# Modelling situated human-robot interaction using IrisTK

**Gabriel Skantze and Martin Johansson**
Department of Speech Music and Hearing, KTH
Stockholm, Sweden
{skantze,vhmj}@kth.se

## Abstract

In this demonstration we show how situated multi-party human-robot interaction can be modelled using the open source framework IrisTK. We will demonstrate the capabilities of IrisTK by showing an application where two users are playing a collaborative card sorting game together with the robot head Furhat, where the cards are shown on a touch table between the players. The application is interesting from a research perspective, as it involves both multi-party interaction, as well as joint attention to the objects under discussion.

## 1 Introduction

Recently, there has been an increased interest in understanding and modelling multi-party, situated interaction between humans and robots (Bohus & Horvitz, 2011; Mutlu et al., 2012: Johansson et al., 2014; Al Moubayed et al., 2014). In situated interaction, the system is typically embodied and the space in which the interaction takes place is of importance. By modelling the physical situation, the system can track multiple users (and possibly system agents) that enter and leave the interaction. Also, the discussion can involve objects in the shared space. The possibility to model this kind of interaction is facilitated by the many affordable sensors that are becoming available, such as Microsoft Kinect. However, while there are many examples of research systems that can engage in situated interaction (Bohus & Horvitz, 2011; Mutlu et al., 2012), the combination of all these techniques together with spoken dialog technology is not trivial, and it might be hard for a novice to put such systems together. Face-to-face interaction involves a large amount of real-time events that need to be orchestrated in order to handle phenomena such as overlaps, interruptions, coordination of head pose and gaze in turn-taking, etc. Also, the knowledge to develop and put together the necessary modules is of a very interdisciplinary nature. This calls for a dialog system toolkit for multi-party face-to-face interaction, which provides necessary modules for multimodal input and output and allows the developer or researcher to author the dialog flow in a way that is simple to understand for the novice, yet powerful enough to model more sophisticated behaviours.

At KTH, we are developing the open source Java-based framework IrisTK (www.iristk.net), which has exactly this purpose (but can of course also be used for speech-only systems). Since we first presented it (Skantze & Al Moubayed, 2012), the framework has matured and has been applied in many different settings (Johansson et al., 2014; Al Moubayed et al., 2014; Skantze et al., 2014). In this demonstration, we will show a system that was implemented using IrisTK, and which was exhibited at the Swedish National Museum of Science and Technology, in November 15-23, 2014[1]. As can be seen in Figure 1, two visitors at a time can play a collaborative game together with the robot head Furhat (Al Moubayed et al., 2013). 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

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



*Proceedings of the SIGDIAL 2015 Conference*, pages 165–167,
Prague, Czech Republic, 2-4 September 2015. 

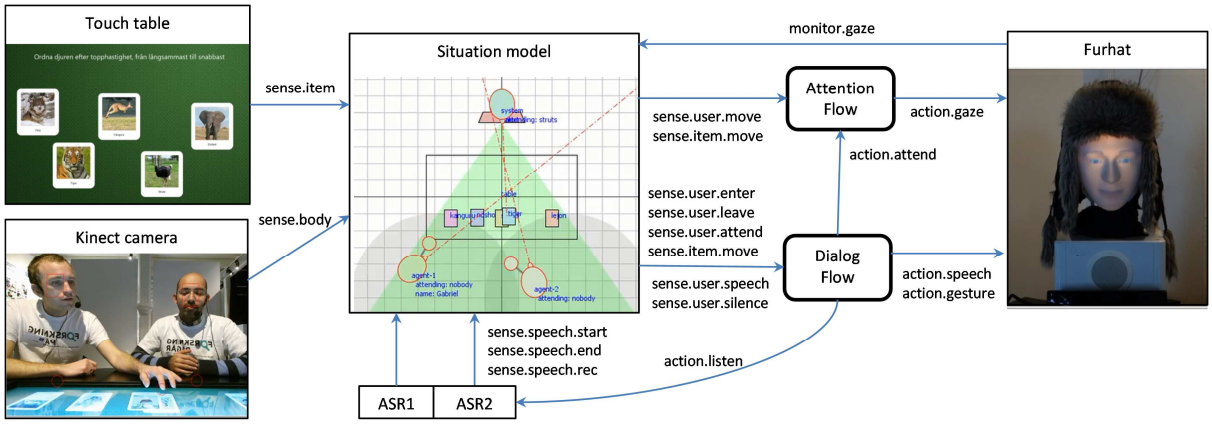


**Figure 1.** A schematic illustration of some of the modules and events used in the card sorting application.

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 (cf. Al Moubayed et al., 2014).

