Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue

Gabriel Skantze
KTH Royal Institute of Technology
91 PUBLICATIONS **1,157** CITATIONS

SEE PROFILE

Joakim Gustafson
KTH Royal Institute of Technology
154 PUBLICATIONS **1,293** CITATIONS

SEE PROFILE

Jonas Beskow
KTH Royal Institute of Technology
186 PUBLICATIONS **2,378** CITATIONS

SEE PROFILE

Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue

Gabriel Skantze, Samer Al Moubayed,
Joakim Gustafson, Jonas Beskow, Björn Granström

Department of Speech Music and Hearing, KTH, Stockholm, Sweden

{gabriel,sameram,jocke,beskow,bjorn}@speech.kth.se

Abstract. This paper presents a large scale study in a public museum setting, where a back-projected robot head interacted with the visitors in multi-party dialogue. The exhibition was seen by thousands of visitors, resulting in a corpus of about 10.000 user utterances. The task of the system was to collect information on peoples' beliefs about the future of robots, in the form of a survey. The data analysis shows that the head and dialogue design allows the system to regulate the turn-taking behaviour, that it is indeed possible for a robot to effectively obtain information from the general public, and that this is facilitated by a multi-party setting.

1 Introduction

Spoken dialogue technology has mostly been applied to tasks where the system allows a human to seek information (such as restaurant recommendations) and/or perform some action (such as booking a ticket). However, we can also imagine situations where the system needs to obtain information from humans, and where spoken interaction would be a useful means for doing this. A typical case where this applies is robotics. Even if robots are able to learn from experience, sufficient information will not always be available in the environment to fill the knowledge gaps. Humans, however, are a rich source of information. If robots are equipped with the knowledge of how to obtain this information, it will give them a powerful means to improve their adaptability and cope with new situations as they arise. Another case where a system may gather information from humans is surveys. As noted by [1], surveys have not been a common application for dialog systems, despite the commercial potential.

This study is part of the IURO project[1], which aims at exploring how robots can be endowed with capabilities for obtaining missing information from humans through spoken interaction. The test scenario for the project is to build a robot that can autonomously navigate in a real urban environment, approach crowds of pedestrians, and enquire them for route directions. In December 2011, the IURO project was in-

[1] Interactive Urban Robot (www.iuro-project.eu)

vited to take part in the Robotville exhibition at the London Science Museum, showcasing some of the most advanced robots currently being developed in Europe. In order to explore how a robot may gather information from humans through multi-party dialogue, we put the interactive robot head Furhat [2], developed within the project, on display. During the four days of the exhibition, Furhat's task was to collect information on peoples' beliefs about the future of robots, in the form of a survey. The exhibition was seen by thousands of visitors, resulting in a corpus of about 10.000 user utterances. This setup allowed us to explore a number of issues in a challenging public environment. First, we wanted to explore to what extent it is possible to obtain information from humans without full understanding, and how this is affected by a multi-party setting. Second, we wanted to verify what we had previously found in controlled experimental settings: that the design of the robot head allows for accurate turn-taking in multi-party interaction. Third, we wanted to test a new control framework for multi-modal, multi-party interaction.

2 Motivation and related work

One thing that makes information gathering systems special is that it is up to the system to determine the value of the information that the user provides. For example, the route directions that the IURO robot will retrieve from human interlocutors are only possible means for accomplishing the task; there is no end in itself in following them or understand all the details. When faced with comprehension problems, the system may encourage the user to elaborate, and then possibly find more useful information in later turns. In a multi-party setting, the system could also turn to other humans to complement or verify the information gained. Studies on human-human dialogue with an error prone speech recognition channel have shown that humans that have a very clear goal of the interaction may accurately pick out pieces of information from the speech recognition results that are relevant to the task, and ask relevant task-related questions [3]. Similarly, in [4], a study is presented which shows how attentive feedback can be used in the call routing domain to encourage the user to provide more information, even if the system has a limited initial understanding of what the user says. For a survey application, the system could collect information systematically, without full understanding, which is later transcribed by human annotators. In order to do this, however, we need to understand what kinds of dialogue strategies may encourage people to provide information, and how they can be used without full understanding of what the user says. The setting of a public exhibition has allowed us to collect a large corpus of interactions, and thus to do quantitative analyses of how users react to an information gathering system.

