Multimodal Multiparty Social Interaction with the Furhat Head

Samer Al Moubayed, Gabriel Skantze, Jonas Beskow, Kalin Stefanov, Joakim Gustafson
Department of Speech, Music and Hearing
KTH Royal Institute of Technology
Lindstedtsv. 24, 10044 Stockholm, Sweden
{sameram, skantze, beskow, kalins, jocke}@kth.se

ABSTRACT
We will show in this demonstrator an advanced multimodal and multiparty spoken conversational system using Furhat, a robot head based on projected facial animation. Furhat is a human-like interface that utilizes facial animation for physical robot heads using back-projection. In the system, multimodality is enabled using speech and rich visual input signals such as multi-person real-time face tracking and microphone tracking. The demonstrator will showcase a system that is able to carry out social dialogue with multiple interlocutors simultaneously with rich output signals such as eye and head coordination, lips synchronized speech synthesis, and non-verbal facial gestures used to regulate fluent and expressive multiparty conversations.

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
Multiparty interaction, Gaze, Gesture, Speech, Spoken dialog, Multimodal systems, Facial animation, Robot head, Furhat, Microphone Tracking

General Terms
Design

1. The Furhat Head
Furhat [1] is a robot head that deploys a back-projected animated face that is realistic and human-like in anatomy. Furhat relies on a state-of-the-art facial animation architecture allowing accurate synchronized lip movements with speech, and the control and generation of non-verbal gestures, eye movements and facial expressions.

Furhat is built to study, implement and validate patterns and models of human-human and human-machine situated and multiparty multimodal communication, a study that demands the co-presence of the talking head in the interaction environment, something that cannot be achieved using virtual avatars displayed on flat screens [2,3]. In Furhat, the animated face is back-projected on a translucent mask that is a printout of the animated model. The mask is then rigged on a 2DOF neck to allow for the control of head movements. Figure 1 shows snapshots of Furhat in different contexts.

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.
ACM 978-1-4503-1467-1/12/10.

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

2. Multimodal Input
For speech recognition, we used the Windows 7 ASR, running in two separate modules, one for each microphone. This allowed the system to process simultaneous speech in both microphones.

In addition to the speech signal from several microphones, the system uses the SHORE™ 2 real-time, robust and multi-person face tracking developed by Fraunhofer [4]. The tracker provides the system with information about the location and the pose of the different visible faces. The tracker also provides non-verbal information about the faces such as estimates of the age, gender, and facial expressions. Such information is utilized by the dialogue system: estimates of the facial expressions are used for example to establish facial mimicry between Furhat and the interlocutor.

To enable the system to engage in multi-party dialogue, it must be able to detect the speaker in the set of possible speakers in front of Furhat. To establish this, two microphones were rigged with two different patterns of Infra-Red reflective markers, and an Infra-Red camera is used in combination with the video camera to detect and track the microphones and align them to the faces. Figure 2 shows the setup of the microphones and the cameras, and Figure 3 shows the face and microphone trackers in action.

[1] For more info on Furhat, see http://www.speech.kth.se/furhat
[2] Fraunhofer SHORE™
http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst

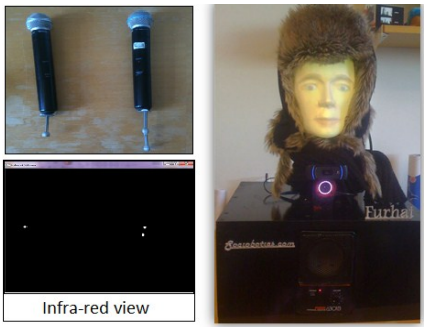

Figure 2. Left top: a snapshot showing two microphones rigged with Infra-Red markers. Left bottom: A screen capture showing a processed Infra-Red view of the microphones. Right: Furhat rigged with the Infra-Red and video cameras.

Figure 3. A snapshot of the face tracker and the microphone tracker in action.

