Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium

Mole Madness – A Multi-Child, Fast-Paced, Speech-Controlled Game

Jill Fain Lehman, Samer Al Moubayed
Disney Research
4720 Forbes Ave, 15213 Pittsburgh, PA, USA
{jill.lehman, samer}@disneyresearch.com

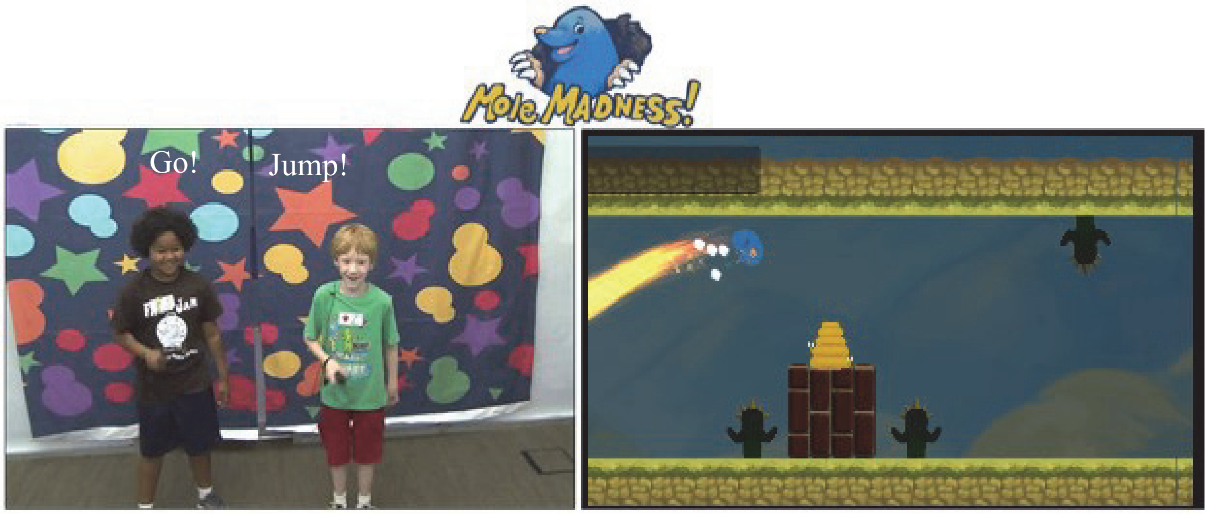

Figure 1. Left: Two children playing Mole Madness. The child to the left controls the horizontal movement of the mole by saying "go" and the child to the right controls the vertical movement by saying "jump." Right: A snapshot of the game during gameplay.

Abstract

We present *Mole Madness,* a side-scrolling computer game that is built to explore multi-child language use, turn-taking, engagement, and social interaction in a fast-paced speech-operated activity. To play the game, each of the two users controls the movement of the mole on one axis with either *go* or *jump.* We describe the game and data collected from 68 children playing in pairs. We then present a preliminary analysis of game-play vs social turn-taking, and engagement through the use of social side-talk. Finally, we discuss a number of interesting problems in multiparty spoken interaction that are encompassed by *Mole Madness* and present challenges for building an autonomous game player.

Introduction

Unrestricted, social, face-to-face interaction remains an unsolved problem because it requires the broadest semantic model and largest vocabulary, mechanisms for resolving ambiguity and reference to the physical environment, and intricate rules for turn-taking based on a rich model of the world, social stature, prior context, history and culture. To gain power and constraint in dialog systems, researchers recast these problems in task-specific terms: the semantics and vocabulary of a travel schedule (Raux et al. 2005), the turn-taking rules of a tutor and student (Al Moubayed et al. 2013; Huang and Mutlu, 2014), the conventions of a meeting format (Gatica-Perez, 2005).

But among ourselves, in all but the most stringent linguistic contexts (e.g. pilot and control tower), we tend to expand the task to include the social. For an autonomous agent, then, the possibility of unrestricted conversation lurks at the edge of every well-defined task interaction.