There are several previous examples of multimodal dialog systems put to the test in public settings [5,6,7]. Allowing spoken interaction in a public setting is indeed a very challenging task – the system needs to cope with a lot of noise and crowds of people wanting to interact at the same time. To make the task feasible, different restrictions are often applied. One example is the virtual museum guide Max [6], which only allowed written input. Another example is the museum guides Ada and Grace [7], which did not allow the visitors to talk to the agents directly, but instead used a "handler" who spoke to the system, that is, a person who knew the limitations of the system and "translated" the visitors' questions. Also, in that system, the dialogue was very simplistic – basically a mapping of questions to answers independent of any dialog context. What makes the Furhat at Robotville exhibition special, apart from allowing the visitors to talk directly to the system, is that the visitors interacted with the system in a multi-party dialog, allowing several visitors to talk to the system at the same time. While there are examples of systems that have engaged in multi-party dialogue in more controlled settings, such as the virtual receptionist presented in [8], we are not aware of any other multi-party dialogue systems put to the test in a public setting, interacting with a large number of users.

3 Multimodal, multi-party interaction

3.1 Furhat: a back-projected robot head

The use of facial animation for interactive agents has been investigated over many years. However, when it comes to situated, multi-party interaction, the use of a flat screen with an animated head suffers from what is known as the Mona Lisa effect [9], since the agent is not spatially co-present with the user. This means that it is impossible to establish exclusive mutual gaze with one of the observers – either all observers will perceive the agent as looking at them, or no one will. While mechanical robot heads are indeed spatially co-present with the user, they are expensive to build, inflexible and potentially noisy. As part of the IURO project, we have developed a robot head called Furhat [2], as seen in Figure 1. Furhat can be regarded as a middle-ground between a mechanical robot head and animated agents. Using a micro projector the facial animation is projected on a three-dimensional mask that is a 3D printout of the head used in the animation software. The head is then mounted on a neck (a pan-tilt unit), which allows the use of both headpose and gaze to direct attention. We have previously shown in an experimental setting that such a 3D projection increases the system's ability to regulate the turn-taking in multi-party dialogue, as compared to a 2D screen [10]. The present study will explore the turn-taking accuracy in a real-life setting.

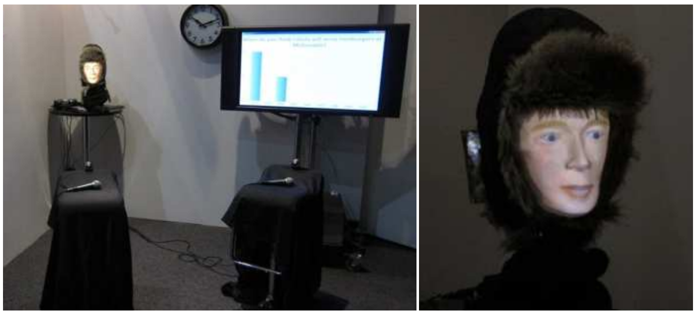

Fig. 1. The museum setup and a close-up of Furhat.

3.2 Technical setup in the museum

The setting of a public exhibition in a museum poses considerable challenges to a multimodal dialog system. In order to engage in a multi-party, situated interaction, the system not only needs to cope with the extremely noisy environment, but also be able to sense when visitors are present. For robustness reasons, we used two handheld close-range microphones put on podiums with short leads, forcing visitors to walk up to one of the microphones whenever they wanted to speak to Furhat. To sense whether someone was standing close to a microphone, ultrasound proximity sensors were mounted on the podiums. Furhat and the two podiums formed an equilateral triangle with sides of about 1.5 meter. In order to make the exhibition more interesting, a screen was mounted on the wall next to Furhat with charts showing the real-time results of the survey. The setup can be seen in Figure 1.

