Toward Better Understanding of Engagement in Multiparty Spoken Interaction with Children

Samer Al Moubayed, Jill Fain Lehman

Disney Research
Pittsburgh, PA, USA
{samer, jill.lehman}@disneyresearch.com

ABSTRACT

A system's ability to understand and model a human's engagement during an interactive task is important for both adapting its behavior to the moment and achieving a coherent interaction over time. Standard practice for creating such a capability requires uncovering and modeling the multimodal cues that predict engagement in a given task environment. The first step in this methodology is to have human coders produce "gold standard" judgments of sample behavior. In this paper we report results from applying this first step to the complex and varied behavior of children playing a fast-paced, speech-controlled, side-scrolling game called *Mole Madness*. We introduce a concrete metric for engagement—willingness to continue the interaction—that leads to better inter-coder judgments for children playing in pairs, explore how coders perceive the relative contribution of audio and visual cues, and describe engagement trends and patterns in our population. We also examine how the measures change when the same children play *Mole Madness* with a robot instead of a peer. We conclude by discussing the implications of the differences within and across play conditions for the automatic estimation of engagement and the extension of our autonomous robot player into a "buddy" that can individualize interaction for each player and game.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – Natural Language; D.2.2 [**Software Engineering**] Design Tools and Techniques – State diagrams; D.2.11 [**Software Engineering**] Software Architectures – Languages

General Terms

Design, Human Factors.

Keywords

Multi-party interaction, child-computer interaction, group task engagement, dialogue systems, spoken interaction, child-robot interaction.

ICMI '15, November 09 - 13, 2015, Seattle, WA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2820733

1. INTRODUCTION

Researchers in interaction all begin from a fundamental assumption: to interact one must first engage and to continue to interact one must stay engaged. Working from this assumption we tend to instantiate a methodology that, in the abstract, follows the same plan: gather interaction data, post-process and/or code the data with respect to engagement to create a "gold standard," find the observable or "sense-able" features in the environment that best predict the engagement states we want to distinguish, and use models based on those features to modify the agent's behavior in the interaction. What the commonality of this description tends to ignore is that what each of us means by engagement is a function of how we want to use the feature in both sensing and acting. For some the term is closely related to "attention" or "joint attention," is measured with respect to observable phenomena over short durations, and effects only immediate actions in the agent. This is a rich viewpoint; many researchers who work on conversation analysis and dialog systems have contributed to our understanding of this kind of engagement and identified observable cues such as facial expressions (Sidner et al. 2005), gaze (Nakano & Ishii, 2010), and gesture and posture (Sanghvi et al. 2011) that can be used by an agent to make effective changes in its behavior. Bohus & Horvitz (2009) focus in further on a single critical instance of this kind of engagement—the signals surrounding the moment when users in open spaces show the intention to interact.

The short-duration view of engagement is not unique to researchers in dialog interaction. Leite, for example, uses engagement as a binary feature to code the behavior of groups of children watching and listening to robots acting out a social scenario (Leite et al. 2015). Her goal in the work is to give the robots the ability to sense when children are disengaged. So, although she derives the feature for every 500 msec time slice, she does not consider engagement as something different from disengagement's opposite. That is precisely the change in viewpoint we present in this work—engagement as a continuous, multi-valued phenomena that is expected to wax and wane over an interaction. We see children as more or less engaged in our activities rather than engaged or disengaged, and expect that there are patterns of engagement over time that

