Furhat goes to Robotville: A large-scale multiparty human-robot interaction data collection in a public space

Samer Al Moubayed, Jonas Beskow, Björn Granström, Joakim Gustafson, Nicole Mirnig*, Gabriel Skantze, Manfred Tscheligi*

KTH Speech, Music and Hearing, Stockholm, Sweden, *HCI&Usability Unit ICT&S Center, University of Salzburg, Austria

E-mail: {sameram|beskow|bjorn|jocke|gabriel}@speech.kth.se, {nicole.mirnig|manfred.tscheligi}@sbg.ac.at

Abstract

In the four days of the Robotville exhibition at the London Science Museum, UK, during which the back-projected head Furhat in a situated spoken dialogue system was seen by almost 8 000 visitors, we collected a database of 10 000 utterances spoken to Furhat in situated interaction. The data collection is an example of a particular kind of corpus collection of human-machine dialogues in public spaces that has several interesting and specific characteristics, both with respect to the technical details of the collection and with respect to the resulting corpus contents. In this paper, we take the Furhat data collection as a starting point for a discussion of the motives for this type of data collection, its technical peculiarities and prerequisites, and the characteristics of the resulting corpus.

1. Introduction

In December 2011, a spoken dialogue system featuring the back-projected physical talking head Furhat (Al Moubayed et al., 2012) was on display at the Robotville exhibition at the London Science Museum. During the four days of the exhibition, Furhat was seen by almost 8000 museum visitors, including many children, which took the opportunity to chat with the system. All in all, the system collected 10000 utterances of unrehearsed, unscripted interaction. The Furhat data collection in London is an example of a type of data collection where the main effort is spent capturing *large-scale* corpora of *situated* human-machine interactions that take place in *authentic public environments*. In order to achieve this, sacrifices must be made on many levels.

This paper discusses the motivation for capturing this type of corpus, its merits, and the necessary trade-offs in data collections like Furhat at Robotville.

2. Background and related work

Although collection of large-scale situated data in public spaces is a cumbersome task, several successful attempts have been made.

The multimodal spoken dialogue system August (Gustafson and Bell, 2000) was used to collect spoken data for more than half a year in 1998 at the Culture Centre in Stockholm, Sweden. August could answer questions about for example restaurants in Stockholm or about his namesake, the author August Strindberg. More than 10000 utterances were collected from 2500 visitors.

Pixie (Gustafson and Sjölander, 2002) collected data from museum visitors, starting in 2002 and lasting for more than two years. Pixie was part of the futuristic exhibition "Tänk om"(What if), which consisted of a full-scale future apartment, in which Pixie appeared as an assistant and an example of an embodied speech interface. Pixie was introduced to the visitors in a movie portraying a future family living in the apartment. Next, the visitors were allowed to enter that same apartment, in which they interacted with Pixie in a computer game setting, helping her perform tasks in the apartment, such as changing the lighting in the apartment. The visitors were also encouraged to ask Pixie general questions about herself or the exhibition. The resulting corpus contains about 100 000 utterances.

In 2004, the life-sized multimodal dialogue system Max was displayed for several years in the Heinz Nixdorf Museums Forum, a public computer museum in Paderborn, Germany (Kopp et al., 2005). Max took written language as input and responded with synthesized speech. In its first seven weeks at the museum, Max recorded over 50000 inputs.

Finally, Ada and Grace, two multimodal spoken dialogue system designed as twins first greeted the visitors to the Museum of Science, in Boston, US in December 2009 (Swartout et al., 2010). The twins acted as museum guides, and spook both to each other and to visitors and human guides. In early 2010, the twins collected over 6000 utterances in a little over a month.

Figure 1. The previous large-scale data collection systems: From left to right: August, Pixie, Max, Ada and Grace.

3. Motivation

Just about all development in speech technology relies heavily on data these days, and the type of data we analyse and base our models on will be reflected strongly in our results and in the behaviour of our systems. When we gather human-machine interaction data, we would ideally like it to be as realistic as possible: real users with real systems in real settings performing real tasks. And we want large quantities of data as well - the more the better. In reality, this set of requirements is unrealistic, and sacrifices have to be made one place or another. In the type of data collection discussed here, the requirements that lay firm are that the dialogues be situated - that they take place in a real, public setting with real people - and that they be sizeable, capturing large quantities of data. As is the case with Wizard-of-Oz data collections, where the system is partially or wholly replaced by a human (the "Wizard"), these data collections are in a sense a window onto the future - they reveal what will happen when we have systems that can handle what our current systems cannot (such as exceedingly noisy environments or multiple speakers with diverse goals).

4. Technical considerations

Wizard-of-Oz collections are often not feasible in these settings. In order to get large quantities of data, the systems must run full-time over extended periods of time, and having a Wizard work all hours is simply too expensive. Instead, these systems work by employing every available trick to make their interlocutors feel at home and to make them continue speaking for as long as possible. The following examples from the systems cited in the background are by no means exhaustive, but serve to illustrate that spoken dialogue designers utilize a wide range of tricks.

