THE FURHAT BACK-PROJECTED HUMANOID HEAD – LIP READING, GAZE AND MULTI-PARTY INTERACTION

SAMER AL MOUBAYED
GABRIEL SKANTZE
JONAS BESKOW

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

In this article, we present Furhat – a back-projected human-like robot head using state-of-the art facial animation. Three experiments are presented where we investigate how the head might facilitate human-robot face-to-face interaction. First, we investigate how the animated lips increase the intelligibility of the spoken output, and compare this to an animated agent presented on a flat screen, as well as to a human face. Second, we investigate the accuracy of the perception of Furhat's gaze in a setting typical for situated interaction, where Furhat and a human are sitting around a table. The accuracy of the perception of Furhat's gaze is measured depending on eye design, head movement and viewing angle. Third, we investigate the turn-taking accuracy of Furhat in a multi-party interactive setting, as compared to an animated agent on a flat screen. We conclude with some observations from a public setting at a museum, where Furhat interacted with thousands of visitors in a multi-party interaction.

Keywords: Robot Head, Humanoid, Android, Facial Animation, Talking Heads, Gaze, Mona Lisa Effect, Avatar, Dialog System, Situated Interaction, Back-projection, Gaze Perception, Furhat, Lip reading, Multimodal Interaction, Multiparty Interaction.

1. Introduction

Building 3D animated computer models of faces is a technology that has advanced significantly in recent years, providing the possibility to build highly realistic, accurate and dynamic avatars. That is partly due to their direct deployment in the moving picture and gaming industries and their flexibility as a research tool in human-human and human-machine face-to-face interaction.

However, designing computer programmes that interact with humans have also been targeted by building humanoid physiognomies not only as software based simulations of the human physiology, but also as situated physical humanoid robots. Physical robot heads are situated in the interaction environment with the humans they are built to interact with, and allow for the possibility to combine them with humanoid robotic bodies that manipulate and navigate their physical environment and interact with humans in a humanlike manner. However, compared to their computer simulation counterparts, robot heads have seemingly been lagging behind. The control and animation of computer models do not easily and robustly map directly into control of mechatronic physical heads, which allow much less detailed control of facial expressions. In addition to that, mechatronic robotic heads are significantly heavier, noisier and demand more energy and maintenance compared to their digital counterpart, while they are expensive and hence exclusive.

In the study of face-to-face communication models, the face and the head have a particular importance due to the large number of important verbal and non-verbal communicative roles they can play in the interaction. This study has been greatly facilitated by the use of animated faces (or talking heads, e.g. [1]) due to their ease of use and flexibility giving them direct access for researchers from different fields and being employed in controlled perception and production experiment research(e.g. [2]).

To bridge the gap between animated heads and robotic heads, several researchers have investigated the use of an animated face, back-projected on a translucent mask. The technology of back-projecting a computer model of a face on a physical three dimensional surface is a technology that is far from new in theory, but has recently gained rapid ground as technological advancements started to allow for the building of functioning prototypes of these heads, bringing the two research fields of facial animation and humanoid robotics closer.

The first traceable implementation of a face projection is the Grim Grinning Ghosts at the Disneyland Haunted Mansion ride opened in 1969[1]. The ghosts are lit simply by projecting a previously recorded movie of the faces matching the face models of the ghosts. Influenced by the experiment at the Disneyland Haunted Mansion, the MIT Architecture Group in 1980 released their first talking head projection, which was driven by tracking a person's head and projecting it on a mask of that same person [3][2].

Recently, more groups have put efforts into creating back-projected humanoid heads as an alternative to mechatronic humanoid heads and for different purposes. Morishima et al. [4] built the prototype system *HyperMask* which aims at projecting a face on a mask worn by a mobile actor, and Hashimoto & Morooka [5] use a spherical translucent surface to back-project an image of a humanoid face. *Lighthead*, built by Delauney et al. [6] is more elaborately designed, and maintains a stylized cartoonish face, built for research on non-verbal interaction. *Mask-bot*, created by Kuratate et al. [7], projects a photo-realistic animation of a real face on a generic plastic mask.

Here, we present *Furhat* [8], a robot head that deploys an animated face that is realistic and human-like in anatomy. Furhat relies on a state-of-the-art facial animation architecture that has been used in a large array of studies on human verbal and nonverbal communication (e.g. [9,10,11]). Furhat was built to study and implement patterns and models of human-human and human-machine situated and multiparty multimodal communication, a study that demands the co-presence of the talking head in the interaction environment.

Using a micro projector, an animated face is back-projected on a three-dimensional mask that is a 3D printout of the head used in the animation software, as illustrated in Fig 1. The mask and the projector are then rigged onto a pan-tilt-roll unit (neck) allowing Furhat to direct its attention using eye gaze and head pose. The animation has a three dimensional model of both eyes with a pupil, iris and eyelids that can be controlled sepa-

---

[1] Walt Disney's Wonderful World of Color, Season 16, Episode 20, Walt Disney Productions.

[2] A video can be seen on: http://www.naimark.net/projects/head.html.

rately. The animation also allows precise control of the eye brows and mouth, as well as accurate lip synchronization to the spoken output.

The development of Furhat has been partly motivated by the lack of co-presence in animated avatars on 2D displays. Since the agent is not spatially co-present with the user, 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. The same problem arises in situated interaction, where the agent might need to look at different objects in the surroundings. This phenomenon, known as the Mona Lisa effect, has been investigated in previous experiments at our lab. In a perception experiment [12], five subjects were simultaneously seated around an animated agent, which shifted the gaze in different directions. 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 (a prototype of Furhat). 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. Then the eyes were not looking frontal, none of the subjects perceived any mutual gaze. In the 3D setting, the Mona Lisa effect was eliminated and the agent was able to establish mutual and exclusive gaze with any of the subjects, in a way similar to physical situated human-human interaction.

In this article, we present three experiments with Furhat, where we investigate how the head might facilitate human-robot face-to-face interaction. The experiments also show how Furhat, an instantiation of the advancing technology of back-projection, can be used as a research platform to study cues and features of human-computer face-to-face communication.

In the first perception experiment, we investigate how the animated lips affect the intelligibility of the spoken output, and compare this to an animated agent on a flat screen, as well as a human face presented on a flat screen. In the second experiment, we investigate the accuracy of the perception of Furhat's gaze in a setting typical for situated interaction − where Furhat and a human are sitting around a table − depending on eye design, head movement and viewing angle. The third study can be regarded as a follow-up to our previous perception experiment on the Mona Lisa effect [12], mentioned above. While that study provided important insights and proves the principal directional properties of gaze through a 2D display surface, it 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. The study then investigates the *interactional* effects of the Mona Lisa effect compared to its basic perceptual properties, and shows whether Furhat is indeed necessary for situated interaction using gaze, or an agent presented on a 2D display will suffice during interaction. In this experiment, we have investigated to what extent the Mona Lisa effect affects the turn-taking accuracy in an interactive multi-party setting.

