Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction

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

Department of Speech, Music, and Hearing. KTH Royal Institute of Technology,
Lindstedtsvägen 24,
10044SE Stockholm, Sweden
{sameram, beskow, skantze, bjorn}@speech.kth.se
http://www.speech.kth.se

Abstract. In this chapter, we first present a summary of findings from two previous studies on the limitations of using flat displays with embodied conversational agents (ECAs) in the contexts of face-to-face human-agent interaction. We then motivate the need for a three dimensional display of faces to guarantee accurate delivery of gaze and directional movements and present *Furhat*, a novel, simple, highly effective, and human-like back-projected robot head that utilizes computer animation to deliver facial movements, and is equipped with a pan-tilt neck. After presenting a detailed summary on why and how *Furhat* was built, we discuss the advantages of using optically projected animated agents for interaction. We discuss using such agents in terms of situatedness, environment, context awareness, and social, human-like face-to-face interaction with robots where subtle nonverbal and social facial signals can be communicated. At the end of the chapter, we present a recent application of *Furhat* as a multimodal multiparty interaction system that was presented at the London Science Museum as part of a robot festival,. We conclude the paper by discussing future developments, applications and opportunities of this technology.

Keywords: Facial Animation, Talking Heads, Robot Heads, Gaze, Mona Lisa Effect, Avatar, Dialogue System, Situated Interaction, Back Projection, Gaze Perception, Furhat, Multimodal Interaction, Multiparty Interaction.

1 Introduction

There has always been an urge in humans to give machines an anthropomorphic appearance and behavior. This urge, perhaps, comes from the human interest to understand and recreate themselves, since humans can be considered (or at least appear to be) the most intelligent and complex animations of life.

This orientation of giving machines a human body and face has been clear since the beginning of works on robotics. For example, the word “robot” was introduced to the public by the Czech interwar writer Karel Čapek in his play R.U.R. (Rossum's Universal Robots), published in 1920. The play begins in a factory that makes artificial people called robots, though they are closer to the modern ideas of androids, creatures that can be mistaken for humans [1].

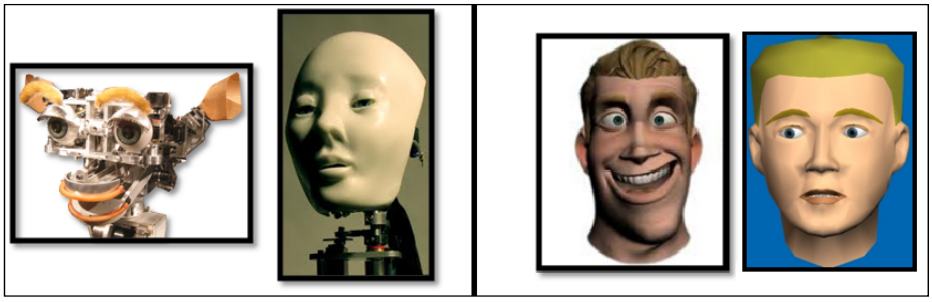
The Holy Grail in the quest for building human-like robots, however, has been the human face. Simulating the appearance and dynamics of the human face has been shown to be an intensely complex matter. The human face, with its subtle and minute movements, carries an incredible amount of information that is designed to be read and interpreted by others. For instance, the human lips carry significant information about speech and intonation [2] [3], the eyes are a mirror to the mind, affect and attention ([4] [5]). The combination of these components provides the human with the possibility to communicate emotions as well as interests. However, it also provides information about more physical parameters such as age and gender, ([6]).

The efforts for building natural anthropomorphic faces has mainly taken two different tracks; one building of physical, mechanical heads that simulate the structure and appearance of a human face; and the other one has been focusing on building three dimensional digital animated computer models. Figure 1 illustrates examples for both tracks.

Building computer simulations of the human face has indeed been a challenging task, but recently making impressive progress. This is mainly due to its major applications in the gaming and moving-picture industries, those being the driving forces behind much of the progress. These models have also been intensively used as a research tool to better understand the functionality of the human face, taking advantage of the flexibility and easy manipulation of these models. An important advantage of these computer models is that they can be replicated at no cost, providing different branches of research and industry with very good accessibility.

Unfortunately, this advancement has not been paralleled in robotics in general: The easy control of computer models is not easily mapped onto control of muscular and mechatronic movements of servos implemented in robotic heads [7], introducing huge limitations in human-looking robotic faces to exhibit smooth and human-like movement, and hence introducing inconsistencies between how the robot looks and how it behaves (usually referred to as the uncanny valley [8]). The other limitation of building human-like robotic faces is their expensive manufacturing and replication. At the moment, there are only a handful of human-like robots, which is making them exclusive and inaccessible to both the research community and the public.

