Towards Rich Multimodal Behavior in Spoken Dialogues with Embodied Agents

Samer Al Moubayed

Department for Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm, Sweden

E-mail: sameram@kth.se

Abstract— Spoken dialogue frameworks have traditionally been designed to handle a single stream of data – the speech signal. Research on human-human communication has been providing large evidence and quantifying the effects and the importance of a multitude of other multimodal nonverbal signals that people use in their communication, that shape and regulate their interaction. Driven by findings from multimodal human spoken interaction, and the advancements of capture devices and robotics and animation technologies, new possibilities are rising for the development of multimodal human-machine interaction that is more affective, social, and engaging. In such face-to-face interaction scenarios, dialogue systems can have a large set of signals at their disposal to infer context and enhance and regulate the interaction through the generation of verbal and nonverbal facial signals. This paper summarizes several design decisions, and experiments that we have followed in attempts to build rich and fluent multimodal interactive systems using a newly developed hybrid robotic head called Furhat, and discuss issues and challenges that this effort is facing.

Keywords—*Multimodal communication; Facial Synthesis; Dialogue Systems, Social Robots, Furhat Robot*

I. INTRODUCTION

Spoken interaction with machines does not only provide an interface to the computer that is hands-free, but having natural and fluent conversations with machines is powerful in applications that are social in nature. While perhaps looking at photos, and booking flights might be done using touch pads, screens and keyboards, other domains such as education and learning, collaboration, negotiation, persuasion, and commerce might be more successful through spoken engaging conversations that resemble more closely the forms of interaction humans are using face-to-face throughout their communication with other humans [1,2,3].

During spoken interaction, people use a large set of signals and forms of interaction that extend beyond the communication of information inside a stream of words. Spoken utterances are modulated and complemented with other paralinguistic non-verbal signals that enhance and even alter the meaning, and provide it with affective and pragmatic context, which in turn make a stream of words more successful and rich with conversational, emotional, and social function [4].

Looking at dialogue exchange from this perspective puts a new level of requirements for building spoken dialogue

systems that are made to serve social and affective functions. Such dialogue systems need to be supported by a high level of cognitive simulations that can understand and generate behaviours that are similar to those of naturally interacting humans. To be able to support a dialogue systems with synthesis models of different non-verbal signals (such as head, gaze, lip, and eyebrows movements), extensive quantification of the functions and effects of the spatio-temporal dynamics of these signals need to be acquired through studies on human-human and human-system interaction.

Multimodal dialogue systems typically have a visual representation (or embodiment) with which they can transfer information through bodily movements, and a set of devices that can monitor and interpret the interaction and the space around it in order to understand context and respond to it, hence aiming at making the system more responsive, aware, and adaptive to the human behaviour, rather than limiting the interaction to a very restricted space where the dialogue can be successful (limiting head movements, speaking through microphones, using push-to-talk recognition, etc.).

While building an artificial and embodied social companion has been a major one 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 that exhibit behaviour similar to that of a human, seems to be an unsolved problem, solutions to several of these technologies involved are becoming more available and robust, so that 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 indispensable tool to study patterns and applications of human-machine interaction that is not possible to investigate from human-human interaction data, that is in addition to providing a control environment where synthesised signals can be manipulated, and their effects can be measured on the interaction, something that can be

significantly more challenging and limiting in human-human interaction.

Recently, several researchers have taken the extra mile and built embodied conversational agents (ECAs) that are able to interact multi-modally with interlocutors while using rich contextual data from the interaction itself and from the environment surrounding the interaction (e.g. [6,7,8,9,10]), however, many design decisions have always been made to manoeuvre around the limitations of the different technologies (such as Automatic Speech Recognition through the use of keyboards for input) on the expense of restriction the naturalness of the interaction and as a result of that, increasing the gap from human-human dialogues. This design problem has always proven challenging and requires a sensitive balance between allowing the interaction to be accurate and meaningful, while making it fluent, engaging and natural so that users can exhibit affective signals similar to when they interact with humans, and would in turn respond to subtle nonverbal signals generated by the system in similar ways they would do when responding to these signals generated by a human.

