Talking with Furhat – multi-party interaction with a back-projected robot head

Samer Al Moubayed, Jonas Beskow, Mats Blomberg, Björn Granström, Joakim Gustafson, Nicole Mirnig, Gabriel Skantze*
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
**HCI&Usability Unit ICT&S Center, University of Salzburg, Austria*

Abstract

This is a condensed presentation of some recent work on a back-projected robotic head for multi-party interaction in public settings. We will describe some of the design strategies and give some preliminary analysis of an interaction database collected at the Robotville exhibition at the London Science Museum

Introduction

This paper describes part of the IURO project, 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 invited 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, developed within the project, on display (Al Moubayed, Beskow, Skantze and Granström, 2012). 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 almost 8.000 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:

- 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.
- 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.
- Test a new control framework for multi-modal, multi-party interaction.
- Obtain a sizeable multimodal corpus of situated spontaneous interaction, for further analysis, including improved strategies for handling speech understanding for children
- Study user reactions to a conversational robotic head

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. 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. Earlier systems have typically used animated agents displayed on a screen, which the visitors interact with one-by-one. What makes the system presented in this paper special compared to these systems – apart from having a clear agenda of gathering information – is that the visitors interacted with the system in a multi-party dialogue, allowing several visitors to talk to the system at the same time.

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, 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, as seen in Figure 1[1]. Using a micro projector, KTH's state-of-the-art 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 mounted on a neck (a pan-tilt unit), which allows the use of both headpose and gaze to direct attention.

Technical setup in the museum

The setting of a public exhibition in a museum poses considerable challenges to a multimodal dialogue 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. In the lab, we have been using a Microsoft Kinect, which includes a depth camera for visual tracking of people approaching Furhat and an array microphone for capturing and localizing speech. However, in the crowded and noisy environment of the museum, a Kinect would not suffice. Instead, 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, we mounted ultrasound proximity sensors on the podiums. Furhat and the two podiums formed an equilateral triangle with sides of about 1.5 meter. On the wall next to Furhat, a screen was mounted with charts showing the real-time results of the survey. The purpose of this was to make the whole exhibition more interesting for the visitors. The setup can be seen in Figure 1.

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. 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.

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

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. The example also illustrates how the system was able to extract partial results from the ASR. 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

[1] For videos of Furhat, see http://www.speech.kth.se/furhat

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

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

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).

1. U1	*(Approaching microphone 1)*
2. F	*(Looking up at U1)* Hi there. Could you perhaps help me?
3. U1	Yes [S: **yes**]
4. F	I have some questions for you. When do you think robots will beat humans in football?
5. U1	Maybe in 10 years [D: that see **in 10 years**]
6. F	That soon! Could you elaborate on that?
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?
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?
12. U1	Sure [S: **sure**]
13. F	Yeah? *(Turning the head towards User-2)* Do you agree?
14. U2	No [S: **no**]
15. F	Mhm. Could you say something more about that?
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?
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?

Table 1: *Example interaction. ASR results shown in brackets, using S: SRGS or D: Dictation grammar.*

Data collection and analysis

Over the four days of the exhibition, we collected about 10.000 utterances, We will use this database to investigate improvements to the ASR component, but we can also use it to study 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? From the corpus, we picked out all utterances that followed one of Furhat's questions (*Initial* question, *Elaboration* request or *Agreement* request). This resulted in a total of 3200 question-answer pairs.

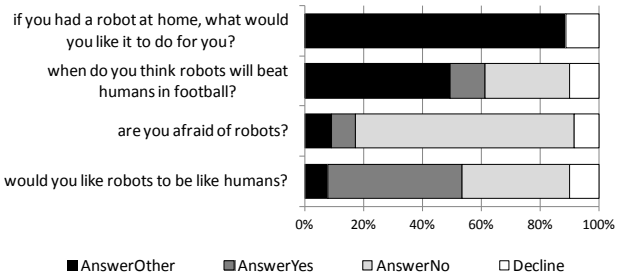

Figure 2: *Examples of answer distributions to some of the initial questions.*

User reactions

86 visitors who actively interacted with the robotic head were asked to fill in a short questionnaire on their impression of the conversation with Furhat by ranking the system on a number of parameters on a 5-point Likert scale, that ranged from 1 "not at all" to 5 - "very much". The mean age of these visitors was 35 years, ranging from 12 to 80 years. The participants' overall impression of the system was very positive: they liked Furhat a lot (mean = 4.08, SD .76), they enjoyed talking to the robot (mean = 4.13, .84), and they liked Furhat's response behaviour (mean = 3.80, SD .71). All questions got mean ratings above 2.5, and questions such as "*How much do you like Furhat?*" and "*Did you enjoy talking to Furhat?*" received scores in excess of 4.

Recognition of children's speech

A substantial part of the visitors who made a conversation with Furhat during the exhibition were children and adolescents. This is expected also to be the case in the intended scenario of the IURO project. For this reason, it is important that the speech recognition component functions well also for this speaker category.

However, recognition of children's speech is known to be a significantly harder problem than that of adult speech. Several factors contribute to this effect. Most importantly, the vocal tract size and proportions differ from those of adults and there is high intra- and inter-speaker pronunciation variability caused by the developing articulation skills, vocabulary size and grammatical usage as a function of age. It has been shown that the recognition word error rate of a system which has been trained on adult speech is increased by an order of magnitude for young children compared with the rate of adults (Elenius and Blomberg, 2005).

An additional limitation is the lack of large children's speech corpora for training the systems. For this reason, speaker independent systems are mainly trained on adult speech. The recognition component used by Furhat, belongs to this category and the exhibition utterances displays low accuracy for young speakers. It is obvious that this problem needs to be addressed by Furhat to reduce the error rate for young speakers in many applications.

The normal technique to raise the recognition accuracy for an unknown speaker is speaker adaptation, in which a limited number of utterances from the speaker are used to move the trained acoustic models towards the position of the target speaker in the acoustic space. Another possibility is feature normalization, where the extracted acoustic features of the utterance to be recognized are transformed in order to remove as much as possible of speaker and environment characteristics from the utterance.

Unfortunately, neither of these techniques is applicable, since we don't have access to these functions of the Furhat black-box recognizer. The only input we can modify is the microphone signal.

A procedure which does not require detailed knowledge of the internal structure of the recognizer is speech conversion. It has been shown that such techniques can improve the accuracy (Sjölander and Gustafsson, 2000). In that work, the frequency scale of the input child speech signal was compressed by about 20%, which reduced the error rate significantly. We will extend their work with speaker-specific frequency scaling. The scale factor can be estimated in different ways: by maximization of the likelihood of an external phoneme recognition module, by tracked fundamental frequency, or by using visual information from the video recordings.

Conclusions

There are not many previous examples of large scale multi-party human-computer dialogue data collections done in public spaces. Despite this challenging environment, the system proved to be very robust during the four days of the exhibition. This real-world setting has confirmed that the 3D design of Furhat allows for accurate turn-taking regulation. 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.

Acknowledgements

This work is partly supported by the European Commission project IURO, grant agreement no. 248314.

References

Al Moubayed, S., Beskow, J., Skantze, G. and Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito, A. et al. (Eds) Cognitive behavioural systems Lecture Notes in Computer Science Springer

Gustafson, J. & Sjölander, K. (2000). Voice transformations for improving children's speech recognition in a publicly available dialogue system. In Proceedings of ICSLP, Denver, Colorado.

Elenius, D., & Blomberg, M. (2005). Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In Proceedings of Interspeech. Lisbon, Portugal.