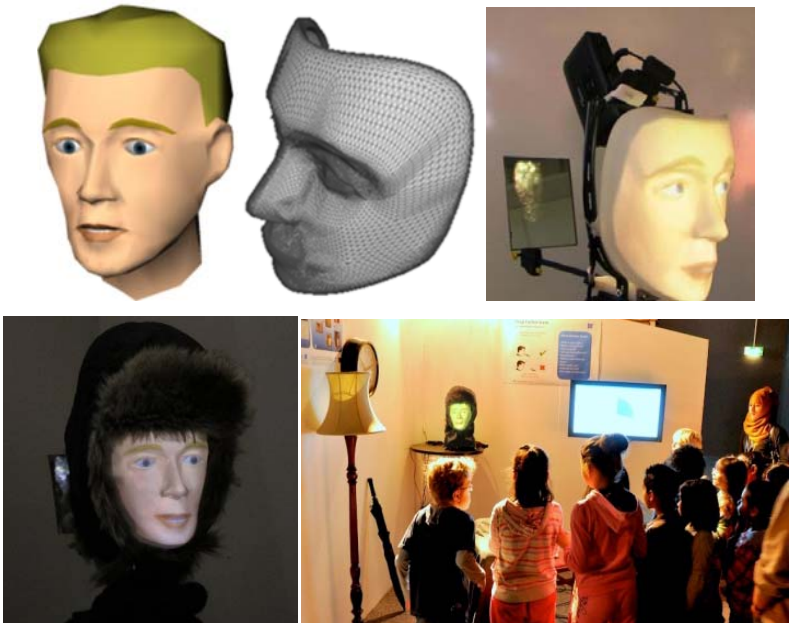


Fig. 1. The 3D animation model used for Furhat (top left), the mask (top middle) and the projection (top right). Below are shown pictures where the Furhat interacts with visitors at the London Science Museum.

These three experiments were all done in the lab, which provides the possibility to do controlled comparisons with different conditions. But it is also interesting to look at the interaction in an uncontrolled setting, where people spontaneously walk up to and talk to the robot, a setting that puts together the different technologies under the test. In December 2011, we were 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. During the four days of the exhibition, Furhat's task was to collect information on peoples' beliefs about the future of robots, through multi-party interaction (see Figure 1). The exhibition was seen by thousands of visitors (including many children), resulting in a corpus of about 10.000 user utterances [13]. We will conclude with some observations from this exhibition.

2. Study I: Visual Perception of Speech (Lip Reading)

One of the benefits of using back-projection (as compared to a mechanical head) is that the lips may be animated very accurately. This does not only enhance the illusion that the talking head itself is the source of the sound signal the system is communicating (rather than a separate process), but the lip movements also play an important role in speech perception and comprehension. The visible parts of the human vocal tract (the lips, tongue, teeth and jaw) carry direct information about the sounds the vocal tract is producing, and this information can be decoded from the movements of these articulators (a process usually referred to as *speech-reading*, or less accurately *lip-reading*). The

contribution of lip movements to speech perception has been thoroughly studied and quantified. Studies by Summerfield [14] show that the information carried in the face compared to only the acoustic signal can equal up to 11db benefit in Signal to Noise Ratio (SNR). The contribution of lip movements is especially evident when the quality of the speech signal is degraded, where looking at the lips can help substitute the loss in the speech signal.

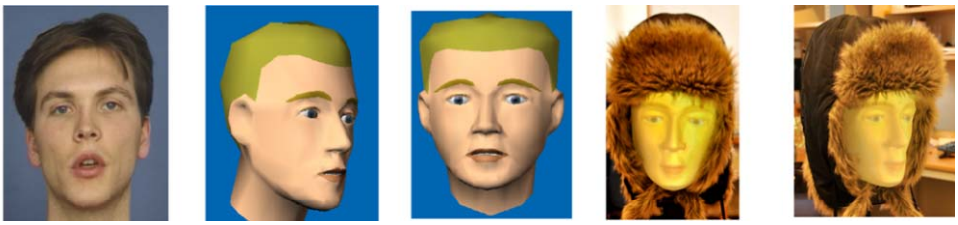
The relation between the speech signal and lip movement was shown early by McGurk & McDonald in their seminal article [15]. The studies showed that the perception of sounds is deeply connected with the information received by the ears and the eyes - consonants are perceived differently if the lip movements and the audio signal are incongruent. These important advantages of lip movement have been taken into account since the early developments on animated talking heads, and different models and techniques have been proposed and successfully applied to animate and synchronize the lips with the speech signal (e.g. [16, 17,18]).

The lip animation model used in Furhat has been previously tested, and found to enhance speech intelligibility in noise [9], when visualized onto a 2D screen. However, since Furhat's plastic mask is static (and therefore the jaw and lips), this might introduce inconsistency and non-alignment between the projected image and the projection surface. Hence, the fact that the animated lips do contribute to speech perception does not need to naturally hold with Furhat. Another possible hypothesis is that, since Furhat is physically three dimensional and more human like, this might decrease the cognitive effort spent reading the lips of the animated agent under high levels of noise, and could therefore perhaps even additionally improve speech perception beyond the original model.

The following study presents an evaluation experiment comparing audiovisual speech intelligibility of Furhat against the same animated face used in Furhat but visualized on a typical flat display.

2.1. Method

The setup used in this experiment introduces subjects to acoustically degraded sentences, where the content of the acoustic sentence was partially intelligible when listening only to the audio. A vocoder (identical to the one used in [11]) was applied to degrade the signal quality, which used band-pass filtering and replaced spectral details with white noise. This vocoder simulates the type of filtering in cochlear implants, and gives a predictable and systematic noise to the signal (as opposed to other types of noise such as babble). The sentences were then enriched by a lip-synchronized talking head to increase their intelligibility, using forced alignment between the phonetic labeling and the original audio. The audiovisual stimuli consisted of a collection of short and simple Swedish sentences, which varied in length between three to six words, and was designed to be phonetically balanced. e.g. *"Den gamla räven var slug"* (The old fox was cunning). The audio-visual intelligibility rate (accuracy) of each sentence was calculated as the number of words correctly identified by the subject, divided by the number of content words in



(a) Video (b) Screen45 (c) Screen0 (d) Furhat0 (e) Furhat45

Fig. 2. Snapshots of the different conditions of the visual stimuli.

the sentence. The sentences were presented in 6 different experimental conditions and the perceived intelligibility of the sentence (the number of correct words recognized) is compared across conditions. For each condition, 15 sentences were used. The conditions were systematically permutated among 10 normal hearing subjects, with normal or corrected to normal vision.

2.2. Conditions

Figure 2 shows snapshots of the stimuli associated with the conditions.

1 - Audio Only: In the audio-only condition, subjects listened to the acoustically degraded sentences without any visual stimuli. This condition is used as a baseline.

2 - Screen0: Talking head on a flat screen viewed at 0° angle: In this condition, the animated face was presented to the subjects along with the acoustic signal. The subject was seated in front of the screen, looking straight at the talking head. The talking head on the screen is oriented to look frontal (0 degrees rotation inside the screen).

3 - Furhat0: Furhat viewed at 0° angle: In this condition the sentences were presented to the subjects with the animated model back-projected on Furhat. The subjects were seated frontal to Furhat.