One technology that is at the centre of this problem is the development of human-like embodied agents that are able to synthesize signals (such as gaze movements, lips movements, and facial gestures), that are natural enough, and perceived by users in accurate, easy, and fluent fashion. The development of these embodied agents has seen major progress in the last couple of decades [11,12]. This progress is also coupled with progress in humanoid robotics, creating physical replicas of these animated agents, both coming with their advantages and disadvantages.

II. THE FURHAT EMBODIED AGENT

We built Furhat to study and evaluate rich and multimodal models of situated spoken dialogue. Furhat [13] 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 [14], 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 [15,16], and as a result make spoken dialogue more fluent [17] 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 [18]. 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 [19,20,21], such as studying the effects of gaze movements in multiparty turn-taking [17,22], audio-visual

intelligibility of physically three dimensional avatars [23], and effects of head-pose on accuracy of addressee selection [21].

Although the face of Furhat is a light projection on a static mask, and this in principle might hinder speech intelligibility due to the lack of movement in the mask, in a study in [x], we found out in a speech intelligibility experiment, that looking at Furhat's lips not only does not hinder speech intelligibility, but rather boosts it. This might be due to the fact that Furhat is a situated agent (through physical embodiment), and this might require lower cognitive load (concentration) to read lips, thanks to its similarity to a typical face-to-face human-human interaction.

Fig. 1. Photographs of the Furhat backprojected robot head.

Fig. 2. Different possible types of facial synthesis signals that can be used by an animated talking head in a dialogue.

Furhat also provides a control environment where facial parameters can be manipulated. In a study on the effect of eyelids movements on gaze perception in [24], it was found that connecting the vertical movement of the eyelids to the movement of the eyeballs actually alters the perception of the direction of the eyes. The eyes are perceived to be looking to lower points in space when the eyelids are moved down to follow the iris, compared to when the eyelids are put in neutral position. Similar experiments are done to study the perception of the direction of the eyes as a result of the movements of the head [24]. These study show examples of how facial animation in Furhat can provide a test-bed to study multimodal signals in situated setups where human-human studies cannot be employed (due to the limitations in how controlled they can be designed to be), and neither using virtual characters due to their lack of co-spatiality since they are visualised on two dimensional screens.

Figure 1 shows two photographs of the system with different facial colors. Figure 2 shows some of the different low and high level signals the face can communicate.

III. MULTIPARTY DIALOGUE WITH FURHAT

Getting a conversational interface from the lab and into uncontrolled real-life environments is no easy ride. Capture technologies that appear to work well in a constrained and controlled laboratory are unpredictable and often fail to perform in more natural settings. To get around the limitations of the different technologies involved in the system and arrive at an operational interaction that can target useful research questions, many design tricks and solutions are often employed, and in many cases, while unavoidable, these solutions limit the degree of intuitiveness and naturalness of the systems. For example, to manage multiparty dialogue in the lab, we use Microsoft Kinect™’s depth camera to track the possible interlocutors with the system, and we use Kinect’s microphone array to capture hands-free directional speech.

In this section, we present an example dialogue system design and evaluation that was carried out at the London Science Museum as part of a robot exhibition, and initially defined very high level of requirements and drove the development of a new dialogue framework built for multimodal multiparty dialogue systems.

Museums offer a very unique forum to show novel and undergoing technologies. Museum setups have been a popular venue for multimodal dialogue systems in the past [6,7,8], this is due to that they attract visitors that are enthusiastic about developing technologies, they are forgiving in terms of the requirement on a stable functioning and task-oriented systems. On top of that, they also offer a sizable source for data collection of large number of naïve users. Aside from the benefits of museum setups, they also come with high requirements in terms of control over the environment. While in the lab, one can carry out highly controlled research investigations, with control over movement, equipment, noise, and demographics of users, a museum setup might suffer very high levels of noise, as well as often tens of simultaneous users and changing lighting conditions.

