Design and Architecture of a Robot-Child Speech-Controlled Game

Samer Al Moubayed, Jill Fain Lehman
Disney Research Pittsburgh
Pittsburgh, PA, USA

{sameram,jill.lehman}@disneyresearch.com

ABSTRACT

We describe the conceptual design, architecture, and implementation of a multimodal, robot-child dialogue system in a fast-paced, speech-controlled collaborative game. In *Mole Madness*, two players (a user and an anthropomorphic robot) work together to move an animated mole character through its environment via speech commands. Using a combination of speech recognition systems and a microphone array, the system can accommodate children's natural behavior in real time. We also briefly present the details of a recent data collection with children, ages 5 to 9, and some of the challenging behaviors the system elicited that we intend to explore.

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

Keywords

Child-Robot Interaction, Gaze, Gesture, Speech, Spoken dialog, Multimodal systems, Facial animation, Natural Spoken Dialog.

1. MOTIVATION

In *Mole Madness*, two players control the movement of an on-screen character through coordinated speech. One player uses "go" to move the mole horizontally and the other uses "jump" to move it vertically, together avoiding obstacles and collecting rewards as the mole travels through each level (see Figure 1). The simplicity of the game's language and visuals allow even very young children to interact freely and exhibit natural behaviors, so we can study issues like engagement, addressee identification and turn-taking [Lehman, 2014].

As a two-player multimodal interactive game, *Mole Madness* provides a platform for studying collaborative language behavior and coordination to accomplish a task that is both time-sensitive and fast-paced. The game affords two distinct kinds of utterances: task speech and social talk. Task speech includes not just "go" and "jump" but also variants (e.g., "g- g- go" and "juuuuump") intended to enlarge the expressivity of the vocabulary. All other speech is social talk, which can be directed either to the mole or the other player. Despite the fast-paced nature of *Mole Madness* and the need to use the vocal channel to achieve game goals, social talk is ubiquitous in child-child gameplay and seems to be related to establishing and maintaining the players' emotional synchrony [Lehman and Al Moubayed, 2015].

HRI'15, March 2–5, 2015, Portland, Oregon, USA.

Figure 1. Screen shot of *Mole Madness* – the blue mole rolls and leaps via the speech commands, "go" and "jump".

This paper reports initial efforts in building an autonomous, robotic play companion for *Mole Madness*. Our goals are to explore and model language behavior, engagement, adaptability, and social bonding with children, as well as to develop technologies that can handle the kind of multimodal interaction children display in collaborative, interactive environments.

2. AUTONOMOUS COLLABORATIVE PLAY IN *MOLE MADNESS*

A robotic play companion needs both task-related and social skills to provide an engaging, fun and meaningful experience. Minimally, the robot must be able to: (1) play the game efficiently, (2) distinguish between utterances that are actionable game-control commands and those that are social, and (3) understand and adapt to the play style of the child. With respect to (2), the game further demands that task language processing occur in real-time; since the mole is constantly moving along gravity and trajectory velocity vectors, if the recognition of a "jump" needs to wait for 40 msec of silence to be recognized then the executed movement will be different from the one intended. Point (3) is particularly important with players whose preferred play styles—fast vs. slow, accurate vs. risky, collect points vs. progress quickly—may vary widely and for whom the selection of the wrong style may turn a fun experience into a frustrating one.

In the following sections we briefly describe the physical setup and basic modular components of an initial system implemented to meet these needs in games played with 40 children, ages 5 to 9.

3. PHYSICAL SET UP DESIGN

A back-projected anthropomorphic robot head ("Sammy J," based on [Al Moubayed et al. 2012]) was placed in an equilateral triangle with the user and a 40" television. The child was seated to try to keep his/her head aligned with Sammy J's (see Figure 2).

We centered a Dev-Audio conical Microcone array at the bottom of the television and a Microsoft Kinect2 at the top. The microphone array was used to sense voice activity, audio beam, and audio stream of the child, and to separate the audio coming from the child from the audio generated by the robot.

4. DIALOG ARCHITECTURE

We used the IrisTK state-charts dialog authoring tool [Skantze & Al Moubayed, 2012] to coordinate communication and the flow of interaction between the speech recognizers, the visual tracker, the robot head and the game. IrisTK provides a simple XML-based communication protocol for interaction among modules and processes on different platforms. Figure 3 shows the main components, described further below.

