Lip-reading Furhat: Audio Visual Intelligibility of a Back Projected Animated Face

Samer Al Moubayed, Gabriel Skantze, Jonas Beskow

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

Abstract. Back projecting a computer animated face, onto a three dimensional static physical model of a face, is a promising technology that is gaining ground as a solution to building situated, flexible and human-like robot heads. In this paper, we first briefly describe *Furhat*, a back projected robot head built for the purpose of multimodal multiparty human-machine interaction, and its benefits over virtual characters and robotic heads; and then motivate the need to investigating the contribution to speech intelligibility *Furhat*'s face offers. We present an audio-visual speech intelligibility experiment, in which 10 subjects listened to short sentences with degraded speech signal. The experiment compares the gain in intelligibility between lip reading a face visualized on a 2D screen compared to a 3D back-projected face and from different viewing angles. The results show that the audio-visual speech intelligibility holds when the avatar is projected onto a static face model (in the case of *Furhat*), and even, rather surprisingly, exceeds it. This means that despite the movement limitations back projected animated face models bring about; their audio visual speech intelligibility is equal, or even higher, compared to the same models shown on flat displays. At the end of the paper we discuss several hypotheses on how to interpret the results, and motivate future investigations to better explore the characteristics of visual speech perception 3D projected faces.

Keywords: Furhat, Talking Head, Robot Heads, Lip reading, Visual Speech.

1 Introduction

During the last two decades, there has been on-going research and advancement in facial animation. Researchers have been developing human-like talking heads that can engage in human-like interactions with humans (Beskow et al. 2010) realize realistic facial expressions (Gratch et al. 2006; Ruttkay & Pelachaud, 2004), express emotions (Pelachaud, 2009; De Melo & Gratch, 2009) and communicate behaviours (Granström & House, 2007; Gustafson et al. 2005; Kopp et al. 2005). In addition to the human-like nonverbal behaviour implemented in these heads, research has also taken advantage of the strong relation between lip movements and the speech signal, building talking heads that can enhance speech comprehension when used in noisy environments or as a hearing aid (Massaro, 1998; Salvi et al. 2009).

Several talking heads are made to represent personas embodied in 3D facial designs (referred to as ECAs, Embodied Conversational Agents (Cassel et al. 2000)) simulating human behaviour and establishing interaction and conversation with a human interlocutor. Although these characters have been embodied in human-like 3D animated models, this embodiment has almost always been displayed using two dimensional display (e.g. flat screens, wall projections, etc.) having no shared access to the three dimensional environment where the interaction is taking place. 2D displays come with several illusions and effects, such as the *Mona Lisa gaze effect*. For a review on these effects, refer to (Todorovi, 2006).

In robotics on the other hand, the accurate and highly subtle and complicated control of digital computer models (such as eyes, eye-lids, wrinkles, lips, etc.) does not easily map onto mechanically controlled heads. Such computer models require very delicate, smooth, and fast control of the motors, appearance and texture of a mechanical head. In addition to that, mechatronic robotic heads, in general, are significantly heavier, noisier and demand more energy and maintenance compared to their digital counterpart, while they are more expensive and exclusive.

To bring the talking head out of the 2D display, and into the physical situated space, we have built *Furhat* (Al Moubayed et al, 2012a). *Furhat* is a hybrid solution between animated faces and robotic heads. This is achieved by projecting the animated face, using a micro projector, on a three dimensional plastic mask of a face. This approach has been shown to deliver accurate situated gaze that can be used in multi-party dialogue (Al Moubayed et al, 2012b; Edlund et al. 2011). It has also been shown to accurately regulate and speed up turn-taking patterns in multiparty dialogue (Al Moubayed & Skantze, 2011). *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. Siciliano et al. 2003; Salvi et al. 2010; Beskow et al. 2010). Figure 1 shows several snapshots of the *Furhat* head.

The question we address in this work is whether this solution comes with negative effects on the readability of the lips. Since the mask is static, jaw and lip movements are merely optical and might not be perceived as accurately as in a flat display due to that the physical surface they are projected onto (the jaw and lips) is not moving according to their movements. The other question is whether the contribution of the lip movements to speech intelligibility is affected by the viewing angle of the face. In 2D displays, the visibility of the lips is not dependent on the viewing angle of the screen (the location of the looker in relation to the screen), due to the Mona Lisa effect, and hence, if there is an optimal orientation of the face, it can be maintained throughout the interaction with humans, something that cannot be established with 3D physically situated heads.