At the exhibition in London, we wanted to test a dialogue system with Furhat that can naturally interact with visitors,

which includes dialogue initiation by the system, and multiparty interaction to allow several users to simultaneously use the system. To solve the auditory and visual noise constraints, we opted to avoid the use visual input to the system and instead employed two ultrasound proximity sensors that were mounted on two fixed podiums. On the podiums, we placed two microphones, making it clear that for the system to work, the users need to stand by the podium and speak to Furhat into the microphone. Whenever a user approached the podium, the proximity sensor informed the dialogue manager of the presence of a user, and Furhat turned its head towards the podium and engaged the user in a dialogue.

However, the bottle-neck technology in spoken interactive setups is often speech recognition itself. A system that is supposed to interact with the public needs to have an ASR that is age-independent, gender-independent, speaker-independent, accent-independent, and noise robust for the “planned” interactive scenarios to take place. Under these requirements, most systems arrive at design solutions to increase the reliability of the speech recogniser. For example, Pixie [6] used a push-to-talk button to get input from users, Max [7] used written text as input and speech only as output, and Ada and Grace [8] opted to use a human agent. This agent had experience speaking to the system but also knew what questions are supposed to be addressed to Ada and Grace, and how to ask them. Through this agent, users could direct questions to the system.

Furhat in London needed to talk to multiple people at the same time, while utilising different verbal and nonverbal signals to regulate the interaction between the users and the system. This means that the above solutions are insufficient as many conversational phenomena had to be taken into account, such as interruptions, overlaps, multiparty turn-taking, etc. Indeed, one would expect that for such subtle social signals to even take place in a dialogue, a minimum level of fluency and naturalness needs to be present during the interaction.

To avoid the need for high recognition accuracy and to make the system robust against recognition errors without significantly limiting the fluency and spontaneity of the dialogue, we designed the dialogue to focus on collecting information from the users rather than merely to answer their questions. The dialogue relied on survey questions about the public perception of the future of robotics. The system asked the questions and used keyword spotting to extract certain information from the answers. Whenever successful, the system updated statistics charts of the answers on an LCD monitor that was mounted next to Furhat. The papers [25,26] explain further in details all the design “tricks” that were used in the system.

IV. THE IRISTK DIALOGUE AUTHORIZING PLATFORM

To orchestrate the whole system, the IrisTK dialogue platform was used [28]. IrisTK is XML based dialogue platform that was design for the quick prototyping and development of multimodal event-based dialogue systems. The framework is inspired by the notion of state-charts used, for example, in the UML modelling language. The state-chart model is an extension of the notion of finite-state machines (FSM), where the current state defines which effect events in

the system will have. However, whereas events in an FSM simply triggers a transition to another state, state charts may allow events to also result in actions taking place. Another notable difference is that the state chart paradigm allows states to be hierarchically structured, which means that the system may be in several states at the same time, thus defining generic event handlers on one level and more specific event handlers in the sub-state the system is currently in. Also, the transition between states can be conditioned, depending on global and local variables, as well as event parameters. This relieves state charts from the problem of state and transition explosion that traditional FSMs typically leads to, when modeling more complex dialogue systems.

IrisTK is based on modeling the interaction of events, encoded as XML messages between different modules (a module can be a face tracker, that transmits XML messages about the location of the face of a user). The design of module based systems is crucial in this system, and in other multimodal dialogue tasks. Such systems are multidisciplinary in nature, and researchers typically working on one of the technologies involved in such a system can be in isolation of the other technologies. This increases the need for the development of higher level, technology independent dialogue management that can allow for the communication of the different tools and programs involved (ASR, TTS, Face Tracking, Facial Animation, and Source Localisation). These programs can be run in different programs running on one or several machines.

IrisTK comes with several tools that support the communication of XML events in between programs, and over a network. This would allow the dialogue management to rely merely on XML events, while being able to be completely blind to the different programs that generate and consume these events. This also allows for the replacement of one or more program, or technology, without the need for any customization of the dialogue flow. Figure 3 shows example modules used in a multimodal multiparty dialogue system.

V. 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 avatar needs to implement output commands, an attention controller is used that, in principle, translates raw low level input data into abstract messages, such as translating 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 of the dialogue system itself.

Figure 4 shows a conceptual chart of the interpretation levels of signals and actions. In this, the controller would represent a dialogue manager and the interpreter and the generator would be the responsibility of an attention controller that manages the joint attention and the translation from sensor and actuator depend commands to general dialogue states.