Some trials have been carried out to bridge this gap between software animation (virtual agents) and physical robots. One solution has been to use a computer screen as a robot head [9], with a virtual agent embedded into it. This approach offers a face with natural looks and dynamics while preserving a physical robot body. However, it naturally suffers several limitations and problems that come with using a flat display as an alternative to a three dimensional physical head, such as that, (aside from large aesthetic inconsistencies), flat displays are not three dimensional and suffer from lacking absolute direction of what is presented into them in relation to where the screen is placed (more detailed discussion in Section 2).



(a) (b)

Fig. 1. *(a) Two examples of physical robot heads. (b) Two examples of computer animated facial models.*

In this chapter, we are presenting a highly natural and effective hybrid solution for using animated agents for robotic heads. We are building on two previous studies that demonstrate the limitations of flat screens in delivering accurate direction of gaze, and hence limit the capabilities of animated agents to carry out situated, multiparty interaction. After that, we present *Furhat*, a three dimensional back-projected robot head that utilizes a computer animated face. We describe the details on how *Furhat* was built and what advantages it offers over, both in-screen animated agents, and mechanical robotic heads. After that we discuss possible applications of using *Furhat* for multimodal, multiparty human-machine interaction, and demonstrate a system for a three-party dialogue with *Furhat* which has recently been showcased at the London Science Museum as part of a European robot festival.

2 Animated Agent and Mechanical Robots

As discussed earlier, interactive agents that are made to look and act as humans can come in two instantiations. First as virtual characters (where the body and face of the agent is a computer software), or second, as physical robots.

One may think of robots as situated physical agents: At the time of interaction, the agent and the human are co-present spatially and temporally, which ultimately simulates the human-human communication setup. However, virtual agents are computer software that are, clearly, not co-present spatially with the interactive partner (the human) in the same space, but can be thought of as living in a virtual space. Many approaches have been tried to optimally bridge these two physical and virtual worlds, and bring the human and the virtual agent into the same world. Those being virtual reality interfaces (Figure 2 left), and holographic projections (Figure 2 right).

In virtual reality, pragmatically, the human is transferred into the three dimensional virtual world, while in holographic projection, the virtual three dimensional world is transferred into our own reality, and hence, both co-exist spatially with the human interlocutor.

These two solutions are highly complex, exclusive and expensive, and are seldom used as a user interface with virtual characters. However, the predominant solution to

bridging the virtual and the real worlds has been via projections onto flat displays (such as flat screens, wall projections, etc.); an example is shown in the middle of Figure 2. The flat display functions as a window between the world the human interlocutor is situated in, and the virtual world of the virtual character [10].

(a) (b) (c)

Fig. 2. *(a) An example of a person wearing virtual reality (VR) glasses, so to be immersed in a virtual world. (b) An example of a virtual character that is presented via a flat display, offering a bridge into the physical and virtual realities. (c) An example of a holographic display of a person, to bring the virtual character into the physical space.*

It is known that the perception of three-dimensional objects that are displayed on two-dimensional surfaces is guided by, what is commonly referred to as the Mona Lisa effect [11]. This means that the orientation of the three-dimensional objects in relation to the observer will be perceived as constant, no matter where the observer is standing in the room or in relation to the display. For example, if the portrait of a face is gazing forward, mutual gaze will be established between the portrait and the observer, and this mutual gaze will hold no matter where the observer is standing. Accordingly, if the portrayed face is gazing to the right, everyone in the room will perceive the face as looking to their left. Thus, either all observers will establish mutual gaze with the portrait or none of them will. This implies that no exclusive eye-contact between the portrait and only one of the observers is possible. This principle, of course, extends to all objects viewed on 2D surfaces, such as pointing hands or arrows.

This effect can be seen as the cost of bridging the two different, virtual and real, worlds, to allow for direct visual interaction between humans and animated agents. This effect, clearly, has important implications on the design of interactive systems, such as embodied conversation agents, that are able to engage in situated interaction, as in pointing to objects in the environment of the interaction partner, or looking at one exclusive observer in a crowd.

In the following two sections we will present the results from two previous studies showing the limitations of the Mona Lisa effect on interaction, and presenting an approach on extending the use of animated faces from the flat screen onto physical three dimensional head models (and so building a physical situated robotic head).

These two studies represent a proof of concept of this approach to overcome the limitations of flat displays of animated faces.

3 Background Study 1: Perception of gaze

Since the Mona Lisa gaze effect is introduced by 2D projection surfaces, we suggested an alternative to 2D projection surfaces, by which the Mona Lisa gaze effect would be avoided. Our approach in this experiment was to use a 3D physical, static model of a human head (as seen in Figure 3). In order to compare this model with a traditional 2D projection surface, we designed an experimental paradigm that tests for mutual gaze as well as for gaze direction in the physical space of the viewer. The method is used to test the differences in accuracies in predicting gaze direction from a face that is presented through a 2D surface and the 3D projected surface.

