

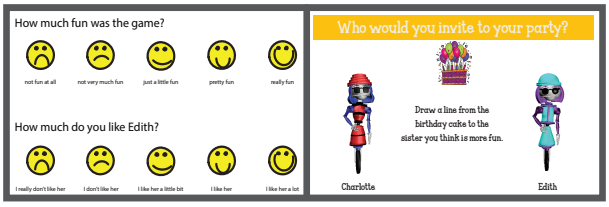

Figure 3. Smiley-o-meter questions and sample comparison card.

4.4 Measures

The study's independent variable was the turn-taking condition of the host: flexible or inflexible. The primary dependent variables included the number of turns taken by each participant, as well as the amount of overlapping speech. Overlapping speech was defined as the percentage of utterances directed toward the agent by one speaker that co-occurred in whole or in part with an utterance by another speaker.

We also collected a number of subjective measures of enjoyment and likability as dependent variables. Participants rated each host and game individually using Smiley-o-meter scales [9] and made forced-choice comparisons between the hosts (see Figure 3). Comparisons took the form of questions written on cards asking the participants to draw a line from a representative item to one of the hosts. For example, participants drew a line from a birthday cake to either Edith or Charlotte on the card that asked, "Which host would you invite to your birthday party?" Other comparison cards asked which sister was smarter, which would be preferred as a teacher, which was prettier, and which should be "hired" as the final host for Robo Fashion World. The text on all cards was read aloud by the experimenter for all groups.

We also modeled the host used in each condition as a covariate in order to determine if either Edith or Charlotte's appearance had an inherent effect on our measures, regardless of behavioral condition. Finally, we looked at the gender and age of participants as covariates, paying particularly close attention to any potential interaction effects between host and condition, gender and host, and gender and condition.

5. RESULTS

In order to measure the effect of the agent's use of flexible behaviors on participation, we analyzed the statistical variance of turns taken within each group using a repeated measures analysis of variance (ANOVA). Average turn variance was 0.61 when the agent used flexible turn-passing behaviors, significantly lower than the turn variance of 6.35 in the inflexible agent condition, $F(1,16) = 8.57$, $p = .010$.

A repeated measures ANOVA indicated that the behavioral condition of the agent had no significant effect on the amount of overlapping speech within groups, $F(1,16) = 0.67$, $p = .43$. We note, however, that the groups in this study tended to have less rambunctious behavior overall than others who have played the game previously [6]. Our procedure did not include a warm-up

task, and 6/10 groups exhibited more overlapping speech in the second game suggesting they may have needed the first game to get used to one another and the environment. In the four groups where there was less overlapping speech in the second game, three were hosted by the flexible version of the character. We consider this to indicate a tentatively positive result.

Turning to our subjective measures, we analyzed the ordinal Smiley-o-meter scales using Likelihood-ratio chi-squared tests, and found that the use of flexible behaviors had no significant effect on either the amount of fun children had playing the game, $\chi^2(1) = 0.99$, $p = .32$, or on the amount that the children liked the agent, $\chi^2(1) = 0.58$, $p = .45$. Children made use of the full scale for each measure. We also found through the use of Pearson's chi-squared tests that the children's choices of agent on the comparison cards were not significantly affected by behavioral condition, including which agent was chosen as smarter, prettier, or the better teacher, or which should be invited to a birthday party or "hired" as the final host of the game.

When including the gender of the children in the analysis, subjective evaluations of liking (but not enjoyment) were more positive for Charlotte than for Edith among boys, irrespective of behavior manipulation, $F(1,59) = 10.36$, $p = .002$. This difference might be explained by Charlotte's red color; recent research has shown adult males to be more attracted to females in red clothing [5]. The gender interaction slightly weakens our claim that the behavior manipulation had no effect on which agent was more liked, as any potential difference may have been washed out by an inherent preference for Charlotte among boys. In future work we will use more neutral colors for our agents.

5.1 Discussion

Overall, our evaluation indicated strong support for the claim that the use of flexible behaviors can equalize participation within small groups of children. Children who were "shut out" of the game by more assertive participants in the inflexible condition were able to take more turns in the flexible condition.

Our evaluation indicated tentative support for the claim that the use of flexible behaviors can decrease overlapping speech in small groups of children. This conclusion could only be reached when accounting for the fact that children were more "warmed up" in the second game they played, and thus more comfortable and excited. We intend to collect more data to strengthen this result, and will include a warm-up session in the experimental design.

Finally, our results indicated that the agent's use of flexible behaviors had no impact on the amount of fun children had while playing the game. In fact, some children picked up on and appreciated the flexible agent's efforts to add turn-taking structure to the interaction:

> P4: *She [flexible host] called on people independently. She would tell us what to do.*

This feeling was not unanimous, of course, as some children seemed to prefer the more free-for-all type of interaction rather than speaking and participating one at a time:

> P11: *She [inflexible host] let us say it as a group.*

Although there was some variability in preferences based on the children's personalities, it was encouraging to see that generally the children had fun with the game and liked the agent, even when it was being slightly more direct in regulating the interaction.

6. CONCLUSION

A virtual agent's verbal and nonverbal behaviors need to be carefully designed to support its context of interaction. We have shown one possible specification for an agent's behaviors that facilitates turn-taking among small groups of children when playing a language-based game. By using flexible behaviors, the agent was able to influence how the conversation unfolded, equalizing participation and decreasing overlapping speech without negatively affecting the children's enjoyment of the game. Ultimately, we believe that our findings will be critical to informing the design of future language-based interfaces involving a virtual character for children, resulting in systems that are sensitive to the limitations of speech technology but still effective and fun for everyone involved.

7. REFERENCES

[1] Black, M., Chang, J., Chang, J., & Naryanan, S. (2009). Comparison of child-human and child-computer interactions based on manual annotations. *Proceedings of the Workshop on Child, Computer and Interaction*. Cambridge, MA.

[2] Chao, C., Lee, J., Begum, M., & Thomaz, A. L. (2011). Simon plays Simon says: The timing of turn-taking in an imitation game. *RO-MAN, 2011 IEEE*. pp. 235-240.

[3] Ervin-Tripp, S. (1979). Children's verbal turn-taking. *Developmental pragmatics*, 391-414.

[4] Goodwin, C. (2007). Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry*, *50*(3-4), 272-302.

[5] Guéguen, N., & Jacob, C. (2012). Clothing Color and Tipping: Gentlemen Patrons Give More Tips to Waitresses With Red Clothes. *Journal of Hospitality & Tourism Research*.

[6] Hajishirzi, H., Lehman, J., & Hodgins, J. (2012). Using Group History to Identify Character-directed Utterances in Multi-child Interactions. *Proceedings of 13th SIGdial conference on Discourse and Dialogue (SIGDIAL'12)*.

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

[8] Oviatt, S. (2000, October). Talking to thimble jellies: Children's conversational speech with animated characters. In *Proc. ICSLP*. Beijing, China. pp. 67-70.

[9] Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. *Proceedings of the 2006 conference on Interaction design and children*, 81-88. ACM.

[10] Ryokai, K., Vaucelle, C., & Cassell, J. (2003). Virtual peers as partners in storytelling and literacy learning. *Journal of computer assisted learning*, *19*(2), 195-208.

[11] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 696-735.

[12] Wiemann, J. M., & Knapp, M. L. (1975). Turn-taking in conversations. *Journal of Communication*, *25*(2), 75-92