Fig. 3. A set of possible modules involved in a multimodal dialogue system. All modules share the same communication protocols and transmit events through a network to each other. Top (blue): Sensory modules, mainly generating events. Middle (orange): Processing modules that translate sensory input to more abstract events, or to action events. Bottom (green): Actuating modules, executing events received from other modules. Actuating modules typically involve a robot and a conversational character (e.g. Furhat), and other peripherals such as an LCD screen.

Fig. 4. A simplified conceptual design scheme of different processes in a multimodal dialogue system.

VI. MAJOR FINDINGS

The exhibition at the London Science Museum enjoyed a very large visibility. Furhat was seen by almost 8.000 users during the 4 days of the robot festival. During the 4 days, the system interacted with hundreds of users. All interactions were audio-visually recorded, and all system events were time-logged.

During the exhibition, a selected set of users were chosen to fill out a questionnaire that focused on their experience with the system, targeting more specific aspects of the quality of the interaction and the system design.

A. Turn-taking strategies

From the analysis of the logged data, 92% of the questions were correctly picked up by the user Furhat was directing its gaze towards, showing very high accuracy of head orientation and gaze for the regulation of multiparty dialogue. Open questions that were addressed to both users (head directed towards the middle and gaze alternated between both users) also resulted in very high neutrality. 54% of these questions were picked up by the user that was previously speaking, showing no bias towards either of the users [28]

B. Multiparty engagement

When Furhat did not understand the user's input, it occasionally passed the question for *elaboration* to the other user saying something like "what do you think about that?" or "can you elaborate on that?", or it passed the question for *agreement* to the other user, saying something like "do you agree with that?". The analysis shows that these strategies were relatively successful in keeping the dialogue going and encouraging the subjects to provide more information to the system. However, repetitive elaboration requests exhibited the lowest efficiency. Only 40% of double-elaboration requests were answered by the users [28].

C. Corpus collection

The exhibition resulted in a multimodal corpus of about 10.000 utterances, out of which were 3200 question-answer pairs [x]. The data that was collected varied in terms of the age of the participants (estimates of age groups were made in [29]) and in gender.

The corpus now is being used for research on building models that are female and child friendly through the use of spectral transformation techniques. For more details, see [29].

D. User experience

During the exhibition, 86 questionnaires targeting the quality of the conversational capabilities and the interaction experience were collected. The questionnaire ranked different questions on a 5 point (1→5) Likert-type scale. 46 of the questionnaires were completed by male users and 39 by females. The results show that people experienced a very positive overall impression of the system. People liked Furhat (mean = 4.08, SD .76). They enjoyed talking to it (mean = 4.13, SD .84), and they liked its response and feedback behaviour (mean = 3.80, SD .71). They also rated the conversation to be rather easy and fluent (mean = 3.17, SD .99). Despite the limited understanding of the system, the score for "did Furhat understand what you said?" was 2.99 (SD 1.05). More details about the questionnaires study can be found in [30].

Following up on the London exhibition, several other presentations of Furhat to the public were carried out (as in [31,32]) where new developments on the sensor technologies and the robustness of the system were significantly enhanced using other technologies such as motion capture and wireless microphone arrays. Such public appearances are invaluable to the testing of the technology to first-time users in a free dynamic environment.

VII. DISCUSSIONS

One of the main characteristics of nonverbal behaviour that makes it very different from verbal behaviour is what is called the *continuity distinction* [4]. In human face-to-face communication, nonverbal communication is always present. As put it in [4] in a grammatically imperfect but rather meaningful way: "When you are in the presence of another human being, you cannot not communicate." For example, silence or a stare gaze in a conversation might signal meanings that are more powerful than words. This fact is highly problematic and complex when it comes to developing and testing full-fledged human-machine embodied interactive systems. Many studies concern the function and behaviour of different nonverbal signals in human communication in certain setups and under certain conditions. However, during interaction with users, the multitude of multimodal input variables humans take into account for shaping their behaviour is highly complex. The output behaviour in terms of nonverbal signals is also highly complex, and it contains several output signals that can be triggered in combinations. If multimodal systems ignore some of these signals or some of their functionalities (e.g. due to limitations in the technology and the state-of-knowledge on human communication), the effects of the multimodal signals the system generates might not reflect their intended functions, due to the large interplay among these interdependent signals (a clear example of this is *sarcasm* where facial expressions could reverse the meaning of the verbal message).