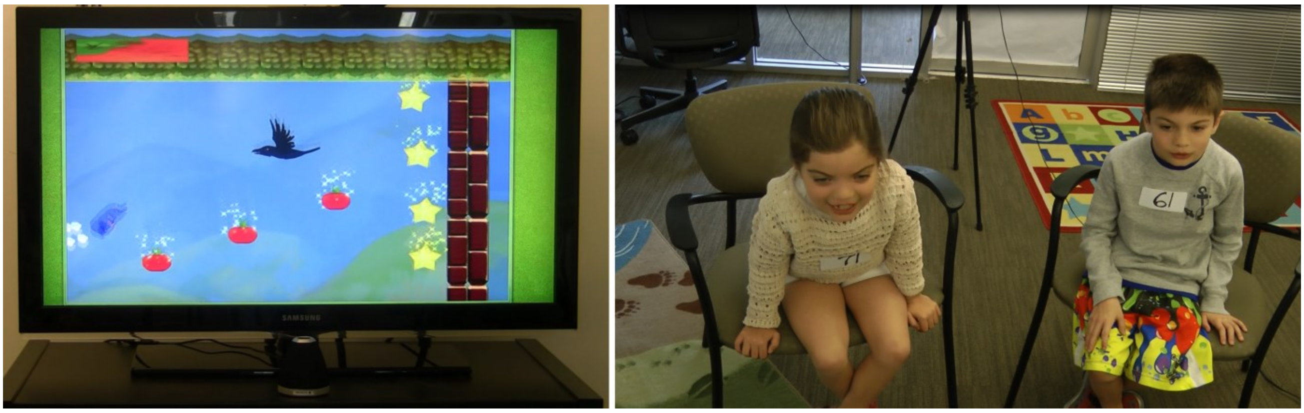

Figure 1. A snapshot of interaction with *Mole Madness*. Left: the game as it appears on a 40" flat screen. Right: Frontal view of two children playing the game.

correspond to successful interactions and patterns that do not. Such a view is clearly related to engagement in the short-term; we also expect that patterns of engagement in the large can be built by sensing different kinds of instances of engagement in the small.

The work reported here is a description of the first two steps of the canonical process toward our goal: gathering interaction data and coding it with respect to patterns of engagement. We describe our interactive game, *Mole Madness*, and the procedure we used to collect data from pairs of children playing together and one-on-one with a non-adaptive robot. We then turn to the coding phase, where we found that the large-scale view of engagement we want to model has unexpected implications and so discuss our eventual decisions and rationales in detail. Having settled on a coding, we examine the patterns of engagement it reveals, primarily with respect to the child-child data we intend to use to build our models, but briefly with respect to the child-robot interactions as well. We conclude with a discussion of the variability in the children's behavior and the challenges it creates for building a robot "buddy" that can adaptively provide a positive experience for each child.

2. *MOLE MADNESS*: A SPEECH-CONTROLLED GAME

Mole Madness is a two-dimensional side-scroller similar to video games like Super Mario Bros®. Each of two players controls an aspect of the mole's movement through its environment using a simple verbal command: *go* for horizontal and *jump* for vertical. Without speech, the mole simply falls to the ground and spins in place.

A close-up of the mole's world, and two children playing the game, can be seen in Figure 1. The environment contains typical kinds of objects for this style of game: walls arranged as barriers to go over or between, items that result in point gain (cabbages, carrots) and point loss (cactuses, birds, rocks), and the occasional special object (star) that acts as a boost to change the mole's normal physics. In addition to providing a familiar and fun experience for the players, the environment is designed to elicit specific patterns of speech. There are flat stretches to evoke isolated consecutive *gos*, steep walls to produce isolated consecutive *jumps*, and crevasses to get through and items to avoid that require coordinated and overlapping sequences of commands. A score bar on the screen updates as the mole touches the various kinds of objects. Although players are not given any specific goal other than to move the mole through the level, players typically adopt maximizing speed and/or points as a goal.

When children play together, task commands (*go*, *jump*) naturally occur in a broader conversational context that includes both non-task utterances directed to the mole ("watch out," "faster") and utterances directed to the other player ("wait, don't say jump yet," "look a star," "he's funny"). For the most part this side-talk is not conversational, in the sense that it rarely requires a verbal response (Lehman & Al Moubayed, 2015). Thus, in robot-child games, the robot periodically generates but does not respond to these kinds of utterances.

With respect to nonverbal behavior, the fast pace and visual processing demands of the game tend to reduce some kinds of expressiveness. Eye and head movements typically seen in face-to-face conversation (such as looking at the person being addressed, looking away to hold the floor, etc. (Abele, 1986)) are impractical when visual attention must remain on the screen. Similarly, facial expressions and body movements that might be interpreted as indicating interest, engagement, and excitement, can be absent in some children who become quite still with intense focus.

From this perspective, and in tasks similar to *Mole Madness*, the perception of small-scale engagement and the sense-able features that signal it might be significantly different from what can be learned from human-human conversation and require an analysis of data that is specific to this type of interaction context.

Figure 2. Snapshots of 6 different video file segments (simultaneous pairs from each of three different games).

3. DATA COLLECTION

In a multi-study data collection, children took part in four activities over the course of a one-hour period, spending approximately ten minutes per interaction with five-minute breaks. *Mole Madness* comprised two of the activities: children played once in pairs, and again, individually with a robot as co-player. All children's families were compensated for their time. Participation took place on four consecutive weekends, at the families' convenience during summer vacation.