4 - Screen45: Talking head on a flat screen viewed at 45° angle: This condition is identical to the *Screen0* condition, except that the head is rotated 45° inside the screen. This condition is designed to compare the audio-visual intelligibility of the sentences with the condition *Screen0* and *Furhat45*.

5 - Furhat45: Furhat viewed at 45° angle: This condition is identical to *Furhat0*, except that subjects were seated at a 45° angle from Furhat. The viewing angle is hence identical to the one in condition *Screen45* (condition 3). This condition is meant to compare to *Screen45* condition, except for the projection surface.

6 - Video: In this condition, subjects were presented with the original video recordings of the sentences, viewed on the same flat display used to show the agent, and the size of the face was scaled to match the size of the animated face.

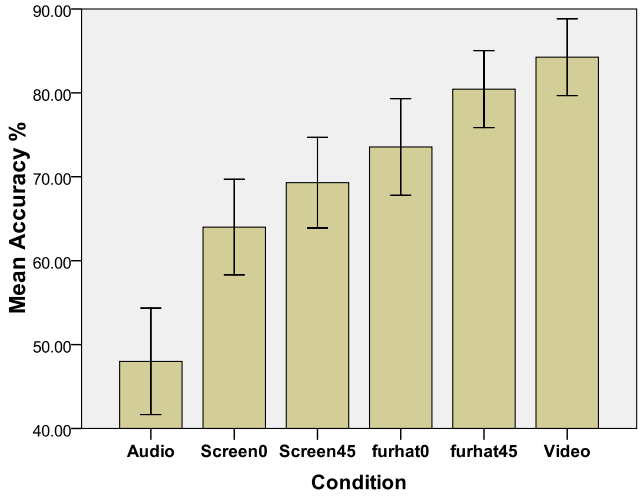


Fig. 3. The average accuracy rates for the different experimental conditions.

Table 1. p-values from the statistical significance test for all combinations of the different conditions.

Condition1	Condition2	*p*-value
Audio	* **Screen0**	***.000***
Screen0	* Screen45	.167
Screen45	* Furhat0	.266
Furhat0	* Furhat45	.079
Furhat45	* Video	.335
Screen0	* **Furhat0**	***.015***
Screen45	* **Furhat45**	***.004***
Furhat0	* **Video**	***.007***
All other	**combinations**	***<.0001***

2.3. Analysis and Results

A repeated measure ANOVA analysis was carried out on the sentence recognition rate (accuracy rate) as a dependent variable and the condition as an independent variable. The test shows a significant main effect ($F(5,45)=21.890$; $p<.0001$). The mean accuracy for each condition is shown in Figure 3, along with the standard error bars. A post-hoc LSD analysis was carried out to measure the significance values between the accuracy rates of each of the conditions (with a significance threshold of 0.05). Table 1 shows the *p-values* for all the conditions pairs, as a result from the LSD analysis.

The analysis firstly shows that the *Screen0* condition (and all other conditions), provide an audio-visual intelligibility that is significantly better than the *Audio* condition alone. This confirms the validity of the lip animation model using either a flat screen or the Furhat mask. The analysis also shows that there is no significant difference in the audio-visual intelligibility of the face being looked at either frontal or at a 45° (no signif-

icant difference between *Screen0* and *Screen45*, or between *Furhat0* and *Furhat45*). This is in line with the findings in [19], where it was found that the lip-reading contribution drops down when looking beyond 45 degrees, but is not significantly different between 0 and 45 degrees.

More importantly, the analysis shows that there is no loss in the audio-visual intelligibility when using the Furhat physically-static mask as a projection surface compared to using a flat screen. In fact, the intelligibility of Furhat is actually significantly *better* than the flat display (the rate is significantly higher for *Furhat0* over *Screen0*, and for *Furhat45* over *Screen45*).

The video condition, however, holds its place as the richest source of facial information, providing an audio-visual intelligibility that is higher, in average, than all the other experimental conditions. That difference is significant for all conditions except for *Furhat45*.

2.4. Discussion

From the findings of the study it appears that Furhat's static mask does not hinder intelligibility, but rather increases it significantly. This is indeed surprising and needs more experimentation to quantify its source. In the design of the Furhat mask, the details of the lips were removed and substituted by a smooth protruded curvature in order to not enforce a static shape of the lip. Because of this the size of the lips, when projected, is perceived slightly larger than the lips visualized on the screen. This enlargement in size might be the reason behind the increase in intelligibility. Another possibility is that looking at Furhat is cognitively easier than looking at a flat display (several subjects reported that: "It's easier and more comfortable to look at Furhat"). Furhat's face is spatially situated and more human-like than on a 2D screen, and due to the high sensitivity of the test to the cognitive state of the subject, using Furhat might have increased the level of focus and attention of the subjects, which could have resulted in a more efficient lip reading. To study this, we intend to carry out an experiment in which we record Furhat with a camera and show the video on a 2D screen, in which case the only variable is whether Furhat itself is 2D or 3D.

3. Study II: Perception of Gaze Direction for Situated Interaction

Moving from the movements of the lips to the movements of the eyes, or eye gaze, the function of gaze for interaction purposes has been investigated in several studies [20,21]. Gaze direction and dynamics have been found to serve several different functions, including turn-taking control, deictic reference, and attitudes [22]. In a multi-party or situated dialogue, gaze may be an essential means to address a person in a crowd, or pointing to one specific object out of many. These functions have been investigated in experiments and models have been proposed on how to control gaze movements in, for example, robots and embodied conversational agents [23]. However, very little research has been done to investigate the perception of these movements by observers, especially in situated settings.

(a) (b) (c)

Fig. 4. Snapshots of (a): the human condition. (b) Frontal viewing of Furhat moving only the eyes. (c) Frontal viewing of Furhat moving the head instead of the eyes.

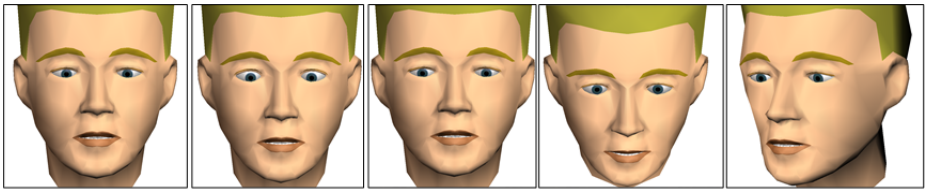
Several studies have explored situated human-robot interaction, where the interlocutors sit around a table with objects that can be referred to, thus constituting a shared space of attention [24,25,26]. These studies provide significant and important findings on the advantageous role of gaze on interaction. However, since they focus on more high level effects of gaze on the overall interaction, they do not provide enough detail and resolution to study and quantify the precision and properties of the perception by humans of the gaze direction generated by robots.

In this study, we explore how accurately humans can perceive the target point in space of the gaze of Furhat in a setting that is typical of situated interaction. The methodology we have employed here is useful to calculate a psychometric gaze function to convey the intended gaze target. But it can also be used to evaluate the gaze perception accuracy of different eye designs and gaze cues in robots.