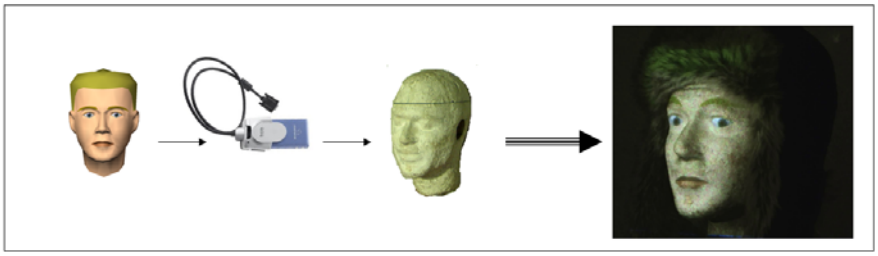


Fig. 3. *An earlier approach for front projecting an animated face onto a physical head model using a micro laser projector.*

The technique of manipulating static objects with light is commonly referred to as the *Shader Lamps* technique [12] [13]. This technique is used to change the physical appearance of still objects by illuminating them using projections of static or animated textures, or video streams.

In the perception experiment in [14], five subjects were simultaneously seated around an animated agent, which shifted its gaze in different directions (see Figure 4). After each shift, each subject reported who the animated agent was looking at. Two different versions of the same head were used, one projected on a 2D surface, and one projected on a 3D static head-model (see Figure 5). The results showed a very clear Mona Lisa effect in the 2D setting, where all subjects perceived a mutual gaze with the head at the same time for frontal and near frontal gaze angles.

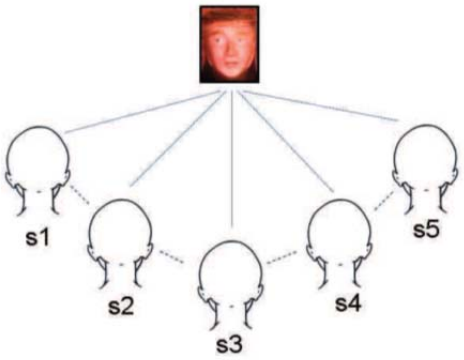


Fig. 4. *Schematic setup and placement of the subject and stimuli point*

While the head was not looking frontal, none of the subjects perceived mutual gaze with the head. In the 3D setting, the Mona Lisa effect was completely eliminated and the agent was able to establish mutual and exclusive gaze with any of the subjects. The subjects achieved a very high agreement rate on guessing on which subject the gaze of the agent was directed at for all the different gaze shifts.

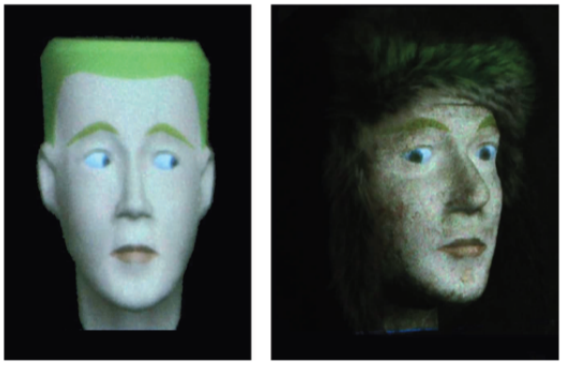


Fig. 5. *A snapshot of the animated agent displayed on a 2D white board (left), and on a 3D head model (right).*

This study provides important insights and proves the principal directional properties of gaze through a 2D display surface. The study also shows that using the simple approach of optically projecting the same face model onto a 3D physical head model would eliminate that effect. However, the study does not show whether this effect will hold during interaction, or whether people are able to cognitively compensate for the effect, and correctly infer the *intended* direction of gaze.

4 Background Study 2: Interactional effects of gaze

In order to explore the interactional effects of gaze in a multi-party conversational setting, a similar experiment was carried out, but with spoken interaction between the head and the participants [15]. Unlike the previous perception experiment, which focused on the *perceived* gaze, this experiment investigated how gaze may affect the turn-taking *behavior* of the subjects, depending on the use of 2D or 3D displays.

Two sets of five subjects were asked to take part in the experiment. In each session, the five subjects were seated at fixed positions at an equal distance from each other and from an animated agent (just as in the previous experiment, see Figure 4). The agent addressed the subjects by directing its gaze in their direction. Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface, as shown in Figure 5. The conversational behavior of the animated agent was controlled using a Wizard-of-Oz setup. For each new question posed by the agent, the gaze was randomly shifted to a new subject. The subjects were given the task of watching a video from a camera navigating around the city of Stockholm, after which the animated agent asked them to describe the route they had just seen. After each video was finished, the animated agent started to ask the subjects about directions on how to reach the landmark the video ended with, starting from the point of view the video started with. Each set of subjects did four dialogs in both the 2D and the 3D condition (i.e. a total of eight videos).