Population. Twenty-eight children, ages 5 to 10 (50% female), played in pairs and one-on-one with Sammy J (a back-projected robot head developed by Furhat Robotics, Al Moubayed et al. 2012). Due to hardware failure, the results discussed in this paper are for the 26 children for whom we have complete data sets (child1-child2, child1-Sammy, and child2-Sammy): 12 females and 14 males with a mean age of 8.3 years (SD = 1.2 years).

Children who are unfamiliar with each other or are of substantively different ages can have very different play styles and patterns of engagement than pairs who are familiar and/or developmentally close. Because we are interested in understanding the dynamics of children playing together and how that can guide behavior when each child plays with a robot who should act like a "buddy," we tried to eliminate these most obvious sources of variability across pairs. Thus as a recruiting strategy, the initial family contacted in each pair was asked to bring both their own child and a friend or sibling who was close in age. All thirteen player pairs were either friends or siblings, with a mean age difference of 7.5 months. Eleven of the thirteen pairs were single sex.

Procedure for child-child games. Children were seated in front of, and in an equilateral triangle with respect to, a 40" flat screen where the game was displayed. The basic principles of the game were explained by a confederate. They were told to interact with the game using their voices, and that the mole could be controlled by the words *go* and *jump*. Each child was assigned an initial role arbitrarily, and at a point in the middle of the game, the pair was asked to switch roles/use the other word. They were also told to use the word *next* when the mole reached the target flag and disappeared into his hole at the end of a level in order to bring the mole back out on a new level. The children were not told to avoid speaking to each other or addressing the mole using other words.

Children were unaware that a trained wizard, located in another room, was using a game controller to remotely control the mole's movements in response to their *go*s, *jump*s, and *next*s. A wizard was used in order to avoid any discrepancies in the quality of the game or the interaction that would stem from problems in recognizing overlapping speech. With only auditory access to the room where the game was played, the wizard had no ability to anticipate the children's speech, and no feedback on the results of his actions or the status of the game play.

The data collection took place in a child-friendly room with minimal hardware intrusion (Figures 1 and 2). Interactions were audio-visually recorded using two high definition cameras, and two stereo microphones. One camera and a stereo microphone were placed on top of the screen capturing a frontal view of the game players (Figure 1-right, Figure 2), and the other capturing a back view with the screen and the game (Figure 1-left). All events from the wizard and from the game were automatically time-logged along with the audiovisual recordings.

Procedure for child-Sammy games. The robot was placed in the right child's position of the same triangular arrangement, and the procedure was identical to the child-child case with two exceptions. First, no wizard was needed because Sammy is able to play autonomously using output from the game and a small microcone array as its only sensory input (details of the dialog design and structure of the system architecture can be found in (Al Moubayed & Lehman, 2015). The second difference was that Sammy always played in the *go* role throughout the activity.

4. DATA PREPARATION

Data recordings comprised fourteen different child-child games, and 28 child-robot interactions. On average, child pairs played for 354 seconds (SD = 65 seconds). Child-Sammy games tended to be shorter (mean=270 seconds, SD=49 seconds).

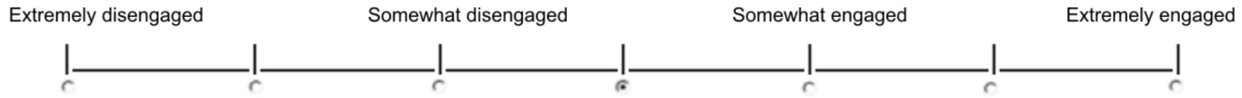

Figure 3. A 7-point scale with four degrees of engagement labeled and separated by non-labeled points.

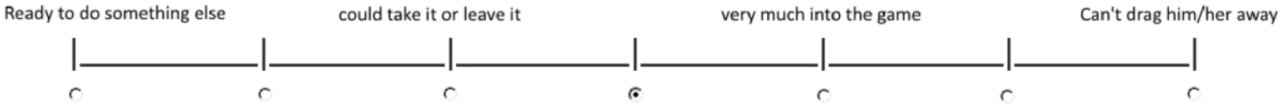

Figure 4. A proxy engagement scale coders used to indicate the "willingness to continue the interaction" based on the segment.

All data streams (audio, video, and logs) were manually synchronized. The recordings were then segmented into game levels, level transitions, and "off-task," the latter defined as any talk or interaction with the confederate. Such events occurred, for example, when a level ended and the children had to be reminded to use the word *next* to continue, or during play when a child turned to the confederate to ask questions or request help. Because we are interested in understanding and modeling engagement in the game, level transitions and off-task segments were removed from further analysis.