3.1. Method

We employ in this experiment a configuration that to some degree simulates a situated human-robot task-oriented setup. In the setup, the human and Furhat sit around a squared table that is assumed to be the center of attention of Furhat. In theory, the robot is hypothesized to be looking at areas of interest on the table, such as physical objects. However, in order to be able to make high resolution quantifications of the perception of gaze targets, the table was covered by a squared grid (partly shown in Figure 4).

The experiment is designed so that Furhat and the human will sit at two different sides of the table, and the rotation of the head or the eyes of Furhat is controlled to look at different parts of the grid. The subjects are instructed to place a small object (a glass disk, shown in Figure 4) on the square where they best believe the robot is gazing at. In order to cover the different areas of the grid, without generating a very large number of gaze stimuli, the grid was divided into 9 equally sized virtual regions (split into 3 rows and 3 columns), and Furhat would choose a random square from each of the regions to look at.



(a) NORMAL (b) STATICEYELIDS (c) NOVERGENCE (d) NECK (e) SIDE

Fig. 5. Example photos of the different conditions. The snapshots were taken from the same animated agent used in Furhat but were taken using a flat screen, for better visibility.

In order to automate the experiment, and to increase the level of interaction, Furhat prompted each answer by moving the gaze to a new target and by saying something like: "where am I looking now?"

The subject would then place the tagging object on the perceived target square and verbally dictate the number of the square to the experiment instructor who in turn entered the number into the system, which then triggered the next stimulus. Although intuitive, it is important to note that in such a setup, it is expected that it is more difficult to precisely detect the target square on the grid the further it is from the eyes of the robot, because the same rotation in degrees of the eyes would cover a larger area if the robot is looking towards the further edge of the grid compared to looking towards the closer edge.

A geometric computer model of the setup was implemented, taking into account the design of the head, the mask, the neck, and the physical dimensions and placements of the table and the grid. The model was used to estimate the required horizontal and vertical rotations of the eyes and/or the neck of Furhat, to gaze at any given square on the grid. Although this model is estimated as accurately as possible, and has taken into account as many details of the setup as possible, it should be noted that this is only an estimate of the exact angle between the eyes of Furhat and the focal point. However, since this model is used for all the experimental conditions in the experiment, it is not the exact location of the perceived target points of the subjects that is under interest, but rather the differences in the perceived direction of gaze across the conditions.

6 different conditions (described below) were presented in random order to 18 subjects. All subjects had normal, or corrected to normal vision. Each condition consisted of 9 stimuli gaze points, each corresponding to one of the nine regions on the grid (divided by three rows columns). Thus, each subject answered to 54 stimuli points (9 stimuli * 6 conditions), and every condition received 162 answers (9 stimuli * 18 subjects).

3.2. Conditions

Figure 5 shows snapshots of some of the conditions taken from the same animated agent used in Furhat

1 - NORMAL: This is the baseline condition against which we compare all other condition. In this condition, Furhat only used the gaze to shift between targets (the head always tilted 15 degrees down, facing straight ahead). The subject was sitting in front of Furhat. We also employed two different techniques to enhance the perception of gaze, which are not always used in robotics and animated agents. First, the distance of the target was taken into account when rotating the eyes so that they were not looking in parallel. The eyes of a human will be looking in parallel lines (both equally rotated) only when gazing at infinity. Humans focus on spatial points in space by rotating the eyes differently, so that both eyes are gazing the focal target, and hence, depending on the distance of the focus target, the eyes will rotate with two different angles – a phenomenon known as *vergence*. However, to our knowledge, it is not yet known whether humans can and do take advantage of this cue in order to more precisely locate the target other humans are gazing at. Second, we implemented dynamic eyelids that move along with the vertical movement of the eyes. When the human eyes move vertically, the upper eyelids move so that they are localized exactly on top of the iris. This property of the eyes, although very easy to employ, has mainly been ignored in the design and synthesis of animated agents and robots [27]. It is not trivial to investigate these factors using human subjects, but the current setup, using an animated face, allows us to do that due to the easy control of the animated models.

2 - STATICEYELIDS: This condition was exactly the same as NORMAL, except that the eyelids did not dynamically move with the vertical movement of the eyes.

3 - NOVERGENCE: This condition was exactly the same as NORMAL, except that the eyes always looked in parallel.

4 - NECK: This condition was exactly the same as NORMAL, except that the eyes were always centered, and the neck was used to shift the gaze target.

5 - SIDE: This condition was exactly the same as NORMAL, except that the subjects were seated at a 45 degrees angle from the head, hence having a side view of the eyes.

6 - HUMAN: In this condition, we explore the gaze perception accuracy when observing a human. This condition is important as a baseline of the perceptual granularity of gaze direction in a human-human setting. A human was seated opposite to the subject, at a similar height and distance from the table as Furhat (see Figure 4). During this condition, the human agent always sustained a frontal head pose, while looking at the different target points on the grid with the eyes only, which made the condition similar to the NORMAL condition. Note that in this condition the focal point is the actual target the human is told to look at – it is not calculated using the geometric computer model described above.

3.3. Analysis and Results

The answers to all the gaze stimuli were transformed from the numbers of the squares on the grid, into their physical rotation angles from the eyes of Furhat, and into centimetres distance from Furhat (on the table plane), for both the horizontal and vertical axes (*x* will correspond to the horizontal dimension of the grid from the point of view of Furhat, and *y* will correspond to the vertical dimension of the grid from the point of view of Furhat).

Figure 6 shows a bird's-eye view of the results for all conditions except NOVERGENCE, since it did not deviate significantly from the NORMAL condition (as we will come back to). The dots in the plot represent the focal points, while the ellipses represent the estimate distributions of the answers of the subjects for each of the 9 regions. The horizontal radius corresponds to one standard deviation along the *y* axis, and the vertical radius corresponds to one standard deviation on the *x* axis. The distance on either axes between the center of the ellipses and the dots equals the shift between the average location of the perceived gaze and the focal point. Table 2 shows the averages of the answers for all subjects and regions for both axes, calculated in degrees. The left columns show the distance between the average answers and the focal points (positive values meaning exaggerated answers), and thus correspond to the distance between the center of the ellipses and the dots in Figure 6. The right columns show the average distance between each answer and the mean of the distribution, thus corresponding to the size of the ellipses in Figure 6. All values were compared against the NORMAL condition and tested for significant difference (as indicated by the shades of the cells), using two-tailed paired t-tests ($p<0.05$; dF=161). It is important to note that the left columns only show the shift in the location between the perceived gaze and the focal point – they do not indicate the precision of the gaze perception. However, they do indicate how the conditions affect the location of the perceived direction of gaze, and thus the need for calibration. On the other hand, the right columns provide a measure of the agreement between subjects. Thus, lower values in these columns indicate a higher precision.