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. The mean response time was also calculated for each condition, i.e. the time between the gaze shift of the question and the time takes for one of the subjects to answer, which showed a significant difference in response time between the two conditions: 1.86 seconds for the 2D condition vs. 1.38 seconds in the 3D condition.

The results show that the use of gaze for turn-taking control on 2D displays is limited due to the Mona Lisa effect. The accuracy of 50% is probably too low in settings where many users are involved. By using a 3D projection, this problem can be avoided to a large extent. However, the accuracy for the 2D condition was higher than what was reported in the previous experiment. A likely explanation for this is that the subjects in this task may to some extent compensate for the Mona Lisa effect – even if they do not "feel" like the agent is looking at them, they may learn to associate the agent's gaze with the intended target subject. This comes at a cost, however, which is indicated by the longer mean response time. The longer response time might be due to the greater cognitive effort required making this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

The subjects were also asked to fill out a questionnaire after the interactions, in which they compared the two versions of the head along three dimensions, as shown in Figure 6. As the figure shows, the 3D version was clearly preferred, perceived as more natural, and judged as less confusing when it comes to knowing whose turn it was to speak.

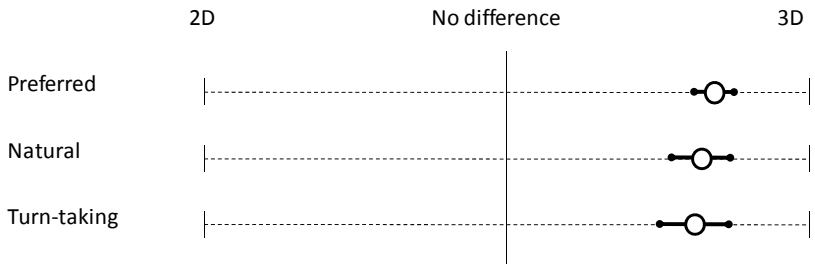


Fig. 6. *The subjective assessment of the 2D and 3D versions of the talking head, showing mean and standard errors.*

5 The Furhat Robot Head

As shown in the previous studies and discussions above, the paradigm of using a physical head model as a projection surface for animated computer models, would not only bring the face outside of the traditional two-dimensional screen, but will also eliminate the Mona Lisa effect and allow for multiparty interaction. From the study above in Section 4, it also appears that people perceive the projected face as significantly more natural than the face shown inside the screen. In addition to that, using the animated computer model as an alternative to a physical robot head solves major difficulties for building naturally looking and moving robot faces, since the technology behind facial animation has reached impressive advancements, and the control of these faces is highly simple and flexible. (Refer to [16] for a short review on the benefits of this approach).

Building on these encouraging findings, we have started building a natural and human-like robotic head that is based on the principle of optically projected computer models. A main modification was applied to the previous approach; that is to back-project the face onto the mask, so that the projector is hidden behind the mask. This means that if the mask is placed onto a robotic neck, the mask and the projector will be attached together and the projected image will not be displaced.

To build the head, several factors had to be taken into account. For example, micro projectors have a small projection angle, and hence if the projector is placed too close to the mask, the projected image will not be big enough to cover the entire projection area of the mask. Another factor was to use a material that will diffuse the light over the mask so that the light projected on the mask will be equally illuminated. One last important factor that had to be taken into account is to be able to acquire a mask model that would exactly fit the design of the projected face, so that no calibration and transformations of the model will be needed, and subtle facial areas, like the eyes, will naturally fit the area of the eyes on the mask.

Figure 7 shows a flow chart of the process of how the back-projected head is built. We call the head *Furhat*, as it got a fur hat that covers the top and the sides of the mask. Following is a detailed description on how *Furhat* has been built, so that it would provide more insights into the properties of the head, and comes as a guide for others to replicate it.