5. THE PERCEPTION OF ENGAGEMENT

The work reported here has two primary goals: (1) to characterize the patterns of engagement we see in children playing together and (2) to provide the data from which we will try to determine detectable multimodal features that predict different levels of engagement. To those ends we use human coders to produce judgments and labeling of our audio-video recording.

To prepare the data for coding, full recordings of game play were split into audio-video and video-only ten second segments showing either the child on the left or the child on the right. Separation of left and right players was done so that judgments would be based on the individual child's behavior without the co-player's behavior for comparison. Figure 2 shows examples from three games. Video-only segments were created both to understand the relative contribution of audio cues to the perception of engagement and because audio from the unseen player could not be erased from the audio-video versions.

Child-child sessions produced 822 segments under each of the audio-video and video-only conditions. Coders labeled all video-only segments in random order first, followed by all randomly ordered audio-visual segments. Video-only segments were coded first to isolate any bias in comparing silent segments (due to the intentional removal of the audio track) to files where the verbal contribution of the player was minimal. Coders used a tailor-made annotation tool that allowed efficient progression through the segments and easy use of the engagement scale. Three female coders who have experience with children received training on the tool and were instructed to take a break every 15 minutes to avoid decreased attention over time.

Following standard practice, our coders were initially given a small calibration subset of the data (a 99 segment mixture of video-only and audio-video data randomly distributed across sessions) and asked to rate engagement using a 7-point scale that ranged from "extremely disengaged" to "extremely engaged" with four specified labels and three intermediate points (see Figure 3). The unlabeled values in the scale allow coders flexibility in judging the behavior as lying between the meanings of the labeled points. Annotation values for the calibration set resulted in an inter-rater agreement (Krippendorff's *alpha*) value of a 0.39. Repeated experiments using the same files (with a gap of 24 hours), also showed low test-retest reliability for each annotator.

Low inter-rater and within-rater agreement indicates that whatever observable features our coders were focusing on, they weren't focusing on the same features, weren't focusing on them consistently and/or weren't mapping them consistently to scalar values. To counter this problem we needed a proxy for engagement that they could use more consistently. The kind of engagement we are interested in—something that isn't binary but has an intensity that changes over time—is a subjective experience that we want to render predictable through observable features. Our coders may experience this kind of engagement themselves, but they do not regularly judge it, as "engagement" per se, in others. What they do regularly anticipate and judge is the willingness of a child to be moved between activities. Thus we created the scale shown in Figure 4 and asked coders to re-evaluate each segment based on how willing they thought the child they had watched in the segment would be to continue with the current activity or move to another.

There are three points to be made about the proxy scale. First, although this measure might differ in quality from what engagement is (as an internal subjective experience), it seeks to measure a consequence of engagement in this kind of interaction. Second, it asks for a situated judgment that our coders have experience making with the population and which is relevant in its own right to the issue of creating an enjoyable experience through an adaptable co-player. Third, because it is situated and it does take advantage of our coders' backgrounds, it is not an all-purpose proxy for modeling long-term, variable-intensity engagement. We believe that attempts to model engagement-in-the-large in some other types of task might also find low agreement among coders who use a scale that relies on the word "engagement"; but we also believe that a proxy solution will work better only in so far as the new scale makes sense for the interaction, population and coders involved.

To compare the power of the "willingness" scale to the original, the same 99 files were presented to the three coders using the same tool. The newly coded data achieved an inter-rater agreement (Krippendorff's *alpha*) of 0.59. The increase in overall agreement was coupled with an increase in similarity between coders' distributions over the scale. The Minkowski distances between distributions for each pair of coders were 0.25, 0.29 and 0.49 with the "engagement" scale, but 0.11, 0.17 and 0.06 with "willingness."

Given these improvements, the entire data set was then coded by all three coders using the "willingness" scale, video-only first, as previously described. Krippendorf's *alpha* for the complete data set was 0.58, comparable to the training set. While this value indicates better agreement than we would expect to get from the "engagement" scale, and compares favorably with inter-rater agreement for engagement measures reported by Oertel (2010) (k = 0.56, 10-point scale, 30 raters) and Leite (k = 0.41, binary scale, two raters), it is not considered statistically strong. Some of the variability between coders no doubt comes from real differences in how they view the behavior in relation to the labeled portions of the scale. But a closer analysis shows that the reliability statistic may be reflecting primarily differences in degree rather than kind. The unlabeled intermediate points on both scales are implicitly defined as "about halfway between the two labeled values." For the "willingness" scale 90% of all segments had pair-wise coder values that were within one point of each. In other words, for the majority of segments coders were disagreeing about the degree to which the same category applied. For the "engagement" scale this was true of only 26% of the subsample.

