Children and adults in dialogue with the robot head Furhat - corpus collection and initial analysis

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

Department of Speech Music and Hearing, KTH, Stockholm, Sweden
{matsb,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 almost 8000 visitors, out of which several thousand interacted with the system. A considerable portion of the visitors were children from around 4 years of age and adolescents. The collected corpus consists of about 10.000 user utterances. The head and a multi-party dialogue design allow the system to regulate the turn-taking behaviour, and help the robot to effectively obtain information from the general public. The commercial speech recognition component, supposedly designed for adult speakers, had considerably lower accuracy for the children. Methods are proposed for improving the performance for that speaker category.

Index Terms: multi-party dialog, human-robot interaction, children's speech

1. Introduction

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, e.g., 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 [1], 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 almost 8.000 visitors and several thousands of them interacted with the robot. This resulted in a corpus of about 10.000 user utterances.

The 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 multimodal, multi-party interaction.

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 examples of multimodal dialogue systems put to the test in public settings [2,3,4]. Such 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. This makes the setup similar to [5], with one difference being that in the setting presented here, the visitors interacted with a physical robot head instead of an agent on a flat screen.

We expect that a substantial number of the humans interacting with Furhat in the intended applications will be children. The assumption is supported by the large proportion of children that visited the Furhat booth during the exhibition. This necessitates studies on how well the system functions for this user category, and work to improve its performance. The collected speech data will be very useful for this purpose, as well as for future research in automatic recognition of children's spontaneous speech.

2. Multimodal, multi-party interaction

2.1. 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 [6], 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 [1], as seen in Figure 1[2]. Furhat can be regarded as the middle ground between a mechanical robot head and animated agents. 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 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,

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

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

as compared to a 2D screen [7]. The present study will explore the turn-taking accuracy in a real-life setting.

2.2. 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. 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, as seen in Figure 1. 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.

Figure 1: *A close-up of Furhat and pictures from the Robotville exhibition at the London Science Museum.*

The multi-modal dialog system was implemented using a newly developed framework based on the notion of *statecharts* [8]. Statecharts avoid the problem of state and transition explosion that traditional FSMs typically lead to when modelling more complex dialogue systems.

For speech synthesis, we used the CereVoice system developed by CereProc[3], 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[4]), 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.

2.3. 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. 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 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. After answering some questions, or after too many non-understandings, Furhat again tried to get the initiative and ask questions from the survey.

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

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

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

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?

3. Collected data

Over the four days of the exhibition, we collected about 10.000 utterances. The utterances were classified into five categories according to their signal quality and their deviation from the ideal single-speaker, grammatically correct sentence. For speech recognition experiments, we have initially selected the subset with the highest quality, consisting of 5899 utterances. This "proper" subset excludes those with recording errors (e.g., missing beginning or end of the utterance, peak-clipped signal and too low signal amplitude typically caused by not addressing the microphone) and deviating speech (laughters, mispronunciations, non-intelligibility, overlapping speakers, talking to other visitors, etc.). It should also be noted that there was considerable background "babble" noise even in the highest quality subset. In this set, 3599 utterances were spoken by visitors to the exhibition and the remaining 2300 ones were spoken by the 5 male members of the Furhat exhibition staff.

3.1. Speaker categories

The recorded utterances have been classified into speaker categories by means of listening. The categories are: adult male visitor, adult female visitor, child and member of the exhibition staff. For each utterance judged to be spoken by a child, the age of that speaker was roughly estimated into one-year age groups. The estimation accuracy was obviously not error-free but may still be of value for comparison with acoustic vocal tract length normalization methods. Improved classification is expected by the planned use of visual information in the video recordings. This will also enable measurement of the number of utterances for each speaker. The estimated speaker categories and child age distributions are presented in Tables 2 and 3, respectively.

Table 2: *Estimated speaker category distribution and average utterance length.*

	Male adults	Female adults	Children
#Utterances	1122	837	1640
#Words / utterance	3.82	4.35	2.80

Table 3: *Estimated speaker age distribution in the utterances judged to be spoken by a child.*

Age	- 6	7 - 9	10 - 12	13 - 15
#Utterances	215	609	564	252

4. Reply rate and turn-taking accuracy

From the full corpus, we picked out all visitor 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. All subject utterances immediately after a question were annotated into several categories. We have in this analysis merged these into two categories: *Answer* (any kind of answer) 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 individual *Answer* rates for children and adults to each of these types of questions are presented in Figure 2.