The multi-modal dialog system was implemented using a newly developed framework called *IrisTK* [11], based on a variant of *statecharts* [12] called *IrisFlow*. Statecharts is a powerful formalism for complex, reactive, event-driven systems, and lends itself well to visual representations. Statecharts is an extension of finite-state machines (FSM), but with several extensions. The most notable difference is that the statechart paradigm allows states to be hierarchically structured, which means that several states may be active at the same time, allowing the designer to define generic and specific event handlers on different levels. Also, the transition between states can be conditioned, depending on variables on different levels, as well as event parameters. This relieves statecharts from the problem of state and transition explosion that traditional FSMs typically leads to, when modelling more complex dialogue systems.

For speech synthesis, we used the CereVoice system developed by CereProc[2], lip-synchronizing it with the facial animation. 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. Each ASR engine also used two parallel language models, one context-free grammar with semantic tags (SRGS[3]), 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, similar to [13]. This allowed the system to recognize answers to very open questions and then pick out specific parts (such as a year) that could be used to update the survey charts.

Using the statechart framework, we defined generic states, such as *Idle* and *Dialog*, with sub-states to handle specific question types (e.g., *AskYNQuestion*, *AskYearQuestion*, *ReqHold*). The generic *Dialog* state then defined event handlers to handle questions from the user regardless of the current sub-state, allowing mixed-initiative interaction. Low-level sub-states were also defined, such as *Speaking*, *Attending* and *Listening*, with relevant event handlers, for example to handle situations where someone left while Furhat was speaking or listening. The statecharts also mapped specific events in the system to gestures in Furhat's face. For example, when the speech rec-

[2] http://www.cereproc.com/

[3] http://www.w3.org/TR/speech-grammar/

ognizer generated a start-of-speech event, Furhat raised the eyebrows, thereby signalling to the user that the system was listening. For a detailed description, see [11].

3.3 Multi-party survey dialogue

An example dialogue is shown in Table 1, which illustrates a number of typical interaction patterns. As soon as Furhat was approached by a visitor, Furhat immediately took the initiative and started to ask questions, as can be seen in turn 1-4. Depending on the current state in the statechart, the specific event handlers in that state listened for specific fragments in the ASR results. For example, in the *AskYearQuestion* state, the phrase "10 years" was considered as an answer to the question (as seen in turn 5). When the system actually understood an answer, it gave some relevant feedback (as in turn 6), but if it did not understand, it simply continued (as in turn 9 and 17). All answers were recorded and information about the corresponding questions was logged, which made it possible to annotate all answers later on. After each question, the system also made an *elaboration request* (as in turn 6 and 15). All utterances from the system (including questions) were randomly selected from a set of possible utterances, resulting in a varied output.

With two users present, Furhat could either ask a *directed question* – with the head posed in direction towards the addressee, and eyes looking forward (establishing eye-contact) – or an *open question* to both of them – with the head directed between the users, while alternating gaze between them (as in turn 17). Furhat then turned to the person who answered the question. When speech was detected in both microphones at the same time, the audio levels were compared in order to choose who to attend to. If a question was directed to one of the users and the other user tried to take the turn, Furhat would acknowledge this by shifting the gaze towards this user and say something like "could you please wait a second", while keeping the head directed towards the original user. Furhat would then shift the gaze back and continue with the interaction he was previously involved with (as in turn 11). Another option for Furhat in the multi-party setting was to follow up an elaboration request with an *agreement request*, by turning to the other interlocutor (as in turn 13). This could sometimes result in longer sequences on the same topic, such as the one shown in turn 9-16: question-elaboration-agreement-elaboration.

Although Furhat always started the interactions by asking questions from the survey, it was also possible for the visitors to ask questions (as in turn 18) and thereby shift the topic for the dialogue. This would trigger Furhat into awaiting further questions, resulting in mixed-initiative interaction patterns. In order to prepare the system for these questions, we used corpora from interactions with agents previously on display at museums [5,7], as well as from pre-tests in the lab. After answering some questions, or after too many non-understandings, Furhat again tried to get the initiative and ask questions from the survey.

Table 1. Example interaction. ASR results shown in brackets, using S: SRGS or D: Dictation grammar. Right column shows current state in the statechart.