6. RESULTS AND ANALYSIS

Despite the overall improvement in reliability that came from using a proxy for engagement, a conservative view would argue that there remains some degree of consistent difference across coders. In analyzing the results we could eliminate those differences by using a mean or median value for each segment, essentially creating an average coder. Elsewhere it has been argued that the idea of a single correct truth for human annotations in semantic interpretation tasks obscures our ability to get at the nature of the inherent subjectivity in such judgments and the range of reasonable interpretations for complex phenomena (Aroyo & Welty, 2015). With respect to our data, we note that by better situating the scale in the coders' own experiences we may have invited some systematic individual differences in the judgments if each coder based her decisions on the observable features that she has found valuable in redirecting the children she knows. We do not know at this stage how well we will be able to predict a coder's data with the features we are able to sense, nor how much overlap there will be in the features that best predict in each case. It is possible that the differences in the data produce no differences in the sense-able feature sets at all, in which case an average coder model would make sense. But it is also possible that one coder's decisions can be more accurately modeled given current sensors; if that's true we may choose to model a particular way of measuring long-term engagement even though it gives rise to some judgments with which the average coder would not agree. Given this possibility, we choose to preserve the individual differences at this point in the process. To continue here, then, we discuss general results that hold across coders but choose a representative coder, Coder 2, to present the children's patterns from a single coherent view.

6.1 Audio-video versus video-only trends

Recall that each coder judged every segment both with and without audio. Every coder gave an average score across all segments that was higher in the case of audio-visual cues. In particular, Coder 2's values remained unchanged for 57% of the child-child segments, increased for 29% of segments when sound was present, and decreased in 14% of cases. The absolute percent of changed versus unchanged values differed for the other two coders, but the relative split of twice as many increasing as decreasing when there was change, held. Almost all children (at least 25/28 for each coder) received some bump in their mean engagement score from audio, although some children clearly benefited more than others. Further, how much of a bump an individual child received depended on coder, reinforcing the idea that individual coders cared differentially about the presence or absence of specific audio and visual cues.

6.2 Engagement trends

Considered in pairs, we found no correlation between the average engagement level of one child and the average engagement level of the second for any coder. In turning to a more finely grained view, Figure 5 shows the engagement plots for both children in each of the child-child games, using Coder 2's audio-visual values. The plots also show a linear fit for the data of each child.

The graphs show large variability in levels of engagement and in engagement patterns, both within and across games, despite the fact that all pairs of children were familiar with each other, were of similar age, typically the same gender, and played the same game. Four patterns occur to different degrees. The top row of Figure 5 groups players who maintain similar engagement values over time, with both enjoying an increasing trend over the course of the game. The middle two rows show players that also maintain a similar pattern of engagement, but because their values hold steady. The bottom row shows the last two patterns, defined by players who start and/or end at different engagement levels. The two pairs at the left have one child converge toward the other as the game unfolds, while the two pairs on the right show players that diverge over the course of the game. If the values do reflect something meaningful about the children's internal state, then clearly the emotional arcs experienced by our participants are not all positive. Equally clearly, some interactions were more successful than others, at least in the