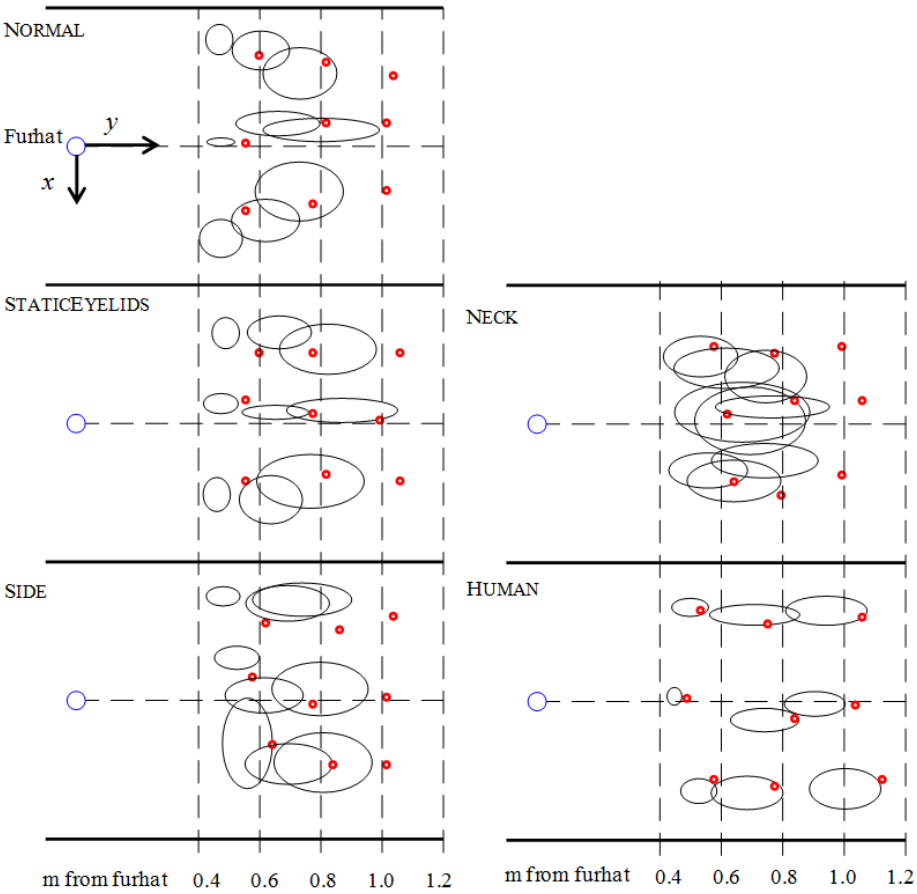


Fig. 6. Distributions (standard deviations) of the answers (ellipses) and focal points (dots), plotted per condition and region.

Table 2. Averages of the answers (in degrees) for all subjects and regions, for both axes. Grey cells indicate significant differences from the NORMAL condition ($p < 0.05$).

	Diff from focal point		**Abs. diff from the perceived mean**	
	x	***y***	***x***	***y***
NORMAL	6.07	-8.71	3.66	4.44
SIDE	5.48	-7.61	4.79	5.10
STATICEYELIDS	5.82	-7.24	3.48	4.27
NOVERGENCE	6.40	-8.36	4.06	4.67
HUMAN	2.79	-3.54	3.18	3.13
NECK	-1.08	-6.44	3.24	5.86

The results indicate that the perception of the human gaze has a higher precision (at least on the x axis), as compared to Furhat (regardless of condition). The mean deviation of about 3 degrees corresponds surprisingly well with the findings on the perception of human gaze reported in [28]. It is a bit surprising to see that the target was in general

perceived as being about 3.5 degrees below the actual target and 2.8 degrees to the side of the target. Since humans should have a lot of experience in perceiving each other's gaze, there is no obvious explanation for this. The HUMAN condition can be regarded as a golden standard for the design of robot heads. One should bear in mind, however, that only one human was used to generate gaze targets. The accuracy and precision of perceiving the direction of human eyes might be different across different humans, possibly depending on the shape and color of the eye; a question that mandates further research to investigate.

Using dynamic eyelids (i.e. lowering the eyelids in relation to the vertical position of the eyes) does not seem to affect the precision of the gaze perception. However, it does affect the perception of the target location, lowering it down, in average 1.47 degrees (difference between the NORMAL condition and the STATICEYELIDS condition over the *y* axis). This is important to take into account when calibrating the psychometric gaze function.

As can be seen in Figure 6, looking at Furhat's gaze from a 45 degrees angle skews the perceived gaze location in an asymmetrical way. This affects the perceived target's location, as well as the precision, which becomes lower on the x-axis (4.79 degrees vs. 3.66 degrees). One possible explanation for this is that the visibility of both eyes is worse when looking from the side, especially when the head looks at the farther side of the table from where the subject is sitting (as seen in Figure 6). This means that, in the case of Furhat, the gaze function should ideally be calibrated differently depending on where the observer is sitting.

No significant differences were found between modeling vergence (different eye rotations) vs. not modeling it (parallel eyes), neither in displacement nor in precision. There are several possible explanations for this. The spatial dimensions of the grid area could be too small to affect large perceivably differences between the relative rotations of both eyes. Another possibility is that the design of Furhat's eyes is not accurate enough to account for these subtle differences. A third possibility is that humans do not utilize vergence when determining gaze direction.

The main difference between the NECK and the NORMAL condition is that the precision on the *y*-axis in the perception of where Furhat is looking when moving the neck is significantly worse than when moving the eyes, which is strikingly obvious when looking at Figure 6. However, the precision on the x-axis is not affected at all (it is actually the condition that is closest to the HUMAN condition). Also, the displacement on the x-axis is the least of all conditions. Thus, perception of horizontal head orientation seems to be very accurate, while perception of vertical head orientation is very inaccurate. We do not have a good explanation of this, but it might have to do with the fact that the head is horizontally symmetrical, but vertically asymmetrical.

3.4. Discussion

While the accuracy of the HUMAN condition is better than the other conditions (at least for the *y*-axis), the difference is not very big, even for the SIDE condition. This shows that there is virtually no Mona Lisa effect when observing Furhat.

The findings presented here have several implications for the design of robots that are supposed to take part in situated interaction with humans. When discriminating between objects on the x-axis, it might be useful to use head orientation, while discrimination on the y-axis might best be done using gaze. Another important implication is that it seems like different psychometric calibration functions should be used depending on the design of the robot, for example depending on whether the eyelids are dynamic and move with vertical gaze shifts or not. The possibility to do this is of course dependent on whether the eyes of the robot are actually used for robot vision (which is not the case with Furhat). Another complicating factor is that the psychometric function seems to be different depending on the viewing angle. Should the system take this into account and compensate for this? This question becomes even more challenging if we take into account that even human gaze would ideally need some calibration to convey the intended target. Thus, the designer has to decide whether the robot should mimic human behavior as much as possible, or maximize the correspondence between intended and perceived target.

Of course, the study also gives rise to several new questions that call for further investigation. There are many combinations of conditions that we haven't tested here. For example, does the perception of human head pose follow the same patterns as we saw here with Furhat (accurate on the *x*-axis but inaccurate on the *y*-axis)? Does side-viewing of human gaze show the same asymmetrical patterns that we saw with Furhat? There are also other individual factors that we haven't explored here. For example, it has been shown in studies of perception of human gaze that the distance of the perceiver may affect the perceived target [29]. Interestingly, they also found that the eyebrows play a role – it seems like humans may use the eyebrows to expand or restrict the availability of their gaze direction to others. The experimental setting presented here could be used to test if similar patterns can be replicated in human-robot interaction.