Previous research has found this to be particularly true of interactions among children at play. Banter, asides, and emotional outbursts conveying excitement and delight do not occur as a separable "chit chat" phase of an otherwise largely rule-following dialog; they are, instead, continuously interwoven into the activity and seem intrinsic to its enjoyment (Lehman, 2014, Yang Wang et al, 2012). Because our main goal is to create experiences that are engaging and fun, we must embrace the intrusion of the social and find our constraint elsewhere. In this paper we describe our work with a two-player game, *Mole Madness*, which was designed to explore issues in task versus social language. First we outline the game, then describe a data collection involving 68 children who played it in pairs. Next we discuss the corpus and present a preliminary characterization of the language and turn-taking behavior in terms of both the game's task demands and the children's social interaction. Finally, we describe the challenges presented in 1) building completely autonomous gameplay with multiple children, and 2) replacing one player with an autonomous agent who is able to adapt to, accommodate, and evoke similar kinds of behavior in a complex and natural environment.

Mole Madness: the 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) or point loss (cactuses, birds), and the occasional special object (star) that acts as a boost to change the mole's normal behavior. In addition to providing a familiar and engaging 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, overlapping, and orchestrated sequences of both commands.

Although players are not given any specific goal other than to move the mole through the environment to the end of the level, there is a score bar on the screen that updates as the mole touches the various kinds of objects. Whether through convention or just visual affordance, players seem to adopt maximizing speed and/or points as a goal.

The game occurs in the broader conversational context shown in Figure 2. On the speech level, any utterance can be directed toward the game on the screen or toward the other player. We expect utterances directed toward the game to fall into one of two categories: *task* commands (*go*, *jump*) or *social mole-directed talk*—out of vocabulary comments addressed to the mole that cannot result in an action (e.g., "go backwards fat mole," "watch out," "faster"). Utterances directed toward the other player are considered to be *social player-directed talk* (or, more simply *side-talk*), and can demand a response that would result in further social turns (e.g., "Are you planning to jump this one?" "Are you ready?"), but do not need to do so (e.g., "Oh you shouldn't have jumped," "Jump now," "Nice move"). Based on our previous experience (Lehman, 2014), we also expect *social side-talk* to sometimes include utterances with emotional content but no obvious addressee (e.g., "That's weird" "I like it").

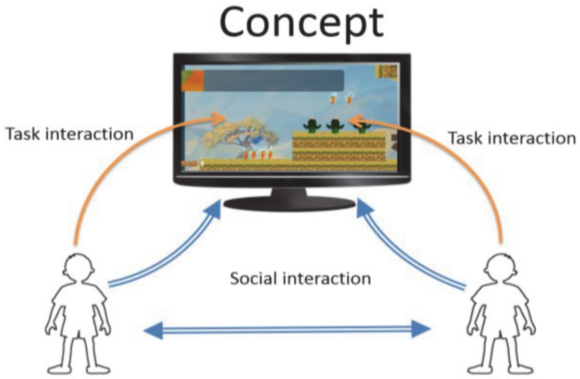

Figure 2. Social versus task interaction.

On the nonverbal behavioral level, the fast pace and visual processing demands of the game are expected to reduce 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 could be absent in an interaction that requires intense focus.

In essence, the speech patterns of the players can be understood as a conversation that is guided by the design of the level. Barriers, rewards, and obstacles translate into rules of turn-taking between the players, demanding instances of very fast turn-change (e.g., a *jump* that is needed immediately after a *go*) and overlapping speech (e.g., a *go* and *jump* together to get the mole to cross an obstacle, or social speech from one player while the other moves the mole along a single dimension).

User Study

Population and procedure. Thirty-four pairs of children, aged 4 to 10 (M = 7.15, s = 2.01 years), played *Mole Madness* as one activity within a larger data collection. Session participation was based on the convenience of scheduling for each family, so the population as a whole contained a mix of players who did and did not know each other with a variety of age gaps.

Game play began with a training phase in which players were told that they could control the mole's movement through its world with a combination of brief horizontal/vertical motions of the Wiimote™ in conjunction with the word *go*/*jump*. (In reality, only the Wiimote™ gesture was sensed; utterances were recorded as data for training speech recognition.) Children practiced