4.1 Game Module

Mole Madness is built in Unity3D. Basic game play uses A* to predict the best next move ("jump" or "go") for the mole according its current location, speed, and acceleration as well the obstacles surrounding it. The A* results were used to predict (recommend) the child's (and Sammy's) next best move. The game was designed with three short tutorials to introduce the child to the game mechanics and environment, followed by three full levels (of about one minute play time each) of play. For the results reported in Section 5, Sammy J always played "go."

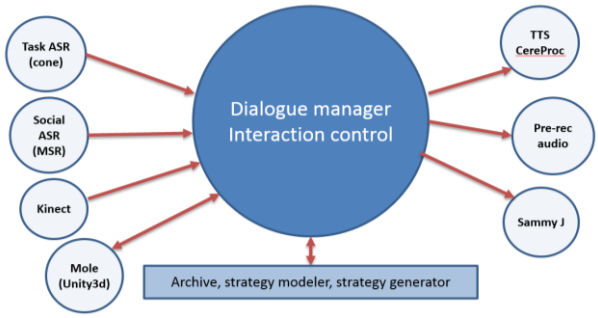

Figure 3. A diagram of the different modules.

4.2 Task ASR

As mentioned above, the system must respond to task speech commands in real time. To accomplish this we currently use simple voice activity detection in conjunction with the mean duration of "go" and "jump" found in a child-child pilot study [Lehman and Al Moubayed, 2015]. When the child's utterance is the approximate length of an average single "go" or "jump" then Task ASR hypothesizes a command, otherwise the decision is left for the large vocabulary speech recognizer, explained below.

4.3 Social ASR

To recognize game commands other than a single "go" or "jump," Microsoft's speech recognizer was used with a grammar containing multiple repetitions of task language (e.g. "go go go", "jumpjump"), and other social talk found in the child-child pilot [Lehman and Al Moubayed, 2015]. Social ASR also provided, with delay, either a confirmation or rejection of the hypothesis that was made by the Task ASR.

4.4 Robot Behavior Generation

Because we wanted to provide speech output with high enthusiasm, variability, and excitement, a voice actor recorded a large set of utterances for game control, comments on the game, tutorial style chitchat, and feedback to the player. The head and gaze control of the robot was also varied systematically depending on the type of feedback the robot gave (e.g., if an over-sized cabbage was captured, when a level was finished, etc.). At the beginning of each tutorial level, the robot engaged in a short dialog with the child to present key information. At the end of each game level, the robot gave excited feedback on how much it enjoyed the level and how well the child had played.

Figure 2. The robot head and a child playing *Mole Madness*.

5. USER STUDY AND FUTURE WORK

The system described here has been used to collect data from 40 children playing with Sammy J for an average of five minutes each. We intend to explore several issues with this data, but a primary focus is to understand and accommodate the large variability we see in the children's preferred style of gameplay. By style we mean more than simply rate of play and level of mastery in traversing the environment efficiently; style also involves things like the ratio of social to task language and the changes in energy and prosody contours that ebb and flow across the levels. It is clear from our pilot study that when children play *Mole Madness* with each other, they co-adopt a style of interaction that both keeps the game moving and feeds their excitement and enthusiasm. Replicating this aspect of the interaction requires the ability to tie a child's behaviors to key game events, interpret them in real-time and adjust the robot's behavior accordingly to maintain emotional synchrony, abilities we look forward to adding to Sammy J's repertoire.

6. ACKNOWLEDGMENTS

We would like to thank Jon Lew for the game implementation and Preben Wik for his work recording the speech of the robot.

7. 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 Cognitive Behavioural Systems. LNCS, pp. 114-130.

[2] Skantze, G., and Al Moubayed, S. 2012. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In Proceedings of ICMI. Santa Monica, CA.

[3] Lehman, J. 2014. Robo Fashion World: A Multimodal Corpus of Multi-child Human-Computer Interaction. In proceedings of ICMI'14 Understanding and Modeling Multiparty, Multimodal Interactions. Istanbul, Turkey.

[4] 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.