1. U1	*(Approaching microphone 1)*	
2. F	*(Looking up at U1)* Hi there. Could you perhaps help me?	*Dialog/Initial*
3. U1	Yes [S: **yes**]	
4. F	I have some questions for you. When do you think robots will beat humans in football?	*Dialog/AskYearQuestion/ AskQuestion3*
5. U1	Maybe in 10 years [D: that see **in 10 years**]	
6. F	That soon! Could you elaborate on that?	*Dialog/ReqElaborate*
7. U1	Well, they are already pretty good at it [D: while they are already predicted owners]	
8. U2	*(Approaching microphone 2)*	
9. F	*(Still looking at U1)* Yeah… I have another question for you. Would you like robots to be like humans?	*Dialog/AskYNQuestion/ AskQuestion7*
10. U2	Yes, absolutely [D: **yes** that see]	
11. F	*(Turning to U2 with gaze only)* Could you just wait a second? *(Looking back at U1)* I'm sorry, where were we. Would you like robots to be like humans?	*Dialog/ReqHold* *Dialog/AskYNQuestion/ AskQuestion7*
12. U1	Sure [S: **sure**]	
13. F	Yeah? *(Turning the head towards User-2)* Do you agree?	*Dialog/ReqAgree*
14. U2	No [S: **no**]	
15. F	Mhm. Could you say something more about that?	*Dialog/ReqElaborate*
16. U2	Well, I want robots to be like robots [D: I want robots to be like Reynolds]	
17. F	Yeah… *(Posing head between users, shifting gaze back and forth between them)* I have another question for you. When do you think robots will serve hamburgers at McDonalds?	*Dialog/AskYNQuestion/ AskQuestion5*
18. U1	Do I need an umbrella? [S: **do I need an umbrella**]	
19. F	*(Turning to U1)* What do you think, you are in London! Maybe you have another question for me?	*Dialog/AnswerQuestion*

4 Data analysis

In this analysis, we will not focus so much on the contents of the visitors' answers, but rather on a more general question: to what extent is it possible for a robot to obtain information from the general public in the form of a multi-party survey? Over the four days of the exhibition, we collected about 10.000 utterances. From this corpus, we picked out all utterances that followed one of Furhat's questions (*Intitial* question, *Elaboration* request or *Agreement* request). This resulted in a total of 3200 question-answer pairs. We have no exact figures of the age and gender of the subjects involved in the analysis, but we collected 85 questionnaires over the whole exhibition from visitors who interacted with Furhat. The mean age of these subjects was 35.5 years (SD 16.2), ranging from 12 to 80 years. 46 of the respondents were male, 39 female. We also asked them to fill in a short questionnaire on their impression of the conversation, using a 5-point Likert scale. The participants' overall impression of the system was very positive. The results indicated that they liked Furhat (mean = 4.08, SD .76), they enjoyed talking to the robot (mean = 4.13, SD .84), and they liked Furhat's response behavior (mean = 3.80, SD .71). For a more detailed analysis, see [14].

4.1 Effect of questions on answers

All answers were then annotated into several categories. For the initial questions, we have in this analysis merged these into four categories: *AnswerYes*, *AnswerNo*, *AnswerOther* (any kind of answer which is not a simple yes or no) and *Decline* (any utterance which does not answer the question, such as "I have no idea", or a change of topic, as exemplified in turn 18 in Table 1). The answers to some of the initial questions are shown in the top chart in Figure 2.