4. Study III: Effect of Gaze on Turn-taking in Multiparty Dialog:

The previous study investigates a situated interaction setup that is popular for human-robot interaction studies, a setup that might be involved in applications of human-robot interaction such as education and collaborative task solving. However, in human-robot interactive setups, robots are not only expected to look at points of interest in the shared space of attention, but also communicate naturally with the users, a process that involves multiparty dialogue regulation using cues such as gaze, mutual gaze and head direction.

As discussed in the introduction of this article, we have shown in previous studies that the perception of gaze on flat displays is affected by the Mona Lisa effect. However, there are other studies which indicate that it may still be possible to have a multi-party

dialog with an agent presented on a 2D display. In [30], a virtual receptionist is presented, which is able to communicate with multiple interlocutors, and to address them individually using gaze. The system uses a flat screen for the animated head, which would theoretically give rise to the Mona Lisa effect. Still, experiments on the use of gaze for controlling turn-taking in a multi-party conversational setting shows that the system may successfully address the different users to some extent. An accuracy of 86.2% between intended addressee and the next person to speak is reported [30]. However, it is not clear to what extent the results are due to the fact that there were only three users, which would make it possible for them to learn and infer the intended direction of gaze, with the cost of an extra cognitive effort. Also, it is not clear whether only the gaze was shifted, or whether also head pose was used (as was indicated in supplemented video material), which may ease the task for the subjects.

In this study, we explore the interactional effects of gaze in a multi-party conversational setting, and investigate the difference between 2D and 3D manifestations of the animated agent. Unlike our previous perception experiment [12], which focused on the *perceived* gaze, this experiment will investigate how gaze may affect the turn-taking *behavior* of the subjects. Thus, it may be possible for the subjects to infer the intended gaze (which may however require extra cognitive resources). Another difference is that the subjects in the perception experiment did not share their vote on where the projected head was looking. In this experiment, the subjects' decisions may affect each other and confusion about the gaze target may have effects on the fluency of the interaction.

4.1. Method

We used an experimental setup similar to the previous experiment reported in [12]. 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 (180 cm from the projection surface, and on a 26 degrees rotation from each other), as illustrated in Figure 7. The agent addressed the subjects by directing its gaze in their direction (without any head rotation).

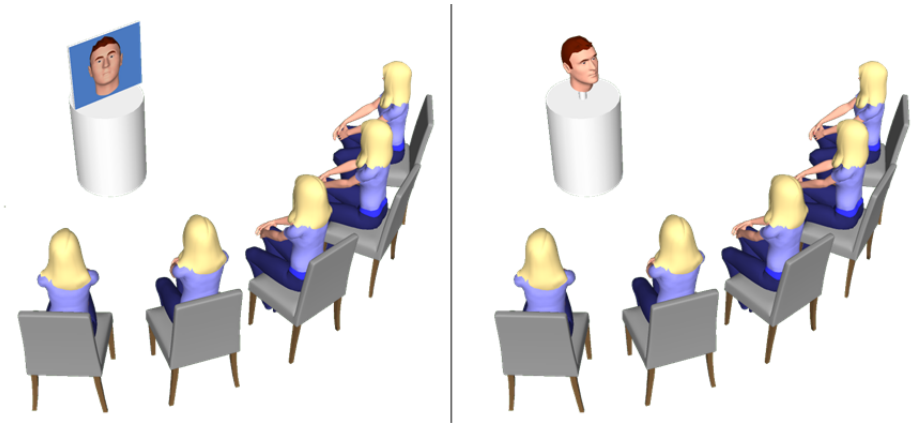


Fig. 7. Schematic setup and placement of the subject and stimuli point in the 2D condition (left) and 3D condition (right).

Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface (using the same 3D computer model), as shown Figure 7. 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 first-person video from a camera navigating around the city of Stockholm (the video was projected on a wall behind the agent), 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. The dialogs were in Swedish. An example dialog is shown in Table 3. Each set of subjects did four dialogs in each 2D/3D condition (i.e. a total of eight videos). The ten subjects were not informed about the purpose of the experiment beforehand. The order of the presentation of the videos and of the conditions was also randomized.

Table 3. The beginning of one of the recorded interactions translated from Swedish.

Agent	*Hi, can you help me to find my way to the Finnish church?*
Subject 1	Yes
Subject 2	Yes
Agent	*Should I continue in this direction?*
Subject 3	Yes, a little bit at least
Agent	*How far should I go then?*
Subject 4	I think it was about 20 meters, maybe a bit longer
Agent	*Is there some landmark I should look for?*
Subject 2	You should look for the Great Church, when you come to Västerlånggatan you will see the Great Church on your left, then should pass it
Agent	*Do you agree?*
Subject 5	Yes

4.2. Analysis and Results

The interactions were logged and the recorded videos were manually annotated. For each answer from a subject, the timing, duration and the subject id of that answer were labeled. Turn takes were labeled to start with the first vocal action made by any of the subjects, which provides a consistent way of labeling turn takes. Using the logged data and the manual labels, the response time was calculated as the time between the gaze shift, and the beginning of the succeeding turn take. The full annotation resulted in 57 turn shifts (question-answer pairs) for the 2D condition, and 56 turn shifts for the 3D condition.

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. This is illustrated in Figure 8. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. By looking at the 2D condition in Figure 8, it is interesting to see that the Mona Lisa effect is present to some extent. This is represented by the fact that all subjects at some point took the turn when the head was looking frontal (gazing at subject 3, who is seated in the middle). This effect is not present in the 3D condition. It is also interesting to see that when the agent was targeting subject 2 and subject 4, the subjects who were seated further away (subject 1 and subject 5 respectively) tended to take the turn. This can be explained by the fact that when the gaze is directed to the side on a 2D surface, the direction is perceived by all subjects to be directed to their side but not at them. However, subject 1 and subject 5 have no more subjects seated further away, which will make them the most likely subjects to actually take the turn.

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. A two sample ANOVA analysis was applied, with the response time as a dependent variable, and the condition as an independent variable. The results show a significant main effect ($F(1)=15.821$; $p<0.001$), with a mean response-time of 1.85 seconds for the 2D condition, and of 1.38 seconds for the 3D condition.

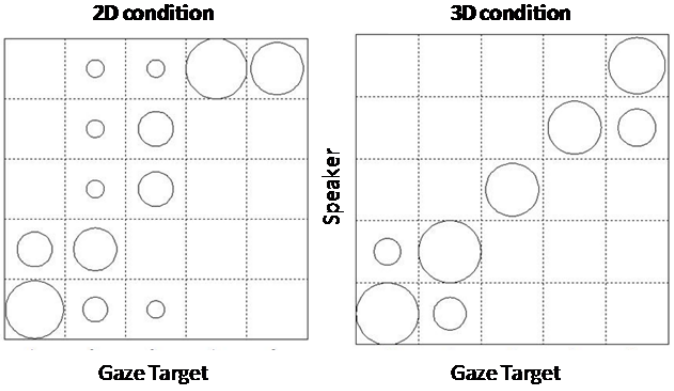


Fig. 8. Confusion between intended addressee (horizontal axis) and the first subject who answered the question (vertical axis).