2 Lip-Reading and Speech Perception – Evaluation of Furhat

One of the first steps in building a believable, realistic talking head is to animate its lips in synchrony with the speech signal it’s supposed to be producing. This is done not only to enhance the illusion that the talking head itself is the source of the sound

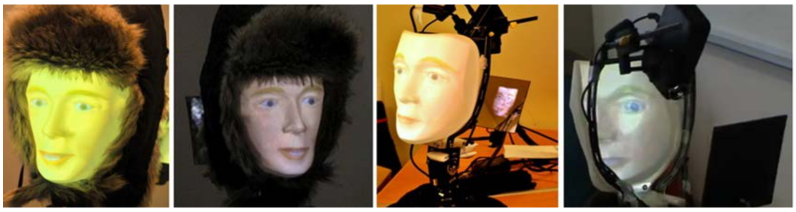


Fig. 1. photos of how Furhat looks like from the inside and outside.

signal the system is communicating (rather than a separate process), but also for the crucial role lip movements play in speech perception and comprehension.

The visible parts of the human vocal tract carry direct information about the sounds the vocal tract is producing. The information the lips carry can be perceived by humans and hence help communicate the information in the speech signal itself (Summerfield, 1992; McGurk & McDonald, 1976). These important advantages of lip movements have been taken into account since the early developments on talking heads, and different models and techniques have been proposed and successfully applied to animate and synchronize the lips with the speech signal itself as input (e.g. Beskow, 1995; Massaro et al. 1999; Ezzat & Poggio, 2000).

However, when it comes to Furhat: Furhat's plastic mask itself is static, although the projected image on Furhat is animated, the fact that the mask itself is static might introduce inconsistency and non-alignment between the projected image and the projection surface, and so the fact that the animated lips do contribute to speech perception does not need to naturally hold with Furhat. The following study presents an experiment comparing audiovisual speech intelligibility of Furhat against the same animated face that is used in Furhat but visualized on a traditional flat display.

3 Lip-Reading Experiment

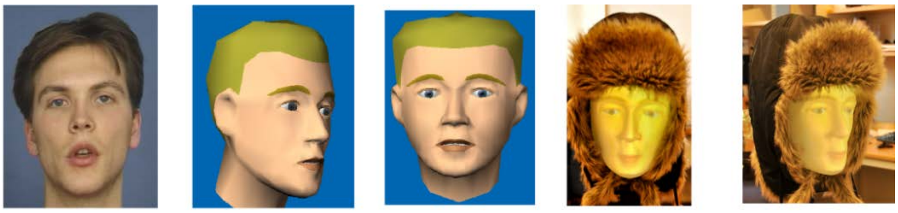
The setup used in this experiment introduces subjects to acoustically degraded sentences, where the content of the acoustic sentence is partially intelligible when listening only to the audio. The sentences are then enriched by a lip-synchronized talking head to increase their intelligibility. The sentences are presented in different experimental conditions (6 in total) and the perceived intelligibility of the sentence (the number of correct words recognized) is compared across conditions.

In the experiment, the audiovisual stimuli consisted of a collection of short and simple Swedish sentences, which vary in length between three to six words, with a basic everyday content. e.g., *"Den gamla raven var slug"* (The old fox was cunning).

The audio-visual intelligibility of each sentence was calculated as *the number of words correctly recognized, divided by the number of content words in the sentence*.

The speech files were force-aligned using an HMM aligner (Sjolander, 2003) to guide the talking head lip movement using the phonetic labelling of the audio file.

The audio signal was processed using a 2-channel noise excited vocoder (Shannon et al. 1995) to reduce intelligibility. This vocoder applies band-pass filtering and replaces the spectral details in the specified frequency ranges with white noise. The



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

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

number of channels was decided after a pilot test to ensure an intelligibility rate between 25% and 75%, as to avoid any floor or ceiling recognition rate effects.

The stimuli were grouped into a set of 15 sentences per group and every set was only used for one condition. The groups were randomly matched to the conditions for each speaker in order to avoid interaction effects between the sentence difficulty and the condition. As a result, each subject was introduced to all the conditions, but was never introduced to the same stimulus more than once. At the beginning of the experiment, one set was always used as training and only in audio mode, as to avoid any training effects during the experiment. During training, subjects were allowed to listen to the degraded audio file as many times as they wished, and feedback was given to them with the correct content of the audio sentence.