Taking into account all possible contextual and situated information during human-machine interaction is not possible yet. This is partly because of the limitations in capture and synthesis technologies and the difficulty in modelling certain contextual information such as personality, culture, background, situation, and so on. Until systems are able to model all the signals that take place during human-human interaction, ECAs developers need in principle to implement all the acquired knowledge from human-human communication in their systems when testing new signals, as testing these new signals independently will not guarantee their exact function when used in combination with other behaviours.

Verbal communication follows a more-or-less clear coding system (language). Although researchers have been trying to construct such a system for nonverbal behaviour (as in some of the work included in this thesis), some believe that a main property of nonverbal behaviour is its ambiguity and unpredictability, and constructing a deterministic coding system for nonverbal communication (i.e. a nonverbal language) strips it from its very important characteristics. This is frequently observed by us in real-life interactions between users and ECAs. As long as users start to observe regular patterns in the behaviour of the system, the system is perceived as "programmed" and their interest in it drops.

It is clear that nonverbal messages follow a meaning and a coding system that underlies when and how they are communicated, but perhaps the inherent unpredictability and ambiguity of nonverbal behaviour is what turns a dialogue from a strict information exchange process into a socially intelligent conversation.

ACKNOWLEDGEMENTS

The work outlined in this paper has been mainly carried out at the Department for Speech, Music and Hearing at KTH, and is a result of several collaborations. Special thanks and credits go to Gabriel Skantze, Jonas Beskow and Joakim Gustafson for their contributions. This work is partly funded by the KTH Strategic Research Areas project Embodied Multimodal Communication.