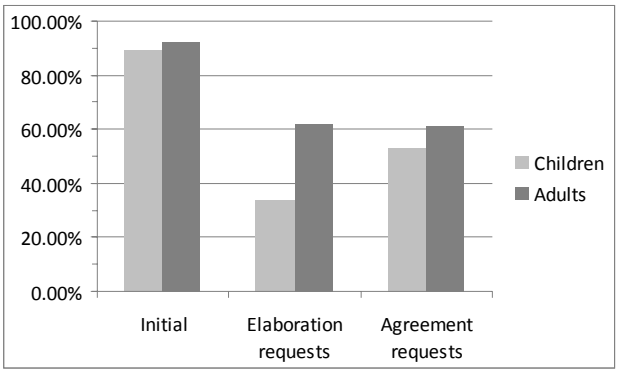

Figure 2: *Answer rates to Initial, Elaboration and Agreement questions from Furhat for children and adults.*

The two subject groups had high *Answer* rates, around 90%, to the *Initial* questions. The rate dropped for the two other question types with a larger decrease for children. A particularly low rate was achieved for children's replies to *Elaboration* requests. This indicates that the dialog needs to be individually designed for this speaker category.

There is some difference between adults and children in the response to Furhat's choice of question direction between the two speakers. Whereas the addressed adult speaker replied in around 96% of the questions, the corresponding number for children was a bit lower, around 92%. This may be caused by a possibly somewhat different turn-taking strategy and maybe also a lower adherence to such rules.

5. Children's speech recognition

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 application 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. Without compensation for these differences, the recognition error rate of systems, which have been trained on adult speech, is significantly increased for young children compared with the rate of adults [9,10].

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 this problem needs to be addressed to reduce the error rate for young speakers. Initial recognition experiments on the exhibition recordings without vocal tract normalization confirm the higher error rates for the children compared to the adult speakers.

Normal techniques to raise the recognition accuracy for an unknown speaker are speaker adaptation and feature normalization. Unfortunately, neither of these techniques is applicable, since we don't have access to the extracted features or the acoustic models of the Furhat black-box recognizer. The only input we can modify is the microphone signal.

A procedure which does not require knowledge of the internal structure of the recognizer is speech conversion. It has been shown that such techniques can improve the accuracy for children's speech [11]. In that work, the frequency scale of the input child speech signal was compressed using Phase Vocoder [12] and TD-PSOLA [13] techniques. A compression factor around 20% reduced the error rate significantly. We will extend this approach with speaker-specific frequency scaling. The compression factor can be approximately estimated in different ways: by maximization of the likelihood of an external phoneme or Gaussian Mixture Model (GMM) recognition component [14], by tracked fundamental frequency, or by using visual information from the video recordings.

Straightforward linear frequency compression results in a bandwidth reduction of the transformed speech signal. If the resulting bandwidth is lower than that used by the recognizer, this will lead to recognition errors. The work in [11] was not exposed to this problem, since the speech recognizer had lower bandwidth (4 kHz) than that of the original microphone signal (8 kHz). In the current exhibition recordings, the recorded signal has the same bandwidth as the recognizer and the problem of reduced bandwidth needs to be addressed.

We have approached the problem by implementing the standard 2-segment piece-wise linear vocal tract length normalization (VTLN) algorithm [15] in the phase vocoder. In this frequency warping function, the high end frequency of the warped spectrum is not changed and there is no bandwidth reduction.

The problem can also be avoided in future data collection by increasing the sampling frequency and, accordingly, the bandwidth of the microphone signal. For example, if the bandwidth is raised to 12 kHz, this will allow linear compression by a factor 0.67 without running into the bandwidth reduction problem in a 8 kHz bandwidth recognizer.

6. Conclusions

There are not many previous examples of large scale multi-party human-computer dialogue data collections done in public spaces, including a large proportion of children. Despite the challenging environment, the system proved to be very robust during the four days of the exhibition. This real-world setting has confirmed what we have found in previous controlled experiments [7] – 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. Further work will study if special dialog design is required for children.

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. Exploiting the multi-party setting and involving other participants on the topic are judged as important for acquiring the requested information.

Initial baseline experiments have confirmed the significantly lower recognition accuracy for children's speech. Further work will be performed to investigate the normalization techniques proposed in the paper.

7. Acknowledgements

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

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

[2] 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.

[3] 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.

[4] 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 10th International *Conference on* Intelligent *Virtual Agents (IVA).*

[5] Bohus, D., & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc ICMI'10. Beijing*, China.
[6] 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.
[7] 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.
[8] Harel, D. (1987). Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, 8, *231-274.*
[9] Gerosa, M., Giuliani, D., Narayanan, S. & Potamianos, A. (2009). A Review of ASR Technologies for Children' s Speech. In *Proceedings of WOCCI*, Cambridge, MA.
[10] 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.
[11] 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.
[12] Wegmann, S., McAllaster, D., Orlo, J., & Peskin, B. (1996). Speaker normalization on conversational telephone speech. In *Proceedings of ICASSP*, Atlanta, GA. pp 339-342.
[13] Dolson, M. (1986). "The phase vocoder: A tutorial" *Computer Music Journal*, vol. 10, no. 4, pp. 14-27.
[14] Moulines, E. & Charpentier, F. (1990). "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communication* Vol. 9 (5/6), pp. 453-467.
[15] Stemmer, G., Brugnara, F. and Giuliani, D. (2005). Adaptive Training Using Simple Target Models. In *Proceedings of ICASSP*, Philadelphia, PA, pp. I–997–1000.