3.1 Conditions

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

1. ***Audio Only:*** In the audio-only condition, subjects were listened to the acoustically degraded sentences without any visual stimuli.
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 is seated in front of the screen, looking straight at the talking head. The talking head in the screen is oriented to look frontal (0 degrees rotation inside the screen), and hence the name *Screen0°*.
3. ***Furhat0°: Furhat viewed at 0° angle:*** In this condition the sentences were presented to the subject with the animated model and 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°* (see further, condition 5).
5. ***Furhat45°: Furhat viewed at 45° angle:*** This condition is identical to *Furhat0°*, except that subjects were seated at a 45° 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.

The conditions were systematically permutated among 10 normal hearing subjects, with normal or corrected to normal vision. All subjects were native speakers of Swedish. During the experiments, the subjects were introduced to all conditions but never to the same sentence twice. Since every condition contained 15 sentences, this resulted with 900 stimuli in total for the experiment (6 condition * 15 sentences * 10 subjects), with every condition receiving 150 sentence stimuli. Subjects were given a cinema ticket for their participation, and the experiment took ~25 minutes/ subject.

4 Analysis and Results

An 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)=21.890, p<.0001]. The mean accuracy for each condition is shows in Figure 3, along with the standard error bars.

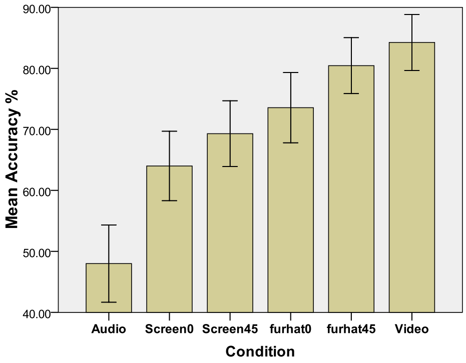


Fig. 3. The average percentage accuracy rates for the different experimental conditions.

A post-hoc LSD analysis was carried out to measure the significance values between the accuracy rates of each of the conditions. The p values for the conditions according to the means are shown in Table 1. All other combinations not included in the table are significantly different from each other.

The results firstly show that the Screen0 condition (and all other conditions), provide an audio visual intelligibility that is significantly higher than the audio condition alone. The results also show that there is no significant difference in the audio-visual intelligibility of the face being looked at either frontal or at a 45° (no significant difference between Screen0 and Screen45, or between Furhat0 and Furhat45). The results show that the Mona Lisa effect would not benefit the audio-visual intelligibility of the face using a flat over a spatially situated head, at least not between 0 and 45° rotation angles.

More importantly, the results show that there is no loss in the audio-visual

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

Condition1		Condition2	*p*-value
Screen0	*	Screen45	.167
Screen45	*	Furhat0	.266
Furhat0	*	Furhat45	.079
Furhat45	*	Video	.335
All other		**combinations**	**< *.01***

intelligibility when using the Furhat‘s physically-static mask as a projection surface compared to using a flat screen, for either 0 or 45° viewing angles of the face. This shows that the Furhat robot head is a valid alternative to the screen in terms of lip readability, and would be a possible interface to aid human speech perception and comprehension. The more surprising finding is that Furhat, not only does not hinder the audio-visual intelligibility of the animated mask, but rather enhances it significantly over the flat display, and for both viewing angles (the rate is significantly higher for Furhat0 over Screen0, and for Furhat45 over Screen45).

5 Discussion & Conclusions

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 lips. 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 increased intelligibility. Another possibility is that looking at Furhat is cognitively easier than looking at a flat display since it is spatially situated and more human-like than a virtual agent presented on a 2D screen.

A main difference between interacting with a face shown on a 2D or 3D surface is that the 2D surface comes with the Mona Lisa effect. For our experiment, this means that the visibility of the face and lips to a subject standing straight in front of the screen or at an angle is the same, and hence if there is an optimal lip reading angle of a face, the face on a 2D screen can maintain that angle and guarantee optimal intelligibility. This is not the same with a 3D head (a physical object). Obviously, the visibility of the lips depends on where the onlooker is standing in relation to the face, and so if looking at the face with an angle is worse than looking at it straight frontal, this would introduce a variable intelligibility depending on the viewing angle. This is found to be the case when reading human lips. In (Erber, 1974) it was found that the lip-reading contribution drops down when looking beyond 45 degrees, but is not significantly different between 0 and 45 degree.

In conclusion, this study aimed at investigating 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.

Acknowledgments. This work has been partly funded by the EU project IURO (Interactive Urban Robot) No. 248314.