3. Multiparty Dialogue

For every microphone, an ASR engine is used. Each ASR engine used two parallel language models, one context-free grammar with semantic tags (SRGS3), tailored for the domain, and one open dictation model. To interpret the dictation results, we have implemented a robust parser that uses the SRGS grammar to find islands of matching fragments.

The multi-modal dialog system is implemented using a newly developed framework based on the notion of statecharts [5] which is a powerful formalism for complex, reactive, event-driven systems. Using multimodal input events such as speech input, entry and movements of interlocutors, the demonstrator will show a social dialogue system that supports mixed initiatives (the system and the user can alternate the initiatives in asking questions). The dialogue is designed to account for multiparty properties, such as interruptions and overlaps, posing open questions to all interlocutors, and context and memory of previous questions and addressees during the dialogue.

4. Multimodal Output

The speech synthesis used is CereVoice developed by CereProc4. The speech output is automatically synced with accurate lips movements using the visual speech synthesis architecture used in Furhat [6]. In addition to speech, head and eyes coordinated movements are used by an attention control module to regulate attentive multiparty interaction in a highly human-like fashion,

[3] http://www.w3.org/TR/speech-grammar/
[4] CereProc ltd: http://www.cereproc.com/

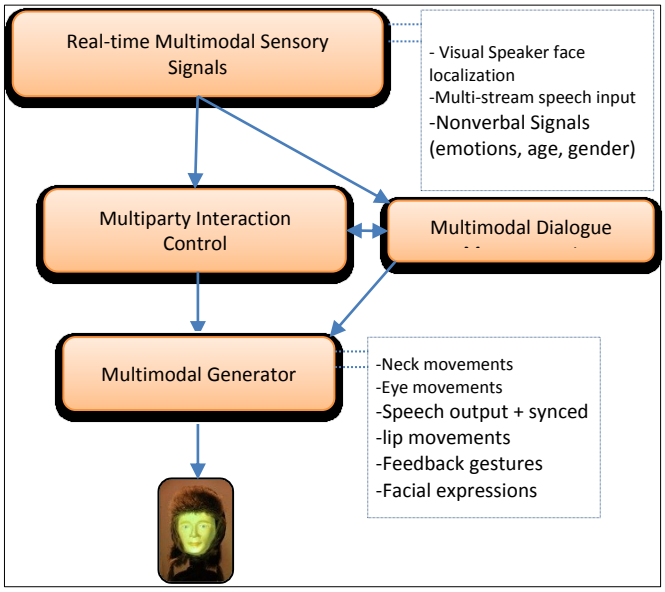

Figure 4. A simplified diagram of the architecture of the system.

such movements serve to the following and alternating between the interlocutors, and used to model cognitive states such as "idle" and "thinking". In addition to that, gestures are used as non-verbal facial signals to feedback, mimicry, attentiveness, and interest.

Figure 4 shows a simplified scheme diagram of the architecture of the system.

5. ACKNOWLEDGMENTS

This work has been partly funded by the EU project IURO (Interactive Urban Robot) No. 248314, and the SAVIR project (Situated Audio-Visual Interaction with Robots) funded by the Swedish Government (strategic research areas).

6. REFERENCES

[1] 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.

[2] 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*.

[3] Al Moubayed, S., & Skantze, G. 2011. Turn-taking Control Using Gaze in Multiparty Human-Computer Dialog: Effects of 2D and 3D Displays. *In Proceedings of AVSP*. Florence, Italy.

[4] Kueblbeck, C. and Ernst, A. 2006. Face detection and tracking in video sequences using the modified census transformation. *Journal on Image and Vision Computing*, vol. 24, issue 6, pp. 564-572, 2006, ISSN 0262-8856

[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.

[6] Beskow, J. 1997. Animation of talking agents. In Benoit, C., & Campbel, R. (Eds.), *Proc of ESCA Workshop on Audio-Visual Speech Processing* (pp. 149-152). Rhodes, Greece.