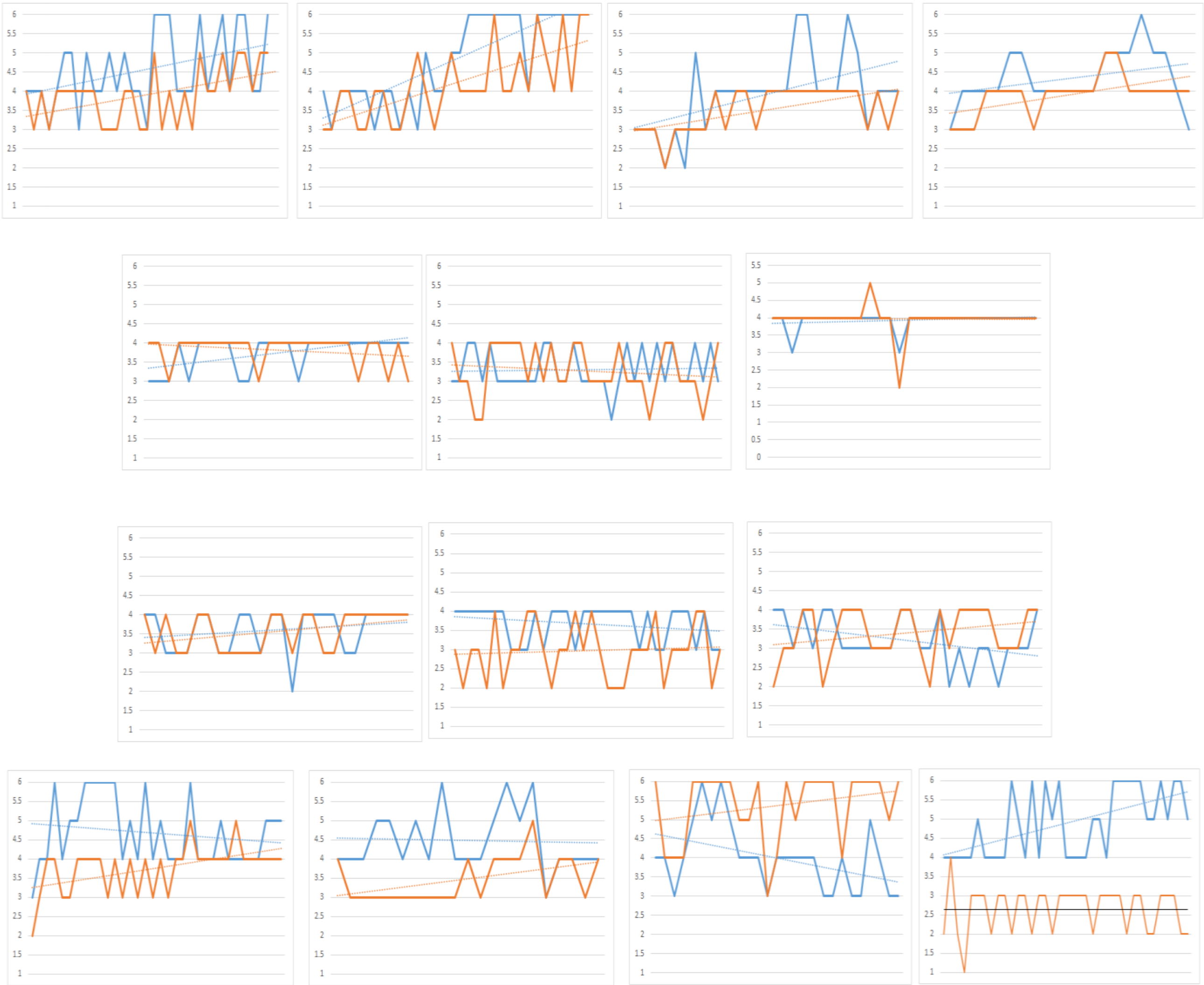

Figure 5. Plots of the engagement trends (engagement over time) for each of the 14 different *Mole Madness* games. Each plot shows the audio-visual data for Coder 2 and a linear fit for a pair of children that played the game together. The plots are arranged in order of similarity and convergence showing large variability across children and games. Two games show divergence in the engagement of the children (the last two plots), while the remainder show a pair with steady or increasing engagement, or one child converging toward the child with the higher engagement.