References

1. Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B.: Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito et al. (Eds.), Cognitive Behavioural Systems. Lecture Notes in Computer Science. Springer. (2012).
2. 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).
3. 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).
4. 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).
5. Beskow, J.: “Rule-based visual speech synthesis,” in Proc of the Fourth European Conference on Speech Communication and Technology, (1995).
6. Beskow, J., Edlund, J., Granström, B., Gustafson, J., & House, D.: Face-to-face interaction and the KTH Cooking Show. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., & Nijholt, A. (Eds.), Development of Multimodal Interfaces: Active
7. Listening and Synchrony (pp. 157 - 168). Berlin / Heidelberg: Springer, (2010).
8. Cassel, J., Sullivan, J., Prevost, S., & Churchill, E. E.: Embodied Conversational Agents. MIT Press. (2000)
9. De Melo, C. & Gratch, J. Expression of Emotions Using Wrinkles, Blushing, Sweating and Tears. In Proceedings of the 9th international conference on Intelligent Virtual Agents IVA 2009: 188-200, Amsterdam. The Netherlands. (2009).
10. Edlund, J., Al Moubayed, S., & Beskow, J.: The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. Proceedings of the Intelligent Virtual Agents 10th International Conference (IVA’2011) (pp. 439-440). Reykjavík, Iceland: Springer. (2011).
11. Ezzat, T. & Poggio, T.: Visual Speech Synthesis by Morphing Visemes. Visual speech synthesis by morphing visemes. In K. A. Publishers, editor, International Journal of Computer Vision, volume 38, pages 45--57, (2000)
12. Erber, N.P.: “Effects of angle, distance and illumination on visual reception of speech by profoundly deaf children”. J. of Speech and Hearing Research, 17, 99-112, (1974).
13. 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).
14. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., and Morency, L.: `Virtual Rapport'. In: 6th International Conference on Intelligent Virtual Agents (IVA 2006). Marina del Rey, CA, USA. (2006).
15. Gustafson, J., Boye, J., Fredriksson, M., Johannesson, L., & Königsmann, J.: Providing computer game characters with conversational abilities. In Proceedings of Intelligent Virtual Agent (IVA05). Kos, Greece. (2005)

16. Kopp, S., Gesellensetter, L., Krämer, N., Wachsmuth, I.: A conversational agent as museum guide -- design and evaluation of a real-world application. In: Panayiotopoulos et al. (eds.): Intelligent Virtual Agents, LNAI 3661, Springer-Verlag, Berlin, 329-343. (2005).
17. Kriegel, M., Aylett, R., Cuba, P., Vala, M. & Paiva, A.: Robots meet IVAs: A Mind-Body Interface For Migrating Artificial Intelligent Agents. In proceedings 10th Int. Conf. On Intelligent Virtual Agents IVA'11, Reykjavik, Iceland. (2011).
18. Massaro, D.: Perceiving talking faces: from speech perception to a behavioral principle. MIT Press, Cambridge. A Bradford Book. ISBN: 978-0262133371. (1997)
19. Massaro, D., Beskow, J., Cohen, M., Fry, C., and Rodgriguez, T.: Picture my voice: audio to visual speech synthesis using artificial neural networks. In Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'99, Santa Cruz, USA, (1999).
20. McGurk, H. & MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746 (1976).
21. Pelachaud, C.: Modeling Multimodal Expression of Emotion in a Virtual Agent. Philosophical Transactions of Royal Society B Biological Science, B 2009 364, 3539-3548, (2009).
22. 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).
23. Ruttkay, Z. & Pelachaud, C. (editors): From Brows till Trust: Evaluating Embodied Conversational Agents, Kluwer, (2004).
24. Salvi, G., Beskow, J., Al Moubayed, S., & Granström, B.: SynFace—Speech-Driven Facial Animation for Virtual Speech-Reading Support. EURASIP Journal on Audio, Speech, and Music Processing, (2009).
25. Shannon, R., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M.: "Speech Recognition with primarily temporal cues," Science, vol. 270, no. 5234, p. 303, (1995).
26. Siciliano, C., Williams, G., Beskow, J., and Faulkner, A.: "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired," in Proceedings of the International Congress of Phonetic Sciences, pp. 131–134, (2003).
27. Summerfield, Q. :Lipreading and audio-visual speech perception. Philosophical Transactions: Biological Sciences, vol. 335, no. 1273, pp. 71-78. (1992).
28. Sjolander, K.: "An HMM-based system for automatic segmentation and alignment of speech," in Proceedings of Fonetik, pp. 93–96, (2003).
29. Todorovi, D.: Geometrical basis of perception of gaze direction. *Vision Research, 45*(21), 3549-3562, (2006).