In August, thought balloons illustrating topics the system could talk about appeared above the character's head at regular intervals, in an attempt to unobtrusively suggest what visitors might say. Another, trick was to place the push-to-talk button such that speakers had to lean in close to the directed microphone to reach it. The system also made use of a video-based person detection system to simulate visual awareness, which was used to encourage approaching users to strike up a conversation.

In the Pixie system the visitors had to register before entering the exhibition, they were then issued RFID tags. Pixie was able to appear at different places in the futuristic home, but in order for her to show up, visitors had to insert their card. This allowed the developers to track the identity, gender and age of each interlocutor, as well as keeping track of their location and progression in the game. The information about the age made it possible to transform children's utterances, before sending them to a commercial speech recognizer, thus improving its performance (Gustafson and Sjölander, 2002).

In the case of Max, the most obvious trick is the use of text input rather than speech. Max also made use of face detection in order to detect users to interact with. The system also simulated its emotional state, making it possible for it to appear aware of its own performance.

The twins Ada and Grace use an entire battery of sophisticated tricks to appear more able than they otherwise would. One of the simpler is to present visitors with a list of things to ask. Another is that the dialogue with them is often human mediated - visitors will tell a guide what they want to ask, and the guide - who has had experience with addressing the twins - rephrases the questions into the microphone. Another trick is that the twins talk between themselves. As they both know exactly what the other is saying, they can often insert clever and timely remarks, which give an impression of robustness and perhaps even intelligence.

Again, these examples serve merely as an illustration of techniques to keep visitors in high spirits, which is essential for getting at the futuristic and currently otherwise unavailable data we aim at in making these data collections. Clearly, tricks are used in other circumstances as well, but to date, they are essential for the large-scale data collection in public spaces.

5. The technology behind Furhat

The robot head called Furhat, uses KTH's state-of-the-art facial animation system. Using a micro projector the animated model is projected, on a three-dimensional mask that is a 3D printout of the head used in the animation software. The back-projection technique has also allowed us to mount the head on a neck (a pan-tilt unit). The mask has been painted with back-projection paint in order to improve visibility of the projection, which makes it possible to use the Furhat head under normal light conditions. Using software-based facial animation in a robot head allows for a flexible generation of advanced facial signals that are crucial for dialogue applications. It also provides the robot with real-time lip-synchronized speech, something which has been shown to increase speech intelligibility in noisy environments (for details on why and how Furhat was built, please refer to Al Moubayed et al 2012). The lip synchronized synthesized speech also lends a sense of authenticity to the head. The laboratory version the system, which was designed to handle two human interlocutors simultaneously to make experiments with the realistic gaze provided by the back-projected talking head, used a Microsoft Kinect[1], which includes a depth camera for visual tracking of people approaching Furhat and an array microphone for capturing speech. In the public space version, these technologies are niceties that, given the current state-of-the-art, must be sacrificed for the sake of simply getting-it-to-work. For speech recognition, the Microsoft Speech API was used and for speech synthesis the William voice from CereProc[2]. CereProc's TTS reports the timing of the phonemes in the synthesized utterance, which was used for lip-synchronization. The voice also contains non-verbal tokens like grunts and laughter that were used to give Furhat a more human-like appearance.

[1] http://kinectforwindows.org/
[2] http://www.cereproc.com/

Figure 2. Pictures from Furhat at Robotville.

To orchestrate the whole system, a state-chart model was used. The framework is inspired by the notion of state-charts, developed by Harel (1987) and used 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 modelling more complex dialogue systems. For the exhibition scenario, the dialogue contained two major states reflecting different initiatives: one where Furhat had the initiative and asked questions to the visitors (“*when do you think robots will beat humans in football?”)* and one where the visitors asked questions to Furhat (“*where do you come from?*”). In the former case, Furhat continued the dialogue (“*why do you think so*?”), even though he often understood very little of the actual answers, occasionally extracting important keywords. To exploit the possibilities of facial gestures that the back-projection technique allows, certain sensory events were mapped to gesture actions in the state chart. For example, when the speech recognizer detected a start of speech, the eyebrows were raised to signal that Furhat was paying attention.

6. Robotville tricks

In the crowded and noisy environment of the museum, with often tens of simultaneous onlookers, a Kinect will not work. In order to cope with this, we used handheld close-range microphones with short leads, forcing visitors to walk up to one of the microphones whenever they wanted to speak to Furhat. Close to each microphone we mounted ultrasound proximity sensors, so the system would know at all times whether someone was holding a microphone. In this way, the methods described below could be used even though the sensor technology with which they were developed could not. The most striking feature of Furhat - his very clear gaze - was utilized to the greatest extent. The setup with one dialogue system addressing two humans was exploited in these ways:

- When nobody was present at a microphone, Furhat would look down, only to look up at each new interlocutor with a greeting as they arrived.
- Newcomers who barged in on one microphone while Furhat was already speaking with someone on the other would face a brief glance and a quick request to wait for their turn.
- When two interlocutors were involved in the same conversation with Furhat, Furhat would deflect some of the utterances he did not understand to the other interlocutor: "What do *you* think about that?"
- Furhat could pose open question to both visitors by directing the head straight in the middle then alternately seeking mutual gaze with the two visitors. By comparing the microphone levels, Furhat could then choose who to attend to and follow-up on.

