A Collaborative Human-robot Game as a Test-bed for Modelling Multi-party, Situated Interaction

Gabriel Skantze, Martin Johansson, Jonas Beskow

Department of Speech Music and Hearing, KTH, Stockholm, Sweden
{skantze,vhmj,beskow}@kth.se

Abstract. In this demonstration we present a test-bed for collecting data and testing out models for multi-party, situated interaction between humans and robots. Two users are playing a collaborative card sorting game together with the robot head Furhat. The cards are shown on a touch table between the players, thus constituting a target for joint attention. The system has been exhibited at the Swedish National Museum of Science and Technology during nine days, resulting in a rich multi-modal corpus with users of mixed ages.

1 Introduction

Recently, there has been an increased interest in understanding and modelling multi-party, situated interaction between humans and robots [1,2,3,4,5]. To develop such models, we think that a test-bed is needed in which data can be collected, and which can be used to test out data-driven models based on this data. The test-bed should be robust enough to be used in a public setting, where a large number of interactions with naïve users can be recorded. In this paper (and demonstration), we present a dialog system that was exhibited during nine days at the Swedish National Museum of Science and Technology, in November 2014[1]. As can be seen in Fig 1, two visitors at a time could play a collaborative game together with the robot head Furhat [1]. On the touch table between the players, a set of cards are shown. The two visitors and Furhat are given the task of sorting the cards according to some criterion. For example, the task could be to sort a set of inventions in the order they were invented, or a set of animals by how fast they can run. This is a collaborative game, which means that the visitors have to discuss the solution together with Furhat. However, Furhat does not have perfect knowledge about the solution. Instead, Furhat's behaviour is motivated by a randomized belief model. This means that the visitors have to determine whether they should trust Furhat's belief or not, just like they have to do with each other. Thus, Furhat's role in the interaction is similar to that of the visitors, as opposed to for example a tutor role which is often given to robots in similar settings [4].

We think that this setup has several features that makes it useful as a test-bed for collecting data and testing out models for multi-party situated interaction. Firstly, it has proven to be fairly robust against speech recognition errors, since the multi-modal

[1] A video of the interaction can be seen at https://www.youtube.com/watch?v=5fhjuGu3d0I

adfa, p. 1, 2011.

nature of the setup allows the system to fall back on other modalities (head pose and movement of the cards) and still take part in a meaningful interaction. This is further enhanced by the multi-party setup which allows the visitors to have a meaningful discussion with each other even in cases where Furhat's understanding is limited. Secondly, the symmetry of such a setting allow us to compare the human behaviour towards each other with their behaviour towards the robot in order to (1) use the data as a model for Furhat's behaviour, (2) investigate to what extent they interact with the robot as if it was a human interlocutor, and (3) evaluate how human-like the robot's current behaviour is, and how it could be improved [3]. A third important feature of this setup is that it involves discussion about objects in the physical space, where the interlocutors' visual focus of attention (VFOA) must be shared between each other and the objects under discussion, which has a clear effect on their turn-taking behaviour [3]. Many previous studies on multi-party interaction have mainly focused on interactions where this is not the case [2,5].

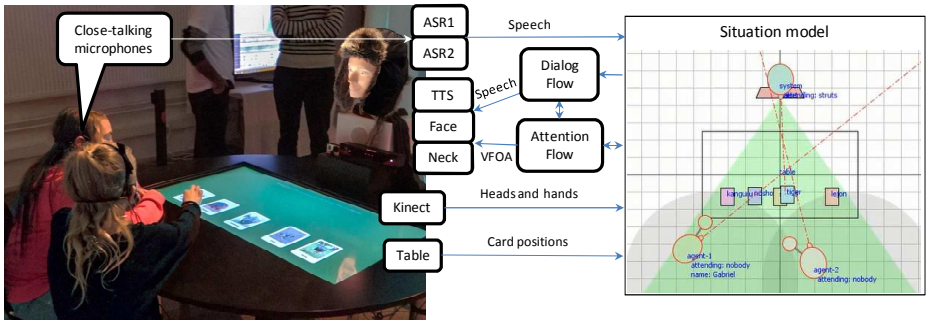

Fig. 1. A schematic illustration of the setting and architecture

2 System description

The system was implemented using the open source dialogue system framework IrisTK[2] [6] and is schematically illustrated in Fig. 1. The visitors are interacting with the Furhat robot head [1], which has an animated face back-projected on a translucent mask, as well as a mechanical neck, which allows Furhat to signal his focus of attention using a combination of head pose and eye-gaze. A Kinect camera (V2) is used to track the location and rotation of the two users' heads, as well as their hands. The low-level events from the different sensors (Kinect, Touch table and ASR) are sent to a *Situation model*, which translates the local coordinates to a common 3D representation, and then generates high-level events for the combined sensory data. This way, speech recognition results from the microphones can be mapped to the right users based on their location, regardless of whether it is a microphone array or a close-talking microphone. Another task of the Situation model is to keep track of when users engage and disengage in the interaction.