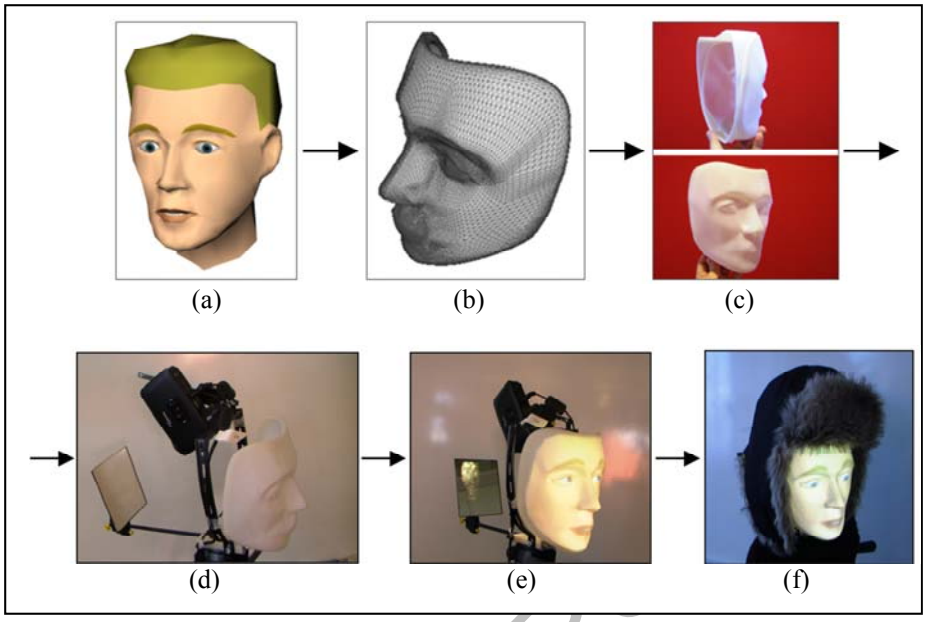


Fig. 7. *A chart showing the process for building Furhat, the back projected robotic head.*

Building Furhat

In the following section we provide a chronological list of the main steps taken to build the robot head:

1- Using an animated face model: The 3D animated face model that is used for this study is detailed in [17]. An animated face model is used due to several reasons: The lips of the face model can be automatically synchronized with the speech signal the system is producing; this is done by using a transcription of the speech utterances to be produced. The lip synchronization system utilized in the face model has proven to enhance speech intelligibility over listening only to the audio signal [18]. This face model also offers flexible control of gestures and facial movements (gaze movements, eyebrows movements, etc.). Gestures played using this face model have also been shown to deliver the communicative functions they are designed for (eyebrows raise to signal questioning, doubt, or surprise [19]); these gestures have also been shown to enhance speech intelligibility [20].
From this evidence, it is clear that this face model can deliver highly accurate and natural movements and would be suitable as a choice for *Furhat*'s face.

2- Printing the 3D mask: The main step is to establish a translucent mask that would allow the back projected light to be clearly visible when looked at from the front. The other important factor is to establish a mask that fits in its shape,

the face model that would be projected on top of it (mentioned in the previous point). To establish this, a 3D copy of the exact face 3D model was printed using a 3D printer, with an equal overall thickness of 1mm. After sample testing, this thickness proved optimal to allow just enough light to be visible on the mask. Figure 7a shows the original 3D computer model of the face. Figure 7b shows the 3D design of the mask acquired by modifying the original face model, and making it suitable for 3D printing. Figure 7c shows the mask after printing. The dimensions of the printed mask were made to resemble the size of an average human head (width 16cm, height 22cm, depth 13cm).

3- Allowing the mask to equally diffuse color: A main problem of back projecting light on translucent objects is that the light-source will be visible (glowing) when looked at from the front. This was an obvious problem when the printed face was used with a micro projector. To solve this problem, a back-projection paint, which is used to create back-projection screens, was used (goo systems Global[1]). This spray paint is used specifically to allow the cured surface to diffuse the light[1] equally over its surface, and hence diminish the problem of unbalanced optical illumination over the mask. Figure 7e shows the back-projected face after applying the back-projection paint on the plastic mask.

Fig. 8. *A front and back view of the mask and the rig of Furhat*

4- Rigging the mask with a micro projector: When the mask was tested and proved ready to use as the back projection mask for the head, the mask then was rigged with a micro-projector that was placed on top of the mask, the projector then projects light onto a mirror that reflects back the face onto the mask. This approach allows for more distance between the projector and the mask, which in turn, allows for the projected image to be in focus and to fit the entire mask.

Figure 7d shows how the head is rigged with a projector and a mirror. Figure 8 shows a front and back view of the head when the mask is rigged with the projector and a mirror, showing how the projected face fits exactly the 3D plastic mask (it is important to note here that the solution of using a mirror is

[1] http://www.goosystemsglobal.com/

probably replaceable by other alternatives such as using a fish-eye lens that widens the projection area of the projector). After the mask was rigged, the head was covered using a fur hat. The fur hat covers the projector and the rig, and hence gives a stronger focus on the facial appearance of *Furhat*. Figure 9 shows *Furhat* with and without its head cover.