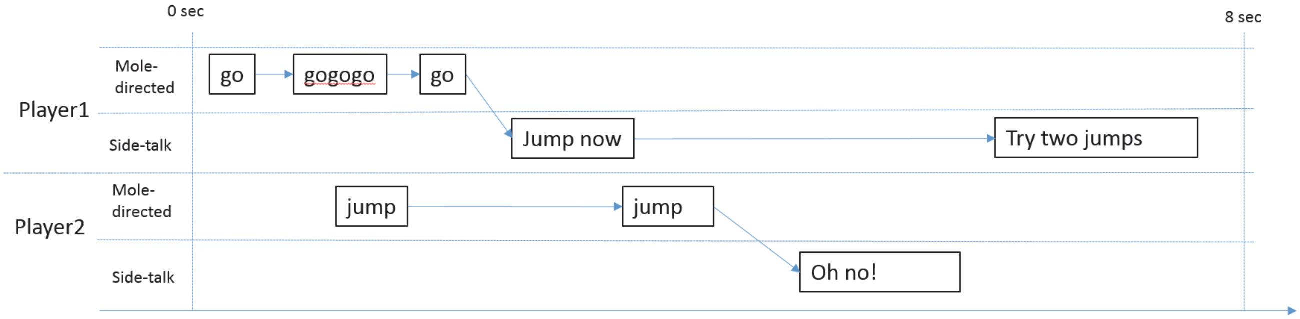

Figure 3. An excerpt of interaction showing social side-talk and task interaction from both players.

the combination a few times before playing a short level with a long stretch (all *go*), a tall wall (all *jump*) and a sequence of short obstacles (*go* and *jump* combined). For children unable to coordinate voice and movement (some four and five year olds), an experimenter controlled the Wiimote™ and the child provided the words.

After training, children played two games: a long level in the role that they had practiced (*go* or *jump*), followed by another long level in which they switched roles. Throughout, a wizard was responsible for detecting when children fell silent or had extended side conversations, and prompted them to continue to tell the mole what to do.

We ignore the data from the training phase, creating a corpus of 68 game sessions, two for each of the 34 pairs of children. Each session lasted an average of 4.3 minutes (s = 1.7 minutes), with a total corpus length of 6:10 hours.

Analysis. The corpus was first annotated in terms of game commands (*go*, *jump*, and any elision, pronunciation or combination of them). Annotators were instructed to transcribe each stretch of speech as a unit (Inter-Pausal Unit or IPU) with a 500ms silence threshold. Utterances that did not contain any instance of *go* or *jump* were segmented but not transcribed.

There were 7929 IPUs in the resulting corpus with a large variance in the number of utterances per session (M = 116, s = 45.2). Out of these turns, 82.9% were single instances or multi-instance strings of *go*s or *jump*s, and 17.1% were *other*. The *other* class includes pure instances of non-task interaction as well as instances of *social mole-directed talk* and *social side-talk* that were interleaved with *go*s or *jump*s without a 500msec pause.

IPUs in the *other* category were further annotated by two coders, with each utterance labeled as either *social mole-talk* or *social side-talk*. Of the 1359 non-task IPUs, only 64 (4.7%) were annotated as *mole-directed talk,* the remaining 95.3% (1295 IPUs) occurred between players. Figure 3 shows an excerpt from one session, visualizing the interleaving of task and *social side-talk* over time.

Although social utterances accounted for 17% of IPUs overall, they were not equally distributed across child pairs; the mean of 9.5 social utterances/game was accompanied by a standard deviation of 11.1, with the least socially-inclined child having no out-of-task utterances, and the most socially-inclined child having 48 in a single game. More important than this variability and range, however, is the significant correlation in the number of social utterances between children within a pair (r = .757, df = 66, p < .001). In other words, children tended to adopt more or less the same degree of social involvement, with some player pairs focused intensely on the game and other pairs engaged in social talk throughout. Where there were large imbalances in amount of social talk, co-players tended to be far apart in age; indeed, difference in age was, itself, significantly correlated with difference in number of social utterances (r = .243, df =66, p < .05). As a result we conjecture that the ability to recognize and respond to social utterances, even if only to match their frequency and general tone, will be of critical importance to both building an autonomous hands-free game and a robot buddy that can engage with the child as a peer that understands and adapts to the child's preference for social interaction.

Discussion

We have presented a brief description of *Mole Madness*, a two-player speech-controlled platform to study multiparty multimodal dialogue in a time-sensitive environment, as well an initial corpus of turn-taking behavior from 34 pairs of young players.

One of our main goals in creating *Mole Madness* is the development of an autonomous robot "buddy" that can interact naturally with a child while playing the game (see Figure 4, with Sammy J, a robot head based on (Al Moubayed et al. 2013a; Al Moubayed et al. 2012) and produced by Furhat Robotics AB).