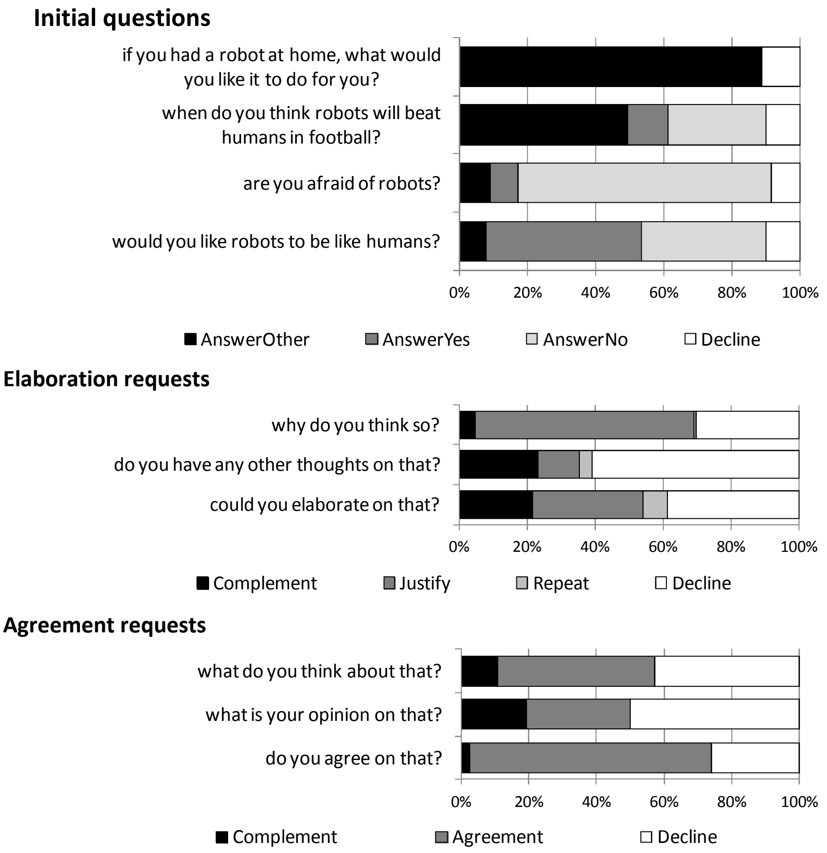

Fig. 2. Answer distributions to some of the questions.

As can be seen in the figure, the visitors seemed to be very cooperative when answering Furhat's initial questions, only 12% questions were declined. The answering pattern for these questions mostly follows the type of question posed. However, many visitors seem to have interpreted "when do you think…" as a yes/no question, probably because they didn’t hear the first word “when”. Content-wise, we can see that the question about whether robots should be human-like divides the visitors in half, and that most of them were not afraid of robots.

The general impression when annotating the answers was that they were very often very brief and without any justification. Therefore, it is interesting to look at the answers to the elaboration and agreement requests, as shown in Figure 2. Here we used the following categories: *Complement* the previous answer (e.g. provide more details and examples), *Justify* the previous answer (e.g. “because it is boring”), *Repeat* the previous answer, to show *Agreement* (including disagreement) to what the previous person said, and to *Decline* to answer. For these requests, the answer rate is not as high as for the initial questions. It is striking how much the actual wordings affect the answering rate (from about 40% to 70%) and the way people elaborate. This shows how important it is to carefully choose the wordings when designing an automatic survey. It is also interesting to see that these requests led to very few *Repeat* actions, which indicates that the elaborations indeed provided new information, but also that the users seem to have thought that their previous answer was understood, despite the very poor speech recognition performance for these kinds of open questions.

4.2 Turn-taking accuracy

Next, we wanted to know to what extent Furhat could regulate the turn-taking when there were two users present. As described above, Furhat could either ask a *directed question* to one of the participants, or an *open question* to both of them. In the first case, the addressee answered the question in 92.2% of the cases. The accuracy is similar regardless of whether Furhat was addressing the same speaker as in the turn before (92.6%), or if Furhat had just switched addressee (91.2%). For open questions, the addressee of the previous question answered the open questions in 54.4% of the cases, which indicates that they were indeed perceived as addressed to both participants. Another possible interpretation could be that the participants were just confused by Furhat’s open questions, resulting in a random behaviour. However, looking at answer type and response time, this is not very likely: only 15.5% of all open questions were declined (which is similar to the general distribution), and the mean response time for these were similar to the directed questions (1744 ms vs. 1868 ms).

Although the settings are not exactly the same, it is interesting to compare the turn-taking accuracy of 92.2% in a public setting to figures reported from more controlled experiments. In [8], an animated head on a 2D screen interacted with three users and gained an accuracy of 86.2%. In [10], a projected 3D head interacted with five users and gained an accuracy of 84% (and a response time of 1.38 s), while a 2D head only gained 50% accuracy (and a response time of 1.85 s).