Fig. 9. *Snapshots of Furhat with and without the head cover (the fur hat).*

5- Giving *Furhat* a neck:

Direction of attention may of course not only be achieved with the eyes, but also by moving the head, using a neck. A neck allows the robot head to use either eye movement, head pose, or both, to direct the attention, but also to do gestures such as nodding. Depending on which behaviors need to be modeled, different degrees of freedom (DOF) may be necessary. To direct the gaze in any direction (if the eyes are centered), 2 DOF are obviously necessary, but in order to perform a wider range of gestures, more DOF may be needed. An example of a very flexible robot neck is presented in [21], where 3 DOF are used: lower and upper pitch (tilting up and down), yaw (panning side to side) and rolling (tilting side to side). Lower pitch is centered where the neck meets the shoulders, and high pitch is centered where the neck is attached to the head.

For *Furhat*, we are currently using a pan-tilt unit. The unit has a no-load speed of 0.162 sec/60° and a holding torque of 64 kg·cm. It has 2 DOF: pitch and yaw, which allows *Furhat* to direct the head in any direction, but also to do simple gestures such as nodding.

6 Example application

The development of *Furhat* is part of a European project called IURO (Interactive Urban Robot)[2]. As part of this project, we were invited to the EUNIC RobotVille Festival at the London Science Museum, December 1st – December 4th, 2011. The purpose of the IURO project is to develop robots that can obtain missing information from humans through multi-party dialogue. The central test-case will be an autonomous robot that can navigate in an urban environment by asking humans for directions. For the exhibition, we wanted to explore a similar problem, but to suit the setting we instead gave *Furhat* the task of asking the visitors about their beliefs of the future of robots, with the possibility of talking to two visitors at the same time and shifting attention between them.

In lab setups, we have been using Microsoft Kinect[3], which includes a depth camera for visual tracking of people approaching *Furhat* and an array microphone for speech recognition. However, due to the crowded and noisy environment in the museum, we chose to use handheld close-range microphones and ultrasound proximity. For speech recognition, the Microsoft Speech API was used. For speech synthesis, we used the CereVoice William TTS from CereProc[4]. CereVoice reports the timing of the phonemes in the synthesized utterance, which was used for synchronization of the lip movements in the facial animation. It also contains a number of verbal gestures that were used to give *Furhat* a more human-like appearance, such as grunts, laughter and yawning.

To control *Furhat*'s behavior, we used an event-driven system implemented in Java, inspired by Harel state-charts [22] and the UML modeling language. This allowed the system to react to external sensory input (speech, proximity data) as well as self-monitoring data, and produce actions such as speech, facial gestures and head movements. The layered structure of the state-chart paradigm allows the dialogue designer to define a hierarchy of dialogue states, and the sensory-action paring that is associated with these states. 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 (i.e., "when do you think robots will beat humans in football?") and one where the visitors asked questions to *Furhat* (i.e., "where do you come from?"). In the former case, *Furhat* continued the dialogue (i.e., "why do you think so?"), even though he often understood very little of the actual answers, occasionally extracting important keywords.

With nobody close to the proximity sensors, *Furhat* was in an "idle" mode, looking down. As soon as somebody approached a proximity sensor, he looked up and initiated a dialogue with "Could you perhaps help me?". The multi-party setting allowed us to explore the use of head-pose and gaze during the dialogue:

- With two people standing in front of him, *Furhat* was able to switch interlocutor using first a rapid gaze movement and then head movement. Often *Furhat* used this possibility to move the dialogue forward, by

[2] http://www.iuro-project.eu/
[3] http://kinectforwindows.org/
[4] http://www.cereproc.com/

switching interlocutor and asking a follow-up, such as "do you agree on that?"

- *Furhat* could either ask a specific interlocutor, or direct the head between the interlocutors and pose an open question, moving the gaze back and forth between the interlocutors. By comparing the audio-level and timing of the audio input from the two microphones, *Furhat* could then choose who to attend and follow-up on.
- If *Furhat* asked a question specifically to one of the interlocutors, and the other person answered, he quickly used gaze to turn to this person saying "could you just wait a second", then shifted the gaze back and continued the dialogue.

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.

Fig. 10. *Furhat at the London Science Museum. The monitor shows the results of the visitors' answers to Furhat's questions. The two podiums with microphones and proximity sensors can also be seen.*

In total, 7949 people visited the exhibition during the course of 4 days. The system proved to be very stable during the whole period. Apart from the video data, we recorded 8 hours of speech from the visitors. We also let the visitors fill out a questionnaire about their experience after the interaction. We have not yet analyzed the data, but it was apparent that many visitors liked the interaction and continued to answer *Furhat*'s questions although he actually understood very little of their answers. The visitors also seemed to understand *Furhat*'s attentive behavior and act accordingly. Videos from the exhibition can be seen at www.speech.kth.se/furhat.