REFERENCES

- [1] B. Reeves, "The Benefits of Interactive Online Characters. Center for the Study of Language and Information," Stanford University, 2000.
- [2] N. Clifford, J. Steuer, and E. Tauber, "Computers are social actors." CHI '94: Proc. of the SIGCHI conference on Human factors in computing systems, ACM Press, pp. 72-78, 1994.
- [3] J. Cassel, J. Sullivan, S. Prevost, and E.E. Churchill, *Embodied Conversational Agents*, MIT Press, 2000.
- [4] V.P. Richmond, J.C. McCroskey, and M.L. Hickson III. "Nonverbal Behaviours in interpersonal relations." (Seventh Edition). Allyn and Bacon, 2011.
- [5] B. Granström, and D. House, "Modeling and evaluating verbal and non-verbal communication in talking animated interface agents," in Dybkjaer, I., Hensen, H., & Minker, W. (Eds.). *Evaluation of Text and Speech Systems* (pp. 65-98). Springer-Verlag Ltd, 2007.
- [6] J. Gustafson, J. Boye, M. Fredriksson, L. Johannesson, and J. Königsmann, "Providing computer game characters with conversational abilities," in *Proceedings of Intelligent Virtual Agent (IVA05)*. Kos, Greece, 2005.
- [7] S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth, "A conversational agent as museum guide - design and evaluation of a real-world application," in *Proceedings of IVA 2005*, Berlin: Springer-Verlag, 2005.
- [8] W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, et al. "Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides," in *10th International Conference on Intelligent Virtual Agents (IVA)*, 2010.
- [9] R. Gockley, A. Bruce, J. Forlizzi, M. Moichalowski, and A. Mundell, "Designing robots for long term social interaction," in *Proceedings of Intelligent Robots and Systems, IROS '05*. 1338-1343, 2005.
- [10] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *Proc. ICMI'10*, Beijing, China, 2010.
- [11] D. Massaro, *Perceiving talking faces: from speech perception to behavioural principles*, MIT Press, 1998.
- [12] P. Cohen, "The role of natural language in a multimodal interface," in *proc. of User Interface Software Technology (UIST '92) Conference*, Academic Press, Monterey, CA, pp. 143-149, 1992.
- [13] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction", in A. Esposito, et al. (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science* (pp. 114-130), Springer, 2012.
- [14] J. Beskow, "Rule-based visual speech synthesis," in *Proc of the Fourth European Conference on Speech Communication and Technology*, 1995.
- [15] S. Al Moubayed, J. Edlund, and J. Beskow, "Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections," *ACM Transactions on Interactive Intelligent Systems*, 1(2) 25, 2012.
- [16] J. Edlund, S. Al Moubayed, and J. Beskow, "Co-present or not? Embodiment, situatedness and the Mona Lisa gaze effect," in Nakano, Y., Conati, C., & Bader, T. (Eds.), *Eye Gaze in Intelligent User Interfaces - Gaze-based Analyses, Models, and Applications*. Springer, 2013.
- [17] S. Al Moubayed, and G. Skantze, "Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In *Proceedings of AVSP*. Florence, Italy, 2011.
- [18] S. Al Moubayed, J. Beskow, J. Edlund, B. Granström, and D. House, "Animated Faces for Robotic Heads: Gaze and Beyond," in Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., & Nijholt, A. (Eds.), *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues*, *Lecture Notes in Computer Science* (pp. 19-35). Springer, 2011.
- [19] S. Al Moubayed, G. Skantze, and J. Beskow, "The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction," *international Journal of Humanoid Robotics*, 10(1), 2013.
- [20] S. Al Moubayed, *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, p75-78, 2012.
- [21] S. Al Moubayed, J. Beskow, B. Bolleballi, A. H. Abdulaziz, et al. "Tutoring Robots - Multiparty multimodal social dialogue with an embodied tutor," in *proceedings of eNTERFACE2013*, Springer, in press.
- [22] J. Edlund, S. Al Moubayed, and J. Beskow, "The Mona Lisa Gaze effect as an objective metric for perceived cospatiality," in *Proc. of the Intelligent Virtual Agents 10th International Conference (IVA 2011)* (pp. 439-440). Reykjavík, Iceland: Springer, 2012.
- [23] S. Al Moubayed, G. Skantze, and J. Beskow, "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., 2012.
- [24] S. Al Moubayed, and G. Skantze, "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, 2012.
- [25] S. Al Moubayed, J. Beskow, B. Granström, J. Gustafson, N. Mirning, et al. "Furhat goes to Robotville: a large-scale multiparty human-robot interaction data collection in a public space," in *Proc of LREC Workshop on Multimodal Corpora*. Istanbul, Turkey, 2012.
- [26] S. Al Moubayed, J. Beskow, M. Blomberg, B. Granström, and J. Gustafson, et al. "Talking with Furhat - multi-party interaction with a back-projected robot head," in *Proceedings of Fonetik'12*. Gothenberg, Sweden, 2012.
- [27] G. Skantze and S. Al Moubayed, "IrisTK: a statechart-based toolkit for multi-party face-to-face interaction," in *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*, Santa Monica, CA, USA, 2012.
- [28] G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, and B. Granström, "Furhat at Robotville: A Robot head harvesting the thoughts of the public through multi-party dialogue," in *Proceedings of IVA-RCVA*. Santa Cruz, CA, 2012.
- [29] M. Blomberg, G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, and B. Granström. "Children and adults in dialogue with the robot head Furhat - corpus collection and initial analysis," in *Proceedings of WOCCI*. Portland, OR, 2012.
- [30] N. Mirnig, A. Weiss, G. Skantze, S. Al Moubayed, J. Gustafson, et al. "Face-to-Face with a Robot: What do we actually talk about?," *international Journal of Humanoid Robotics*, 10(1), 2013.
- [31] S. Al Moubayed, G. Skantze, J. Beskow, K. Stefanov, and J. Gustafson, "Multimodal Multiparty Social Interaction with the Furhat Head," in *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*. Santa Monica, CA, USA, 2012.
- [32] S. Al Moubayed, J. Beskow, and G. Skantze, "The Furhat Social Companion Talking Head," in *Interspeech 2013 - Show and Tell*. Lyon, France, 2013.