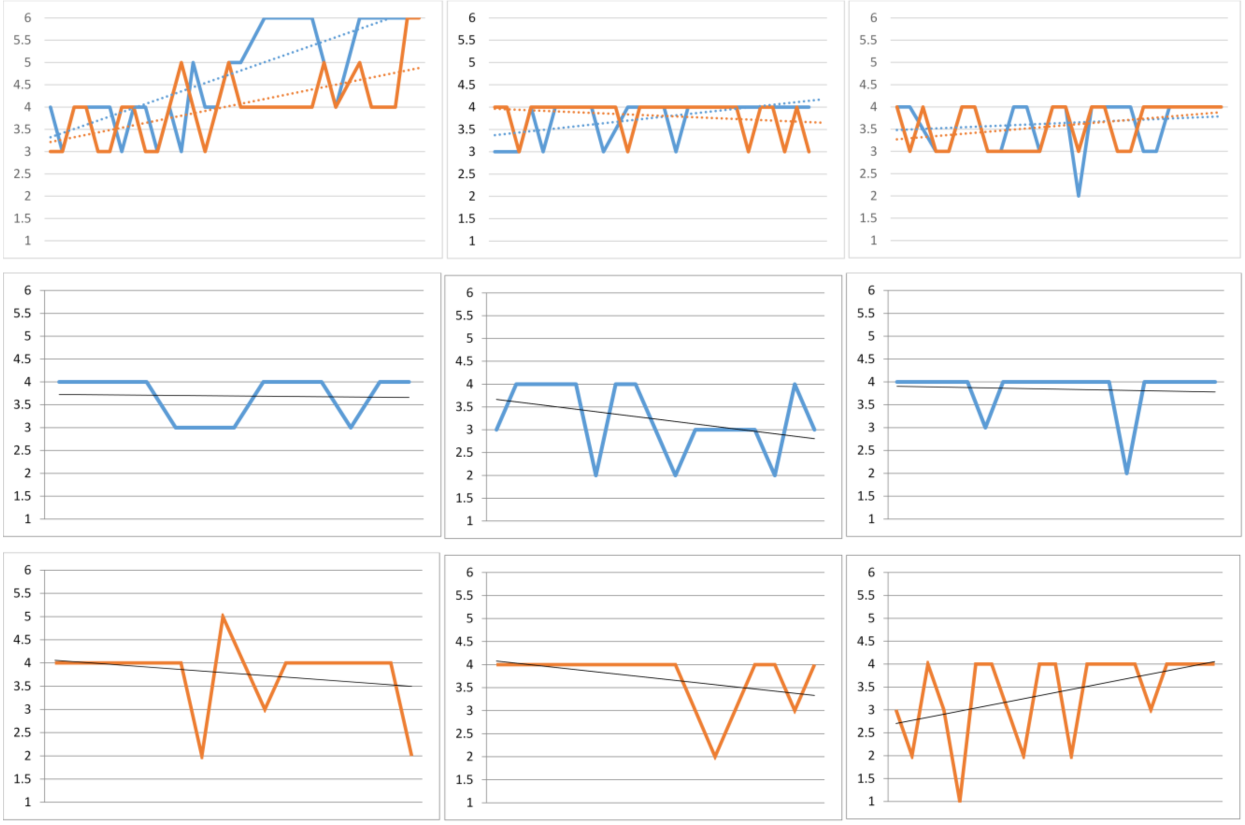

Figure 6. Three examples of trends across peer conditions. Top row is the engagement of two children playing *Mole Madness* together. Lower two rows are the engagement trends of the same children when each played separately with Sammy.

sense that both players enjoyed a high or increasing desire to remain engaged in the experience over time.

The variability we see implies that there is no "average" player and no "canonical" interaction with respect to engagement—the game and co-player afford a range of experiences. A robot peer that cannot detect and track changing levels of engagement and flexibly adapt its behavior to coordinate with them, is unlikely to produce an optimal experience for each child.

7. CHILD-SAMMY INTERACTION

As mentioned, each child also played *Mole Madness* with an autonomous but non-adaptive robot, Sammy J, in the *go* role. In this section we briefly explore the coders' perceptions of the children's willingness to continue to engage with the robot as co-player. Again, we use Coder 2 as our specific representative when one is needed.

The mean engagement score over all segments was lower for the child-Sammy sessions for all coders, although the difference for Coder 2 was negligible (3.78 to 3.75). As in the child-child games, audio-visual scores tended to be higher than video-only judgments for all coders. Coder 2's remained unchanged for 50% of segments, increased for 35% and decreased for 15%. And as before, most children benefited although some children benefited more than others, and which children did depended on the coder. For any given coder, there was no correlation between those who benefited from audio in the child-child condition and those who benefited with Sammy. We conjecture that there may have been individual differences in aspects of the child's vocalizations with friend versus robot, and intend to explore that possibility in future work.

We also examined the engagement trends of children in the child-robot condition. Although space precludes a complete side-by-side comparison, Figure 6 shows examples for six children when playing with Sammy, repeating the child-child graph for each pair. Despite the cumulative statistic being nearly identical, for many of the children, the game felt quite different. For children with flat or decreasing engagement, the ability of the system to detect the situation as it changes (or fails to change) and adapt its behavior to influence the long-term pattern seems critical to an emotionally successful interaction.