7 Discussions

One major motivation behind this work is to build a robot head that can use state-of-the-art facial animation to communicate and interact with humans. These include natural and smooth lip movements, control of perceivable eye and gaze movements.

To make a robot head that is able to capitalize on social signals, its head should be able to generate such signals to highly perceivable accuracy. The first step towards reaching this goal was to use animated talking agents. However, since the robot is supposed to be able to engage in interactive multimodal dialogue with multiple people, the simple solution of using a computer screen as an interface with an animated agent projected onto it became disadvantageous. This is due to the fact that the 2D screen has no direction, and suffers from the Mona Lisa gaze effect (amongst other effects). This effect makes it impossible to establish, for example, exclusive eye-contact with one person out of many.

The solution to reach these goals, while avoiding the hindering effects of flat displays, is *Furhat*, a hybrid solution that can be thought of as bringing the animated face out of the screen and into the real-physical world.

Fig. 11. *Examples showing different instantiations of the colors of Furhat's facial features.*

Clearly, the benefits of using an animated agent as a robot head employing optical projection meets the goal of bringing the smooth and accurate animation of 3D computer models into a robot head. But there are more advantages. The flexibility of using a computer model allows for fast and online control of the face depending, for example, on context. *Furhat* for example can change its facial design on the fly since the colors and shape of its different facial parts is just a software animation (this manipulation is however limited by the design of the mask). Figure 11 shows examples of different facial colors of *Furhat.*

These and other parameters can be controlled depending on context and the environment, for example, *Furhat* can have a different facial design depending on the cultural background, or the age of the interlocutor. It can change its color contrasts depending on the surrounding light.

One expressive and environment-sensitive part of the face that can be controlled in this setup is the eyes. The pupil size for instance, can correspond to the amount of light in the surroundings [23], and can also reflect functions such as affect and interest.

Another context-aware property of the eyes is the corneal reflection. This is when the image of the environment is reflected on the cornea. This phenomenon has been shown to provide significant amount of information about the environment where the interaction is taking place [24].

These features can be easily implemented in *Furhat* on the software side by controlling the size and textures of the eyes and hence the projected image will more accurately reflect the situated context *Furhat* is interacting in.

Other benefits of *Furhat* to be used as a robotic head are its low weight, low maintenance demands, low noise level (only the noise coming from the neck), and its low energy consumption.

8 Conclusions

In this chapter, we have presented *Furhat*, an example of a paradigm for building robot heads using software based animated faces. Based on experimental evidence, this paradigm makes animated faces look more natural and human-like since it brings them out of the screen and onto a human-head-shaped three-dimensional physical object. This, not only makes animated faces look more natural in interactions, but also solves problems that arise when visualizing them onto flat displays. Such potential problems are achieving accurate multiparty interaction using gaze and head direction (since flat displays lack the enforcement of direction).

Looking at what *Furhat* has to offer to robotic heads, the advantages of using software design and animation instead of hardware (physical-mechanical) design and animation are numerous. Robot heads lack the ability to move their facial parts smoothly and accurately enough to simulate human facial movements (eye movements, blinking, eyebrows movement, and specially lip movements), let alone looking like human ones.

Furhat, on the other hand, uses an animated face that can move its facial parts online, in real-time, and to a large degree like humans do. In addition to movement, the design of the face is very flexible. The design of robot heads typically cannot change after manufacturing the head (the color and design of the lips and eyebrows, the color of the eyes, the size of the iris…), *Furhat*'s colors and design, on the other hand, can easily change. This is achieved by using the animated face model it utilizes as its face, while still using the same face mask and hardware, and hence no mechanical or hardware cost is associated with this functionality.

After we presented *Furhat* and how it was built in this paper, allowing for others to possibly replicate the process, we have presented a sample application that uses *Furhat* for multiparty interaction with human, which was presented at the London Science Museum for 4 days and received around 8000 visitors.

We would like to use *Furhat* not only as a natural interactive robot head, but also as a research framework which allows for studying human-human (one can think of

Furhat as a tele-presence device) and human-robot interaction in single and multiparty setups and in turn-taking and dialogue management techniques using face and neck movements, to count a few.

Acknowledgments. This work has been done at the Department for Speech, Music and Hearing, and funded by the EU project IURO (Interactive Urban Robot) No. 248314. The authors would like to thank Simon Alexanderson for designing the 3D mask model for printing, and to thank Jens Edlund, Joakim Gustafson and Preben Wik for their interest and inspiring discussions.