Other tricks included maintaining a fairly strict control over the dialogue. The main goal of the data collection was to learn more about what happens when a system attempt to gather data - more specifically, directions - from people in public places. The dialogue type - to collect information - was kept, but the information asked for was changed to better fit the museum setting. When the system did not understand a response, it would not ask the visitor for a repetition or otherwise admit that it did not understand. Instead it would either respond with "yeah" with positive or negative prosody, followed by “can you elaborate on that” or ask the other visitor to comment on that response. In order to prepare the system for initiatives from the visitors, open questions from users of August, Pixie and the twins Ada and Grace were introduced in the language model and responses to them implemented. Both the system's ability to tell jokes and to sometimes answer with a hint of sarcasm was noted by visitors, who seemed to take it as a sign of "intelligence". Another trick that made children significantly more engaged was the possibility to tell Furhat to change his appearance (colours of his face, lips and eyebrows).

As a final trick, the developers on-site would sometimes take one of the microphones and take part in the dialogue. By doing this, they suggested to spectators what one might successfully say to the system, while they at the same time got the three-party dialogue going. In most cases, the resulting dialogue would be more successful also for the visitor speaking to the system at that time. This data, with an impromptu three-party dialogue between the system, a developer, and a visitor is interesting, since it shows how naïve users can unobtrusively be guided through a dialogue.

7. Robotville results

In four days, the Furhat exhibition collected around 10000 utterances - more than eight hours worth of speech and video - from people that spoke to Furhat in the presence of tens of other visitors - about 8000 all in all. The data is currently being analyzed. The wide press coverage Furhat received often describes the system as "witty", "sarcastic" and "intelligent", statements that bear evidence of the effectiveness of the tricks exploited in the system, since the extremely noisy environment and the sheer amount of visitors resulted in the system only rarely understood what was being said.

We also wanted to get an impression of the perceived quality of the conversational abilities of the system used in the corpus collection. Since thousands of visitors interacted with Furhat it was possible for us to tap into their minds. We selected 86 visitors who actively interacted with the robotic head, and asked them to fill in a short questionnaire on their impression of the conversation and their rating of the feedback and the robot's performance 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. 46 of the respondents were male, 39 female (one participant did not fill in the demographic data section of the questionnaire). The participants rated their general interest in technology on a mean of 4.42 (SD .798) and their interest in robots on a mean of 4.28 (SD .954), which might be an indicator for a generally higher tolerance towards technical systems. 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). Even if the participants stated that they had to concentrate quite a bit to talk to Furhat (mean = 3.38, SD 1.16), they nevertheless could very well understand what Furhat said (mean = 4.25, SD .89) and they rated the conversation with it as rather easy (mean = 3.17, SD .99). 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. In spite of the limited understanding capabilities of the system, the score for "*Did Furhat understand what you said?*" was 2,99 (SD 1.05)3. This might also be seen as an indication that the tricks used actually worked.

8. Conclusions

We have described an audio-visual data collection with a spoken dialogue system embodied by the animated talking head Furhat. The data collection contains situated data in a real-world public place - a museum with thousands of visitors passing by over four days. A novelty with the current corpus is that it contains multi-party dialogues in a public space between two humans and a robotic head. It is an example of a risky and expensive type of data collection, where great attention is paid to keeping the situation and the environment authentic and the quantities of data large, at the expense of control and system performance. A common factor for these data collections is that they collect data of human-machine dialogues that are actually more complex than what state-of-the-art technology can actually accomplish today. There are several reasons to do this - the need for data to further research and development, and the showcasing of future possibilities. Another common way of achieving this is by using Wizard-of-Oz systems, but in these massive public space collections, such systems are not feasible, both for reasons of scale and ethics.

We have described how, in this type of data collection, it is essential to exploit every trick available in order to make the conversations appear better than they actually are, if judged by the systems ability to understand and respond to what its interlocutors say. Although data collected in such setup are rich in natural interactional behaviours from naïve users, it is to some degree limited in how people interact in today's state-of-the-art task-oriented dialogue systems. Instead, the motivation for collecting this type of data is that it is essential for us to gain insights into how people may behave when interacting with and perceive future dialogue systems and technologies. By doing this, our efforts can be guided in the right direction.

9. Acknowledgements

This work is partly supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314. Also thanks to David Traum and Ron Artstein for providing transcriptions from the Ada and Grace interactions and to Jens Edlund for contributing in writing the paper.

10. 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., and Bell, L. (2000). Speech Technology on Trial: Experiences from the August System. In Natural Language Engineering, 6.

Gustafson, J., and Sjölander, K. (2002). Voice transformations for improving children's speech recognition in a publicly available dialogue system. In Proc of ICSLP 2002 (pp. 297-300). Denver, Colorado, USA.

Harel, D. (1987). Statecharts: A visual formalism for complex systems. Science of computer programming, 8(3), 231-274.

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,* Berlin: Springer-Verlag.

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).*