4.3. Discussion

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 our previous perception experiment in a similar setting [12]. 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 don't ”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 in making this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

5. Multi-party interaction in a public setting

The three studies presented above were all done in controlled lab environments. While such conditions are necessary to investigate the differences between different conditions in a controlled way, it is also interesting to look at what happens in a public settings, where people spontaneously walk up to Furhat and engage in multi-party dialogue, without any instructions. In December 2011, we were 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 Furhat on display at the museum [13]. During the four days of the exhibition, Furhat’s task was to collect information on peoples’ beliefs about the future of robots, in the form of a survey. The exhibition was seen by thousands of visitors, resulting in a corpus of about 10.000 user utterances. This setup allowed us to explore a number of issues in a challenging public environment. First, we wanted to explore to what extent it is possible to obtain information from humans without full understanding, and how this is affected by a multi-party setting. Second, we wanted to verify what we had previously found in controlled experimental settings: that the design of the robot head allows for accurate turn-taking in multi-party interaction (as described in Study III above). Third, we wanted to test IrisTK [34] a new control framework for multi-modal, multi-party interaction.

There are several examples of multimodal dialog systems put to the test in public settings [31,32,33]. Allowing spoken interaction in a public setting is indeed a very challenging task – the system needs to cope with a lot of noise and crowds of people wanting to interact at the same time. To make the task feasible, different restrictions are often applied. One example is the virtual museum guide Max [32], which only allowed written input. Another example is the museum guides Ada and Grace [33], which did not allow the visitors to talk to the agents directly, but instead, used a “handler” who spoke to the system, that is, a person who knew the limitations of the system and “translated” the visitors’ questions. Also, in that system, the dialogue was very simplistic – basically a

mapping of questions to answers independent of any dialog context. What makes the Furhat at Robotville exhibition special, apart from allowing the visitors to talk directly to the system, is that the visitors interacted with the system in a multi-party dialog, allowing several visitors to talk to the system at the same time. While there are examples of systems that have engaged in multi-party dialogue in more controlled settings, such as the virtual receptionist presented in [30], we are not aware of any other multi-party dialogue system put to the test in a public setting, interacting with a large number of users.

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 and when they quit the interaction. To do this, 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. The setup can be seen in Figure 1. As soon as Furhat was approached by a visitor, Furhat immediately took the initiative and started to ask questions, such as "when do you think robots will beat humans in football?". 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. 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. In the case of open questions (as defined above), the addressee answered the question in 92.2% of the cases, with a response time of 1.7 s. The accuracy is similar regardless of whether Furhat was addressing the same speaker as in the turn before (92.6%), or if Furhat had just switched addressee (91.2%). For open questions, the addressee of the previous question answered the open questions in 54.4% of the cases (with a response time of 1.9 s), which indicates that they were indeed perceived as addressed to both participants.

Although the settings are not exactly the same, it is interesting to compare the turn-taking accuracy of 92.2% in a public setting to figures reported from more controlled experiments. In the study on a virtual receptionist reported in [30], an animated head on a 2D screen interacted with three users and gained an accuracy of 86.2%. In study III reported above, the projected 3D head interacted with five users and gained an accuracy of 84% (and a response time of 1.38 s), while a 2D head only gained 50% accuracy (and a response time of 1.85 s).

6. Conclusions

In this article, we have presented some initial experiments that we have done with Furhat. Taken together, they indicate some advantages of the back-projection technique, not only over mechanical humanoid heads, for their simplicity and expressivity, but also as compared to animated avatars on flat surfaces. The animation provides accurate lip and eye movements which facilitate face-to-face interaction between humans and robots. For future studies, we intend to compare the back-projection technique with a mechanical head, where it might be harder to precisely control for lip movements.

Study I investigated the differences in audio-visual intelligibility (lip readability) of Furhat compared to its in-screen counterpart. The results are promising, and validate the suitability of the head as an alternative to animated avatars displayed on flat surfaces. The results also show that people benefit from Furhat in terms of lip reading significantly more than showing the same model on a flat display. This is indeed interesting, and motivates future work to investigate the sources of these differences. In study II, we investigated how the perception of Furhat's gaze in a situated interaction setting is affected by various factors. Similar to the lip reading experiment, the accuracy of Furhat is very close to that of a human. Study III showed that the 3D design of Furhat provides a better turn-taking accuracy in a multi-party dialogue, as compared to an animated agent on a flat display, due to the Mona Lisa effect. The turn-taking accuracy was also validated in a public setting in a museum where Furhat interacted with thousands of users through multi-party dialogue. The multiparty multimodal dialogue used at London has been since then extended and tried successfully at other venues (e.g. [35, 36]).

Furhat is an example of a technology that is gaining ground for building expressive and natural humanoid heads. The technology, as in Furhat, makes robotic heads more accessible to research labs, even for those with no experience or interest in robotics hardware, thanks to the fact that very little is indeed controlled mechanically (aside from the neck). This also provides simulation software of the robotic face that is identical to its physical counterpart. The use of software also allows for a large set of possible customizations with very little effort, for example, the colors and the design of the different parts of the face (the size and colors of the eyes, eyebrows, iris, pupil, etc.). The facial animation model used in Furhat also contains a model of the tongue; hence Furhat has a tongue that is automatically controlled using the lip synchronization software, a property that is not easily employable in mechatronic heads. In addition to this, the animation also provides a large variability and complexity of facial expressions.

We believe that Furhat as a research tool will provide wider accessibility to a larger array of researchers and laboratories interested in face-to-face human-robot interaction research and development. In this article, we have presented examples of such potential research made possible by this technology.

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 preparing the 3D mask model for printing, and to thank Jens Edlund, Joakim Gustafson, Björn Granström and Preben Wik for their interest and inspiring discussions, and to all the subjects who took part in the studies.

References

1. D. Massaro, *Perceiving talking faces: from speech perception to a behavioral principle,* (MIT Press, Cambridge, 1998).
2. J. Cassel, J. Sullivan, S. Prevost and E.E. Churchill, *Embodied Conversational Agents* (MIT Press, 2000).
3. M. Naimark, Two unusual projection spaces, *Presence: Teleoperators and Virtual Environments* **14**(5) (2005) 597-506.
4. S. Morishima, T. Yotsukura, K. Binsted, F. Nielsen and C. Pinhanez, HyperMask: talking head projected onto real objects, *The Visual Computer,* **18**(2) (2002) 111-120.
5. M. Hashimoto and D. Morooka, Robotic facial expression using a curved surface display, *Journal of Robotics and Mechatronics,* **18**(4) (2006) 504-505.
6. F. Delaunay, J. de Greeff and T. Belpaeme, Towards retro-projected robot faces: an alternative to mechatronic and android faces, in *Proc. of the International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Toyama, Japan, 2009).
7. T. Kuratate, Y. Matsusaka, B. Pierce and G. Cheng, Mask-bot: a life-size robot head using talking head animation for human-robot communication, in *Proc. of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)* (2011) (pp. 99-104).
8. S. Al Moubayed, J. Beskow, G. Skantze and B. Granström, Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction, in *Esposito, A. Esposito, A. Vinciarelli, A. Hoffmann, R. and C. Müller, (Eds.), Cognitive Behavioural Systems,* (Lecture Notes in Computer Science. Springer, 2012).
9. C. Siciliano, A. Faulkner and G. Williams, Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners, in *AVSP 2003-International Conference on Audio-Visual Speech Processing*, (2003).
10. G. Salvi, J. Beskow, S. Al Moubayed, and B. Granström, SynFace—Speech-Driven Facial Animation for Virtual Speech-Reading Support, *EURASIP Journal on Audio, Speech, and Music Processing*, (2009).
11. S. Al Moubayed, J. Beskow and B. Granström, Auditory-Visual Prominence: From Intelligibility to Behavior, *Journal on Multimodal User Interfaces*, **3**(4) (2010) 299-311.
12. S. Al Moubayed, J. Edlund and J. Beskow, Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections, *ACM Transactions on Interactive Intelligent Systems*, **1**(2) (2012) 25.
13. G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow and B. Granström, Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue, in *Proceedings of IVA-RCVA.* (Santa Cruz, CA, 2012).