4.3 Effect of multi-party involvement

Finally, we wanted to see how the involvement of another interlocutor in a multi-party setting affects the answer rate. Figure 3 shows the answer rate after the initial question, and for sequences of elaborate requests (when the same person is asked) and agreement requests (when the other person is asked). As the figure shows, the answer rate decreases after further elaborations with the same interlocutor (I-E-E), but increases when the other interlocutor is involved (I-E-A) (χ^2=5.34, dF=1, p<0.05). This indicates that it is indeed useful to involve other participants in order to gain more information on the same topic.

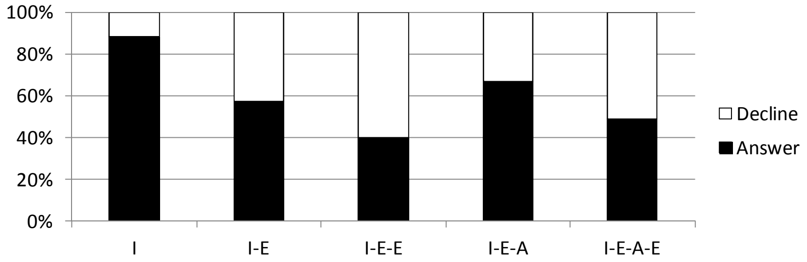

Fig. 3. The answer rate after (I)nitial question and sequences of (E)laborate and (A)greement requests.

5 Conclusions and Discussion

There are not many previous examples of large scale multi-party human-computer dialogue data collections done in public spaces. This real-world setting has confirmed what we have found in previous controlled experiments [10] – that the 3D design of Furhat allows for accurate turn-taking regulation. As an extension to these previous findings, we have also learned that it is possible to pose open questions to multiple participants, without confusion.

As the data analysis shows, it seems to be possible for a robot to effectively make humans provide information, despite a relatively poor speech recognition performance. People seemed to be willing to answer Furhat's questions, and to some extent elaborate on the topic. However, the actual wordings of such requests have a great impact on the answer rate and what kinds of answers are retrieved. The answer rate falls with further elaborations on the same topic and the same participant. This, however, might be mitigated by exploiting the multi-party setting and involve other participants on the topic.

One challenge that this kind of mixed-initiative survey dialogue system poses is how to distinguish answers to the system's questions from counter-questions (where the user ignores the system's question and claims the initiative). In the system presented here, we have relied on hand-coded phrase spotting, but that is not always sufficient, since it is often hard to tell whether a phrase is part of an answer or a question. We are currently looking into the use of machine-learning to distinguish answers from questions in the user's utterances, using features such as prosody and dialogue context.

6 Acknowledgements

This work is partly supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314, as well as the SAVIR project (Situated Audio-Visual Interaction with Robots) funded by the Swedish Government (strategic research areas). Also thanks to David Traum and Ron Artstein for providing transcriptions from [7].

7 References

[1] Stent, A., Stenchikova, S., & Marge, M. (2006). Dialog systems for surveys: The Rate-a-Course system. In *Proceedings of Spoken Language Technology Workshop, IEEE*.

[2] 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, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science*. Springer.

[3] Skantze, G. (2005). Exploring human error recovery strategies: implications for spoken dialogue systems. *Speech Communication, 45*(3), 325-341.

[4] Gustafson, J., Heldner, M., & Edlund, J. (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 240-251). Berlin/Heidelberg: Springer.

[5] Gustafson, J. (2002). *Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.

[6] Kopp, S., Gesellensetter, L., Krämer, N., & Wachsmuth, I. (2005). A conversational agent as museum guide - design and evaluation of a real-world application. In *Proceedings of IVA 2005, International Working Conference on Intelligent Virtual Agents*. Berlin: Springer-Verlag.

[7] 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 *Proccedings of IVA*.

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

[9] 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.

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

[11] Skantze, G., & Al Moubayed, S. (in press). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. To be published in *Proceedings of ICMI*. Santa Monica, CA.

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

[13] Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*. Norwich, UK.

[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. In *Proc of LREC Workshop on Multimodal Corpora*. Istanbul, Turkey.