The Furhat Social Companion Talking Head

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze

Department of Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm, Sweden
{sameram,beskow,Skantze}@kth.se

Abstract

In this demonstrator we present the Furhat robot head. Furhat is a highly human-like robot head in terms of dynamics, thanks to its use of back-projected facial animation. Furhat also takes advantage of a complex and advanced dialogue toolkits designed to facilitate rich and fluent multimodal multiparty human-machine situated and spoken dialogue.

The demonstrator will present a social dialogue system with Furhat that allows for several simultaneous interlocutors, and takes advantage of several verbal and nonverbal input signals such as speech input, real-time multi-face tracking, and facial analysis, and communicates with its users in a mixed initiative dialogue, using state of the art speech synthesis, with rich prosody, lip animated facial synthesis, eye and head movements, and gestures.

Index Terms: Furhat, Robot Head, Human-Robot Interaction, Multiparty Dialogue, Social Signal Processing.

1. Introduction

Building an artificial and embodied social companion is one main vision of the many applications of speech technologies. Putting together a system that is able to multimodally interact with naturally moving and speaking interlocutors, and under human terms, is no easy task. Such systems require the robust and accurate functionality of a large number of technologies, all operating under rules borrowed from models human-human interaction.

Although arriving to such advanced artificial companions, or talking agents seems to be an unsolved problem, solutions to several of these technologies are becoming more available and robust, in a way we believe, when designed to maximize the efficiency of the system for specific tasks, is able to forward current state of the art in artificial conversational agents. These systems would not only be important as a validation to the underlying technologies, but also provide an indispensible tool to study patterns and applications of human-machine interaction that is not possible to investigate from human-human interaction data.

Recently, several researchers have taken the extra mile and built embodied conversational agents (ECAs) that are able to interact multimodally with interlocutors and using rich contextual data (e.g. [1,2]), but as most systems have used in screen animated avatars, this has limited the system in its ability to enable situated and multiparty interaction with its users [5].

2. Furhat

In this demonstrator we present the Furhat[1] robot head [3]. We built Furhat to study and evaluate rich and multimodal models of situated spoken dialogue. Furhat is a robot head that consists of an animated face that is projected using a micro projector on a three dimensional physical mask that matches in design the animated face that is projected on it. The state of the art animation models used in Furhat produce synchronized articulatory movements in correspondence to output speech [4], and allow for highly accurate and realistic control of different facial movements. The head is also supported with a 3DOF neck for the control of its head-pose.

The solution to build a talking head using the technique used in Furhat is superior in that: 1) Using a three dimensional head allows for situated and multiparty interaction that is not possible to establish accurately with avatars projected on two dimensional surfaces, thanks to its ability to eliminate the so-called Mona Lisa gaze effect [5,6], and 2) The use of facial animation instead of other mechatronic solutions to build robot heads enables the use of highly advanced and natural dynamics that are not so easily possible with mechanical servos and artificial skin, thanks to the advanced in facial animation techniques [4]. Furhat, in addition to being a platform to implement models of spoken human-human interaction, has become a vehicle to facilitate research on human-robot interaction [7,8], such as studying the effects of gaze movements in multiparty turn-taking [6, 9], audio-visual intelligibility of physically three dimensional avatars [10], and effects of head-pose on accuracy of addressee selection [12].

Figure 1 shows Furhat in interaction with a user, and Figure 2 shows the insides of Furhat's head and the hardware.

Figure 1. A photograph of Furhat interacting with a subject at the Robotdalen Innovation Challenge 2013.

Figure 2. Photographs showing the inner hardware and system rig of Furhat.