14. Q. Summerfield, Lipreading and audio-visual speech perception, *Philosophical Transactions: Biological Sciences*, **335**(1273) (1992) 71–78.
15. H. McGurk and J. MacDonald, (1976) Hearing lips and seeing voices, *Nature*, **264**(5588), 746–748.
16. J. Beskow, Rule-based visual speech synthesis, in *Pardo, J. (Ed.), Proc. of the 4th European Conference on Speech Communication and Technology (EUROSPEECH«95),* (1995) (pp. 299-302).
17. D. Massaro, J. Beskow, M. Cohen, C. Fry and T. Rodriguez, Picture my voice: Audio to visual speech synthesis using artificial neural networks, in *Proc. of AVSP 99,* (1999) (pp. 133-138).
18. T. Ezzat, G. Geiger and T. Poggio, Trainable videorealistic speech animation, in *Proceedings of the 29th Conference on Computer Graphics and Interactive Techniques*, (New York, USA, 2002), pp. 388–398.
19. N. Erber, Effects of angle, distance and illumination on visual reception of speech by profoundly deaf children, *Journal of Speech and Hearing Research*, **17** (1974) 99-112.
20. M. Argyle, R. Ingham, F. Alkema and M. McCallin, The different functions of gaze, *Semiotica*, **7**(1) (1973) 19-32.
21. P. L. Mirenda, A. M. Donnellan and D. E. Yoder, Gaze behavior: A new look at an old problem, *Journal of Autism and Developmental Disorders*, **13**(4) (1983) 397-409.
22. A. Kendon, (1967). Some functions of gaze direction in social interaction, *Acta Psychologica,* **26**, 22-63.
23. O. Torres, J. Cassell and S. Prevost, Modeling gaze behavior as a function of discourse structure, in *Proc. of the First International Workshop on Human-Computer Conversation*, (1997).
24. J. D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey and J. Ventre-Dominey, I reach faster when I see you look: gaze effects in human-human and human-robot face-to-face cooperation, *Frontiers in neurorobotics*, (2012) 6.
25. Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita and T. Miyamoto, Responsive robot gaze to interaction partner, in *Proceedings of robotics: Science and systems*, (2006).
26. M. Johnson-Roberson, J. Bohg, G. Skantze, J. Gustafson, R. Carlson, B. Rasolzadeh and D. Kragic, Enhanced Visual Scene Understanding through Human-Robot Dialog, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (2011).
27. G. Bailly, S. Raidt and F. Elisei, Gaze, conversational agents and face-to-face communication, *Speech Communication*, **52**(6) (2010) 598-612.
28. J. J. Gibson and D. D. Pick, Perception of another person's looking behaviour, *The American journal of psychology*, **76**(3) (1963) 386-394.
29. R. Watt, B. Craven and S. Quinn, A role for eyebrows in regulating the visibility of eye gaze direction, *The Quarterly Journal of Experimental Psychology*, **60**(9) (1963) 1169-1177.
30. D. Bohus and E. Horvitz, Facilitating multiparty dialog with gaze, gesture, and speech, in *Proc. ICMI'10*, (Beijing, China, 2010).

31. J. Gustafson, *Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction*, Doctoral dissertation, (KTH, Department of Speech, Music and Hearing, KTH, Stockholm, 2002).
32. S. Kopp, L. Gesellensetter, N. Krämer and I. Wachsmuth, A conversational agent as museum guide - design and evaluation of a real-world application, in *Proceedings of IVA 2005, International Conference on Intelligent Virtual Agents,* (Berlin: Springer-Verlag, 2005).
33. W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, K. Bronnenkant, J. Williams, A. Leuski, S. Narayanan, D. Piepol, C. Lane, J. Morie, P. Aggarwal, M. Liewer, J. Y. Chiang, J. Gerten, S. Chu and K. White, Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides, in *Proc. of the 10th International Conference on Intelligent Virtual Agents (IVA)*, (Berlin: Springer-Verlag, 2010).
34. G. Skantze and S. Al Moubayed, IrisTK: a statechart-based toolkit for multi-party face-to-face interaction, in *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*, (Santa Monica, CA, USA, 2012).
35. S. Al Moubayed, G. Skantze, J. Beskow, K. Stefanov and J. Gustafson, Multimodal Multi-party Social Interaction with the Furhat Head, in *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*, (Santa Monica, CA, USA. 2012).
36. S. Al Moubayed, Bringing the avatar to life*. Studies and developments in facial communication for virtual agents and robots,* Doctoral dissertation, (School of Computer Science, KTH Royal Institute of Technology, 2012), p75-78.

Samer Al Moubayed is a postdoctoral researcher at the Department of Speech, Music and Hearing, at KTH, Stockholm, Sweden. He received his PhD from KTH in 2012, for his studies and developments on human-robot face-to-face interaction. Samer has been part of several EU projects, including H@H, MonAMI, and IURO. His main work and interest are embodied dialogue systems, multimodal synthesis, and nonverbal social signal processing.

Gabriel Skantze is a senior researcher (Docent) in speech technology at the Department of Speech Music and Hearing, KTH, Stockholm, Sweden. He holds a PhD in speech communication from KTH. During his studies he specialized in error handling and miscommunication in spoken dialogue systems. His current research focus is on real-time models of spoken dialogue and empirical studies of human-robot interaction. He has participated in numerous EU projects related to dialogue systems and robotics, including CHIL, MonAMI and IURO. He is currently the principal investigator of a nationally funded project in the area of multimodal incremental dialogue processing.

Jonas Beskow is an associate professor in speech technology and communication, with main research interests in the area of audio-visual speech synthesis, talking avatars and virtually- and physically embodied conversational agents. He has participated in numerous EU projects related to multimodal speech technology in human-machine interaction and accessibility applications, including PF-STAR, SYNFACE, CHIL, MonAMI, HaH, IURO and LipRead. He is currently the principal investigator of two nationally funded projects in the area of sign language and gesture in face-to-face interaction. He is involved in two start-up companies in the domain of talking avatars, and is one of the developers of the open source speech processing tool WaveSurfer.