References

1. Dominik , Z. :"Who did actually invent the word "robot" and what does it mean?". The Karel Čapek website. Retrieved 2011-12-10. http://capek.misto.cz/english/robot.html
2. Summerfield, Q. :Lipreading and audio-visual speech perception. PhilosophicalTransactions: Biological Sciences, vol. 335, no. 1273, pp. 71-78. (1992).
3. Al Moubayed, S., & Beskow, J. :Effects of Visual Prominence Cues on Speech Intelligibility. In Proceedings of Auditory-Visual Speech Processing AVSP'09. Norwich, England, (2009).
4. Argyle, M., & Cook, M. :Gaze and mutual gaze. Cambridge University Press, (1976).
5. Kleinke, C. L. :Gaze and eye contact: a research review. *Psychological Bulletin, 100*, 78-100, (1986).
6. Ekman, P. & Friesen, W.V. :Unmasking the face: A guide to recognizing emotions from facial clues. Malor Books. ISBN: 978-1883536367, (2003).
7. Shinozawa K., Naya F., Yamato J., & Kogure K. :Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human Computer Studies*, 62 (2), pp. 267-279, (2005).
8. Mori, Masahiro. :Bukimi no tani. :The uncanny valley (K. F. MacDorman & T. Minato, Trans.). Energy, 7(4), 33–35. (Originally in Japanese), (1970).
9. Gockley, R., Simmons, J., Wang, D., Busquets, C., DiSalvo, Ke., Caffrey, S., Rosenthal, J., Mink, S., Thomas, W., Adams, T., Lauducci, M., Bugajska, D., Perzanowski, and Schultz, A. :Grace and George: Social Robots at AAAI. Proceedings of AAAI'04. Mobile Robot Competition Workshop, pp. 15-20, AAAI Press, (2004).
10. Edlund, J., Al Moubayed, S., & Beskow, J. :Mona Lisa as a measure of co-spatiality. In proceedings of the international conference on Intelligent Virtual Agents IVA'11, (2011).
11. Todorovi, D. :Geometrical basis of perception of gaze direction. *Vision Research, 45*(21), 3549-3562, (2006).
12. Raskar, R., Welch, G., Low, K-L., & Bandyopadhyay, D. :Shader lamps: animating real objects with image-based illumination. In *Proc. of the 12th Eurographics Workshop on Rendering Techniques* (pp. 89-102), (2001).

13. Lincoln, P., Welch, G., Nashel, A., Ilie, A., State, A., & Fuchs, H. :Animatronic shader lamps avatars. In *Proc. of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality (ISMAR '09)*. Washington, DC, US: IEEE Computer Society, (2009).
14. Al Moubayed, S., Edlund, J., & Beskow, J. :Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. *ACM Trans. Interact. Intell. Syst.* 1, 2, Article 11, 25 pages, (2012).
15. Al Moubayed, S., & Skantze, G. :Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In *Proceedings of the* international conference on Auditory-Visual Speech Processing AVSP. Florence, Italy, (2011).
16. Al Moubayed, S., Beskow, J., Edlund, J., Granström, B., & House, D. :Animated Faces for Robotic Heads: Gaze and Beyond. In Esposito, A. et al (Eds.), Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues, Lecture Notes in Computer Science, 2011, Volume 6800/2011, 19-35, DOI: 10.1007/978-3-642-25775-9_2, (2011).
17. Beskow, J. :Talking heads - Models and applications for multimodal speech synthesis. Doctoral dissertation, KTH, (2003).
18. Beskow, J. :Animation of talking agents. In Benoit, C., & Campbel, R. (Eds.), Proc of ESCA Workshop on Audio-Visual Speech Processing (pp. 149-152). Rhodes, Greece, (1997).
19. Granström, B., & House, D. :Modeling and evaluating verbal and non-verbal communication in talking animated interface agents. In Dybkjaer, l., Hemsen, H., & Minker, W. (Eds.), Evaluation of Text and Speech Systems (pp. 65-98). Springer-Verlag Ltd, (2007).
20. Al Moubayed, S., Beskow, J., & Granström, B. :Auditory-Visual Prominence: From Intelligibility to Behavior. Journal on Multimodal User Interfaces, 3(4), 299-311, (2010).
21. Brouwer, D.M., Bennik, J., Leideman, J., Soemers, H.M.J.R., & Stramigioli, S. :Mechatronic Design of a Fast and Long Range 4 Degrees of Freedom Humanoid Neck. In *proceedings of ICRA,* Kobe, Japan, ThB8.2. pp. 574-579, (2009).
22. Harel, D. :Statecharts: A visual formalism for complex systems. Science of computer programming, Elsevier. pp231—274, 8(3), (1987).
23. Blackwell, R.D., Hensel, J.S., & Sternthal, B. :Pupil dilation: What does it measure? Journal of Advertising Research 10, 15–18 (1970).
24. Nishino, K. & Nayar, S. K. :Corneal Imaging System: Environment from Eyes. *Int. J. Comput. Vision* 70, 1 (October 2006), 23-40. DOI=10.1007/s11263-006-6274-9, (2006).