Human-robot collaborative tutoring using multiparty multimodal spoken dialogue

Samer Al Moubayed[1], Jonas Beskow[1], Bajibabu Bollepalli[1], Joakim Gustafson[1], Ahmed Hussen-Abdelaziz[5], Martin Johansson[1], Maria Koutsombogera[2], José David Lopes[3], Jekaterina Novikova[4], Catharine Oertel[1], Gabriel Skantze[1], Kalin Stefanov[1], Gül Varol[6]

[1]KTH Speech, Music and Hearing, Sweden
[2]Institute for Language and Speech Processing, Greece
[3]Spoken Language Systems Laboratory, INESC ID, Portugal
[4]Department of Computer Science, University of Bath, UK
[5]Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[6]Bogazici University, Turkey

sameram@kth.se

ABSTRACT

In this paper, we describe a project that explores a novel experimental setup towards building a spoken, multi-modally rich, and human-like multiparty tutoring robot. A human-robot interaction setup is designed, and a human-human dialogue corpus is collected. The corpus targets the development of a dialogue system platform to study verbal and nonverbal tutoring strategies in multiparty spoken interactions with robots which are capable of spoken dialogue. The dialogue task is centered on two participants involved in a dialogue aiming to solve a card-ordering game. Along with the participants sits a tutor (robot) that helps the participants perform the task, and organizes and balances their interaction. Different multimodal signals captured and auto-synchronized by different audio-visual capture technologies, such as a microphone array, Kinects, and video cameras, were coupled with manual annotations. These are used build a situated model of the interaction based on the participants personalities, their state of attention, their conversational engagement and verbal dominance, and how that is correlated with the verbal and visual feedback, turn-management, and conversation regulatory actions generated by the tutor. Driven by the analysis of the corpus, we will show also the detailed design methodologies for an affective, and multimodally rich dialogue system that allows the robot to measure incrementally the attention states, and the dominance for each participant, allowing the robot head Furhat to maintain a well-coordinated, balanced, and engaging conversation, that attempts to maximize the agreement and the contribution to solve the task.

This project sets the first steps to explore the potential of using multimodal dialogue systems to build interactive robots that can serve in educational, team building, and collaborative task solving applications.

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

HRI'13, March 3–6, 2014, Bielefeld, Germany.
ACM 978-1-4503-1467-1/12/10.

Keywords

Human-Robot Interaction, Multiparty interaction, human-robot collaboration, Spoken dialog, Furhat robot, conversational management.

General Terms

Design, Algorithms, Human-Factors.

Overview

This work attempts to address social and interactional skills required by an embodied dialogue system to control the interaction flow as well as to boost and balance the engagement of the participants in the task they are involved in, while at the same time mitigating dominant behavior and encouraging less talkative interlocutors to equally participate in the interaction. The task and the setup chosen in this work are considered first steps towards understanding the behavior of a conversational tutor in multiparty task solving setup, as an example of a setup that can be used for applications in group-collaboration and negotiations, an activity that is highly dependent on the affective, and social behavior of the interlocutors [1]. Another main criteria that is taken into account while developing this setup is the ability to move directly from the models learnt from the annotations and analysis of the corpus, into an implementation of multiparty multimodal dialogue system, using the robot head Furhat [2]. Furhat was developed to support non-verbally and dynamically rich audio-visual synthesis, and to study human-robot spoken interactions [3,4], together with the newly developed IrisTK dialogue platform [5] both developed and utilized in multimodal multiparty embodied spoken dialogue systems.

The setup consisted of a card ordering game, in which two subjects had to engage in a conversation about the value of each card, and finally arrive at a decision about their order. Subjects were matched depending on their extr-

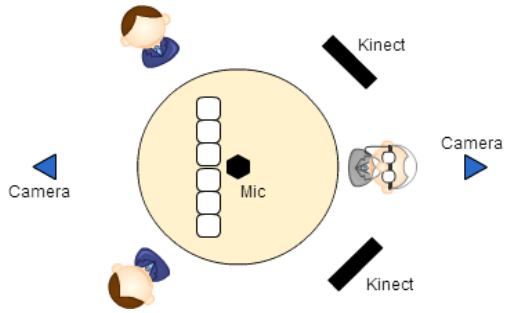

Figure 1. Snapshots of Furhat[1] in close-up and two users in interaction[2]

Figure 2. Top: a snapshot of the human-human data collection showing the tutor and the table, containing the colored cards, and the microphone array. Bottom: A view from one of the Kinects showing overlays of the real-time head-pose and torso-skeletal tracking.

Figure 3. Snapshots of Furhat in close-ups.

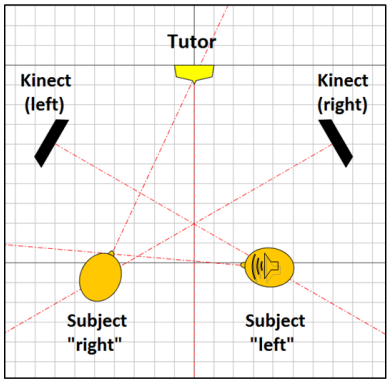

Figure 4. A visualization of the perceived activity of the situated interaction. The figure shows the tracked horizontal head rotation (pan) of each subject, and the voice activity detection (highlighted by an image of a speaker). The figure shows that the subject (left) is currently speaking.

[1] For more info on Furhat, see http://www.speech.kth.se/furhat

oversion (by filling a personality test before the recordings). The subjects were sitting with a tutor that attempted to coordinate the conversation and to regulate the interaction so that both participants get equal chance contributing to the decision making process. The data was recorded using real-time equipment and annotation. Participants' head, hands, and torso, were recorded using Kinects. The verbal activity of each participants and the tutor were captured using the table-top Microcone[3] microphone array, and the cards were automatically captured and tracked using a video camera placed on top. The data was then analyzed to understand and build a model of the head-pose, verbal and nonverbal feedback, and turn-management and interruption strategies the tutor employed. The tutor behavior was measured against the subjects' auto-recorded data, their verbal activity, and their dominance over time.

From this setup and data, that is soon to be publicly available, a highly complex dialogue system was built to simulate the tutor in these aspects. The dialogue setup benefited from the human recording in that it used exactly the same setup and equipment to track and analyze the conversation.

The paper on site will further describe data analysis results and experimental investigations on the behavior of the subjects and the performance of the dialogue system. Figure 1 shows a chart of the physical setup employed. Figure 2 shows snapshots of the human-human corpus recording. Figure 3 shows the robot head Furhat used as the tutor, and Figure 4 shows a real-time visualization of the interaction using the capture equipment.

ACKNOWLEDGMENTS

This work has been partly funded by the KTH ICT Strategic Research Areas Embodied Multimodal Communication.

REFERENCES

[1] Nass, C., Steuer, J., & Tauber, E. (1994), "Computers are social actors", *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, pp. 72–78

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

[3] Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The *Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction*. International Journal of Humanoid Robotics, 10(1).

[4] Al Moubayed, S., Edlund, J., & Beskow, J. 2012. *Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections*. ACM Transactions on Interactive Intelligent Systems, 1(2), 25.

[5] Skantze, G. and Al Moubayed, S. 2012. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*. Santa Monica, CA, USA.

[3] http://www.dev-audio.com/products/microcone/