8. DISCUSSION AND FUTURE WORK

The work presented in this paper constitutes only a first step towards understanding how to detect, model and influence engagement in interactive spoken games with children. We treat engagement as a variable quantity that can rise and fall gradually over the course of an interaction, rather than as a single description of the interaction as a whole, or a binary moment-to-moment correlate of attention. We described our method and decision process in arriving at a more useful way to elicit engagement judgments via a proxy scale that both situates the coder's task in more concrete language and allows us to characterize the rise and fall of the measure over time. Although the "willingness" scale offered better inter-rater reliability among coders, and was comparable to what some other researchers have found, standard statistical measures were not high enough to conclude that all values meant exactly the same thing. We did find that most of the differences between coders corresponded to differences in degree rather than kind, but we do not know whether the differences in judgments ultimately correspond to substantively different mixtures of sense-able features in the environment. A clear next step to this work is to build models based on audio and visual features that predict the different levels of engagement, both to see whether the coders' differences matter in practice and to give Sammy the ability to see what the coders see.

A closer examination of one coder's data revealed variability of engagement level within a game as well as variability in overall patterns across different players as a function of both time and co-player. We have begun the work of exploring whether linguistic entrainment and other types of acoustic coordination are more or less present in what seem to be the desirable patterns and trends. Indeed, for the subset of this data for which we have high quality audio without dropout (eight pairs), we have found that some verbal and acoustic features demonstrate significantly more synchrony when the children are in a high engagement state (Chaspari et al, 2015). To respond in kind, Sammy's repertoire of behavior will need to be extended to produce comparable signals. When added to the ability to sense what path the child is on, such adaptive behaviors can be used to reinforce or change that movement in peer-like ways. The patterns of behavior in the current data, both child-child and child-robot, present a backdrop against which we can measure Sammy's growth toward producing more satisfying interactions for all children.

9. ACKNOWLEDGMENTS.

We would like to thank Preben Wik for advice on designing the personality of the robot character, Jon Lew for help in game design and programming, and Peijin Zhang for help in data collection and preparation.

REFERENCES

[1] Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. 2012. *Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction*. In Esposito, A., et al. (Eds.), Cognitive Behavioural Systems. Lecture Notes in Computer Science, pp 114-130.

[2] Al Moubayed, S. and Lehman, J F. 2015. *Design and Architecture of a Robot-Child Speech-Controlled Game.* In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction HRI2015, Portland, OR, USA.

[3] Aroyo, L. and Welty, C. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation, *AI Magazine,* AAAI Press, Volume 36, Number 1, pages 15-24.

[4] Bohus, D. and Horvitz, E. 2009. *Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings*, in Proceedings of SIGdial'09, London, UK.

[5] Chaspari, T., Al Moubayed, S., and Lehman, J. 2015. *Exploring Children's Verbal and Acoustic Synchrony: Towards Promoting Engagement in Speech-Controlled Robot-Companion Games*. In Proceedings of the First International Workshop on Modeling Interpersonal Syncrhony, ICMI.

[6] Gatica-Perez, D. 2009. Automatic nonverbal analysis of social interaction in small groups: A review, *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, November.

[7] Lehman, J. and Al Moubayed, S. 2015. *Mole Madness – a Multi-Child, Fast-Paced, Speech-Controlled Game*. AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction. Stanford, CA

[8] Leite, I., Mccoy, M., Ullman, D., Salomons, N., Scassellati, B., and Haven, N. 2015. *Comparing Models of Disengagement in Individual and Group Interactions*. In 10th ACM/IEEE International Conference on Human Robot Interaction Conference (2015).

[9] Nakano, Y and Ishii, R. 2010. *Estimating user's engagement from eye-gaze behaviors in human-agent conversations*. In Proc. of the 15th International Conf. on Intelligent User Interfaces, IUI '10, pages 139{148. New York, NY, USA, 2010. ACM.

[10] Oertel, C. 2010. Identification of Cues for the Automatic Detection of Hotspots, *Master's Thesis*, Bielefeld University.

[11] Rich, C., Ponsler, B., Holroyd, A. and Sidner, C. 2010. *Recognizing engagement in human-robot interaction*. In Proc. of the 5th ACM/IEEE International Conf. on Human-Robot Interaction, pages 375{382. IEEE.

[12] Sanghvi, J., Castellano, G., Leite, I. Pereira, A., McOwan, P., and Paiva, A. 2011. *Automatic analysis of affective postures and body motion to detect engagement with a game companion.* In Proc. of the 6th International Conf. on Human-robot Interaction, pages 305{312, New York, NY, USA, 2011. ACM.

[13] Sidner, C., Lee, C., Kidd, C., Lesh, N. and Rich, C. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140{164, 2005.

[14] Yu, Z., Papangelis, A. and Rudnicky, A. 21015. *TickTock: Engagement Awareness in a non-Goal-Oriented Multimodal Dialogue Systems.* In Proceedings of the AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction. Stanford, USA, 2015.