## 2 Overview of IrisTK

The system architecture is schematically illustrated in Figure 1. IrisTK provides a large set of modules for processing multimodal input and output, and for dialogue management, that can be put together in different ways. The framework defines a set of standardized events (as can be seen in Figure 1), which makes it possible to easily switch different modules (such as system agents or speech recognizers), as well as implementing new ones.

### 2.1 Vision and Situation modelling

A Kinect camera (V1 or V2) can be used to track the location and rotation of the two users' heads, as well as their hands. The head pose of the users can for example be used to determine whether they are addressing Furhat or not. This data, together with the position of the five cards on the touch table are sent to a Situation model, which maintains a 3D representation of the situation (as seen in Figure 1). The task of the Situation model is to take all sensor data and merge them into a common coordinate system, assign speech events to the right users based on the spatial configuration, and produce higher-level events.

### 2.2 Speech processing

IrisTK supports different combinations of microphones and speech recognisers. In the museum setup, we used close talking microphones together with two parallel cloud-based large vocabulary speech recognizers, Nuance NDEV mobile [2], which allows Furhat to understand the users even when they are talking simultaneously. However, the modularity of the framework makes it very easy to use the array microphone in the Kinect sensor instead. It is also possible to use SRGS grammars for speech recognition and/or semantic parsing, as well as extending the audio processing chain to add for example prosodic analysis.

### 2.3 IrisFlow

IrisTK also provides an XML-based formalism (IrisFlow) for rapidly developing behaviour modules, based on the notion of Harel statecharts (Harel, 1987) and similar to SCXML[3]. As discussed in Skantze & Al Moubayed (2012), this formalism combines the intuitiveness of Finite State Machines with the flexibility and expressivity of the Information State Update approach to dialogue management. As can be seen in Figure 1, we use two such behaviour modules running in parallel for the museum application: one for dialogue management and one for maintaining Furhat's attention. Thus, IrisFlow can be used to script both higher-level and lower-level behaviours. The Dialogue Flow module orchestrates the spoken interaction, based on events from the Situation model, such as someone speaking, shifting attention, entering or leaving the interaction, or moving cards on the table. The Attention Flow keeps Furhat's attention to a specified target (a user 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

[2] http://dragonmobile.nuancemobiledeveloper.com/
[3] http://www.w3.org/TR/scxml/

Furhat (again taking Furhat's position in the 3D space into account).

### 2.4 System output

For face-to-face interaction, IrisTK provides an animated agent that can be presented on a screen. While this solution suffice when only one person is interacting with the system, it does not work so well for multi-party interaction, due to the Mona Lisa effect (Al Moubayed et al., 2013), which means that it is impossible to achieve mutual gaze with only one of the users, or for users to infer the target of the agent's gaze in the shared space (such as the cards on the table). The preferable solution is to instead use a robot. IrisTK currently supports the Furhat robot head[4], but we are working on supporting other robot platforms. Furhat 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. The animation solution makes it possible to express subtle and detailed facial gestures (such as raising the eye brows or smiling), as well as accurate lip sync. The facial manifestation is completely decoupled from the speech synthesis, so that different agents can be combined with different speech synthesizers.

## 3 Discussion

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. To this end, IrisTK provides many tools for easily logging all events in the system, as well as the audio. Thus, we think that IrisTK is an excellent tool for doing research on situated interaction.

Apart from being used for research, IrisTK has also been used for education at KTH. In the course *Multimodal interaction and interfaces*, given to master students, it is used both for a three hour lab on conversational interfaces, as well as a platform for group projects. Only with two–three weeks of work and with little need for supervision, the students have used IrisTK to implement systems for travel booking, city exploration, cinema ticket booking, an interactive calendar and a virtual doctor[5].

We are still working on several ways to improve IrisTK. Currently it only runs on Windows (although it should be easy to port since it is Java based). We are also working on adding modules for face recognition, so that the system can maintain a long-term relationship with the users. Another improvement will be to add support for other robot platforms, such as NAO, which would also make it possible to explore body gestures. Another extension will be to combine the authoring of the flow with statistical models, such as reinforcement learning, so that some behaviours can be learned through interaction with users.

[4] http://www.furhatrobotics.com

[5] Videos of these system can be seen at http://www.iristk.net/examples.html

## Acknowledgements

This work is supported by the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (2011-6237).

## References

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.

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).

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).

Harel, D. (1987). Statecharts: A visual formalism for complex systems. *Science of Computer Programming, 8*, 231-274.

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.

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.

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.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication, 65*, 50-66.