[2] http://www.iristk.net

Two behaviour controllers based on the Harel statechart mechanism offered by IrisTK run in parallel: The *Dialog Flow* and the *Attention Flow*. The Attention Flow keeps Furhat's attention to a specified target (one or both users, or a card), even when the target is moving, by consulting the Situation model. The 3D position of the target is then transformed into neck and gaze movement of Furhat (again taking Furhat's position in the 3D space into account). This, together with the 3D design of Furhat, makes it possible to maintain exclusive mutual gaze with the users, and to let them infer the target of Furhat's gaze when directed towards the cards, in order to maintain joint attention [1]. Since the public setting of a museum is very noisy, the users were wearing close-talking microphones. The speech recognition is done with two parallel cloud-based large vocabulary speech recognizers, which allows Furhat to understand the users even when they are talking simultaneously.

The Dialogue Flow module orchestrates the spoken interaction, based on input from the speech recognizers, together with events from the Situation model (such as cards being moved, or someone leaving or entering the interaction). The head pose of the users is used to determine whether they are addressing Furhat, and he should contribute to the discussion, or whether they are discussing with each other, and he should just provide backchannels to signal that he is still keeping track of what they are saying. In order to provide meaningful comments about the cards being discussed, The Dialog Flow maintains a *focus stack*. Cards are primed in the focus stack when their names are detected in the speech recognition, or when they are being moved. Since Furhat's role should be similar to that of the visitors', and therefore not have perfect knowledge, his behaviour is motivated by a randomized belief model. This is generated by taking the correct value (e.g. *speed* or *year*) for each card and then apply a random distortion. In addition, a random standard deviation is calculated for each belief, which represents Furhat's confidence in his belief. The outcome of these parameters is governed by two constants: *Ignorance* and *Uncertainty*, which may be used to tune Furhat's general behaviour. This belief model can be used to allow Furhat to compare two cards using a Z-test and assess his belief about their order and his confidence in this belief, which will in turn affect his choice of words. For example, he could say "I am quite sure the kangaroo is faster than the elephant", or "I have no idea whether the telescope was invented before the printing press".

3 Discussion and Future work

During the 9 days the system was exhibited at the Swedish National Museum of Science and Technology, we recorded data from 373 interactions with the system, with an average length of 4.5 minutes. The dataset contains mixed ages: both adults playing with each other (40%), children playing with adults (27%), and children playing with each other (33%). After completing one game, the players could choose to continue playing. 58 % of the pairs who completed the first game chose to do so.

In the current setting, most of Furhat's behaviour is motivated by hand-crafted policies that were tuned during extensive testing. However, we think that the setup described in this paper serves as a very useful test-bed for collecting data on situated

interaction (as we have done), and then use this data to build data-driven models for Furhat's behaviour. The test-bed then allows us to evaluate these models in the same setting that the data was collected. The models we are working on include detection of turn-taking relevant places based on a large range of multi-modal cues (e.g. head pose, words, prosody, movement of cards). Another interesting problem is to determine where Furhat's visual focus of attention should be. Since we have data on the users' attentional behaviour (approximated by head pose), and their roles are similar to that of Furhat's, this could be used to guide a data-driven model.

The exhibition at the museum also showed that both children and adults enjoyed the game, with several of them coming back to play more. We therefore think that the game could be a useful application in its own right, for example in an educational setting. The simple game concept (sorting of cards) allows the content to be easily customized for a particular subject.

Acknowledgements. This work is supported by the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (2011-6237) and *KTH ICT-The Next Generation*. Thanks to everyone helping out with the exhibition: Saeed Dabbaghchian, Björn Granström, Joakim Gustafson, Raveesh Meena, Kalin Stefanov and Preben Wik.

References

[1] 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).

[2] Bohus, D., & Horvitz, E. (2011). Decisions about turns in multiparty conversation: from perception to action. In *ICMI '11 Proceedings of the 13th international conference on multimodal interfaces* (pp. 153-160).

[3] Johansson, M., Skantze, G., & Gustafson, J. (2014). Comparison of human-human and human-robot Turn-taking Behaviour in multi-party Situated interaction. In *International Workshop on Understanding and Modeling Multiparty, Multimodal Interactions, at ICMI 2014*. Istanbul, Turkey.

[4] Al Moubayed, S., Beskow, J., Bollepalli, B., Hussen-Abdelaziz, A., Johansson, M., Koutsombogera, M., Lopes, J., Novikova, J., Oertel, C., Skantze, G., Stefanov, K., & Varol, G. (2014). Tutoring Robots: Multiparty multimodal social dialogue with an embodied tutor. In *Proceedings of eNTERFACE2013*. Springer.

[5] Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational Gaze Mechanisms for Humanlike Robots. *ACM Trans. Interact. Intell. Syst., 1*(2), 12:1-12:33.

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