[1] For info and videos of Furhat: www.speech.kth.se/furhat

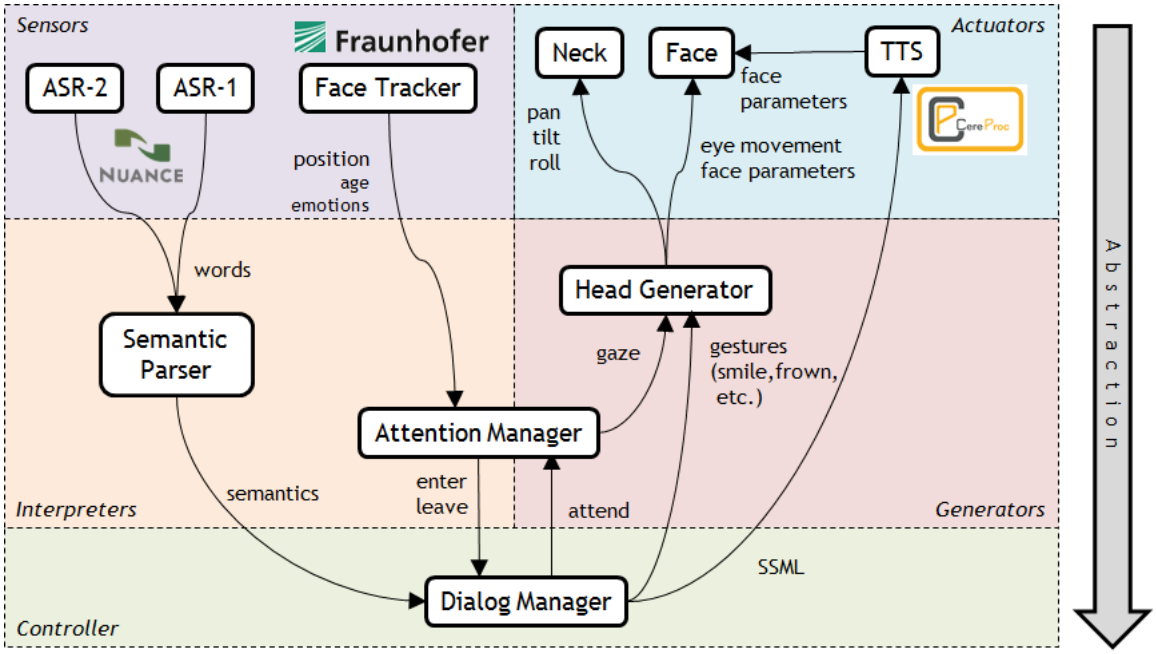

Figure 3. An example flow diagram of the hierarchal event driven system's components, events, and their interaction.

3. System Architecture

We present in this demo a system that is capable of interacting audio visually with multiple interlocutors at the same time. Using input signals such as speech recognition, speech activity detection, face tracking, and facial analysis, the system is capable of generating a rich set of verbal and nonverbal signals (such as gestures, facial colours, eye and head movements, and acoustic prosody) that render the system socially aware, and fluent [13].

The speech synthesis used is the prosodically rich, and highly natural CereVoice developed by CereProc1. The speech output is automatically synced with accurate lips movements using the visual speech synthesis architecture used in Furhat.

In this system, users are able to speak to Furhat using wireless handheld microphones. For every microphone, a nuance[2] ASR engine is used. Each ASR engine uses language models built using previous interactions with Furhat at other demonstrators, such as the one at the London Science Museum [14]. In addition to the speech signal from several microphones, the system uses the SHORE™ 3 real-time, robust and multi-person face tracking developed by Fraunhofer [15]. 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.

The multi-modal multiparty dialog system is implemented using a newly developed framework based on the notion of statecharts [12] which is a powerful formalism for complex, reactive, event-driven systems. Using multimodal input events such as speech input, entry, movements, and the location of interlocutors, and other facial analysis data, 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. Joint Attention

To enable the dialogue manager to handle the dialogue content without paying attention to the specific devices used to capture input, and how the head needs to implement output commands, an attention controller is used that, in principle, translates raw low level input data into abstract messages (such as face tracking data, into presence and location of interlocutors), and output abstract commands such as "attend user 2" or "look away", to low level facial, eye and head movements and gestures. The purpose of building an attention control component in such dialogue systems is to keep track of the attention state of the system and of its interlocutors, providing the system with messages that are independent of how the decision on that attention state is calculated, allowing for simple replacement of devices used in the system without affecting its functionality and the need for any customization.

5. Acknowledgements

This work has been done at the Department for Speech, Music and Hearing, and has been funded by: the EU project IURO (Interactive Urban Robot) No. 248314, The EU EIT-Kic project CaSA (Computers as Social Agents, , RIHA 12124), and the KTH Strategic Research Areas project Embodied Multimodal Communication.

[1] CereProc ltd: http://www.cereproc.com/
[2] www.nuance.com
[3] Fraunhofer SHORE™
http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst

6. References

[1] Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J-Y., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. In Proceedings of IVA.

[2] Bohus, D. & Horvitz, E. Facilitating multiparty dialog with gaze, gesture, and speech, in Proc. ICMI'10, (Beijing, China, 2010).

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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).

[8] Al Moubayed, S. Bringing the avatar to life. Studies and developments in facial communication for virtual agents and robots, Doctoral dissertation, (School of Computer Science, KTH Royal Institute of Technology, 2012), p75-78.

[9] Al Moubayed, S., & Skantze, G. (2012). Perception of Gaze Direction for Situated Interaction. In Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. The 14th ACM International Conference on Multimodal Interaction ICMI. Santa Monica, CA, USA.

[10] Al Moubayed, S., Skantze, G., & Beskow, J. (2012). Lip-reading Furhat: Audio Visual Intelligibility of a Back Projected Animated Face. In Proc. of the Intelligent Virtual Agents 10th International Conference (IVA 2012). Santa Cruz, CA, USA: Springer.

[11] Skantze, G., Al Moubayed, S., Gustafson, J., Beskow, J., & Granström, B. (2012). Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue. In Proceedings of IVA-RCVA. Santa Cruz, CA.

[12] 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.

[13] Mirnig, N., Weiss, A., Skantze, G., Al Moubayed, S., Gustafson, J., Beskow, J., Granström, B., & Tscheligi, M. (2013). Face-to-Face with a Robot: What do we actually talk about?. International Journal of Humanoid Robotics, 10(1).

[14] Al Moubayed, S., Beskow, J., Granström, B., Gustafson, J., Mirning, N., Skantze, G., & Tscheligi, M. (2012). Furhat goes to Robotville: a large-scale multiparty human-robot interaction data collection in a public space. InProc of LREC Workshop on Multimodal Corpora. Istanbul, Turkey.

[15] 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.