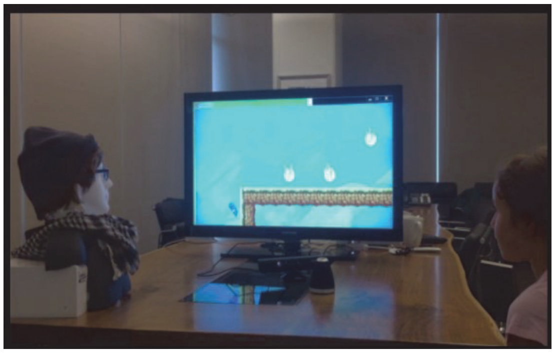

Figure 4. Sammy J and a child playing Mole Madness.

The design of *Mole Madness* raises many largely unexplored issues in multimodal, multiparty dialogue. From our initial analysis, the degree of social involvement within-session was typically well-balanced while the degree of social involvement across sessions was highly variable. This pattern may be due to high levels of alignment and adaptation between the two children, although how such alignment manifests quickly between strangers is unclear. It is also not clear how the amount of social engagement between the players is related to perceived success or level of engagement in the game itself. These are questions at the core of our future work, as developing an autonomous agent that can take the role of one of the players needs to account for these phenomena, adapting its style in response to the child's behavior.

Another turn-taking issue that arises is how to use task design to manage linguistic complexity. In *Mole Madness*, the design of the level affects the ability of the players to engage in social talk. Any particular layout of the environment and physics for the mole allows more or less of a "time window" during which the players have the chance to engage in social talk without affecting the game (losing). If banter, asides, and emotional outbursts are part of what makes the game fun then there needs to be enough time for such events to occur, but not so much time that we invite unconstrained social conversation. It may be that adaptive game-content generation can be used as a support mechanism for the autonomous agent when having difficulties dealing with social input from the child.

A final issue raised by our data seems unique to the intentional language limitations imposed by the game. We find many instances in which children tried to employ consistent speech variations as temporally-sensitive control commands. For example, children produced variants such as *g-g-g-g- go*, *gogogogo*, *juuuuump*, and *jumpjumpjump* in an attempt to influence the physics of the movement and its timing. Although there exist a handful of studies that attempt to explore the flexibility of non-verbal speech in games (Sporka et al. 2006; Harada et al. 2006), what children mean when they vary the loudness, pitch, duration of phones, syllables and words, and the pronunciation of the command itself is a new and interesting question. Taken together, these issues present challenges to building speech recognition systems and dialogue control that might need to be significantly different in approach from those used to process conversational speech.

Acknowledgments

The authors would like to thank Tushar Arora and Jon Lew for their work designing and implementing the game, and Preben Wik for his part in the design of Sammy J.

References

Abele, A. 1986. Function of gaze in social interaction: Communication and monitoring. J. Nonverb. Behav.10, 83–101.

Al Moubayed, S., Beskow, J., Bollepalli, B., Abdulaziz, A.H. et al. 2013b. Tutoring Robots – Multiparty multimodal social dialogue with an embodied tutor." In proceedings of eNTERFACE2013, Springer.

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.

Al Moubayed, S., Skantze, G., & Beskow, J. 2013a. The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. Int. Jour. of Humanoid Robotics, 10(1).

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.

Harada, S., Landay, J.A., Malkin, J., Li, X. and Bilmes, J.A. 2006. The vocal joystick: evaluation of voice-based cursor control techniques. In: Assets '06: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 197–204. ACM, New York.

Huang, C.-M., and Mutlu, B. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In Proc. HRI'2014, Bielefeld, Germany.

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 Workshop. Istanbul, Turkey.

Raux A, Langner B, Bohus D, Black AW, Eskenazi M. 2005. Let's Go Public! Taking a Spoken Dialog System to the Real World. In: Proceedings of Interspeech, September 4–8, Portugal; Lisbon. p. 885–8.

Sporka, A.J., Kurniawan, S.H., Mahmud, M. and Slavík, P. 2006. Non-speech input and speech recognition for real-time control of computer games. In: Assets '06: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 213–220. ACM, New York.

Yang Wang, W., Finkelstein, S., Ogan, A., Black, A., and Cassell, J. 2012. "Love ya, jerkface": Using sparse log-linear models to build positive and impolite relationships with teens. In Proceedings of SIGDIAL, pages 20–29.