ON TALKING HEADS, SOCIAL ROBOTS AND WHAT THEY CAN TEACH US

Jonas Beskow

Division of Speech, Music and Hearing,
School of Electrical Engineering and Computer Science, KTH, Sweden
beskow@kth.se

ABSTRACT

This paper discusses how animated talking heads and physical social robots can benefit investigations into human spoken communication. It gives an overview of a series of experiments carried out in our group in recent years that highlight how these types of technologies can serve as a tools for controlled experiments in phonetics sciences in order to investigate aspects in visual speech perception and face-to-face interaction. It also discusses the role of agent embodiment, both in terms of how physical situatedness influences perception and interaction, and how different robot/avatar embodiments (e.g. head-only vs full-body humanoid) influence the expressivity and suggests how such effects may be studied and compensated for using stylized or exaggerated motions.

Keywords: talking heads, social robots, multimodal interaction, motion capture, conversation

1. INTRODUCTION

Spoken language- and multimodal interaction technologies are developing at a rapid pace. Currently we see a surge in voice-based technologies deployed on smart speakers, in cars and on our mobile phones. While these disembodied agents work well for short query and command/control style interactions, there are applications in fields such as education, service, retail, health, elderly care, simulation, training and entertainment where these simple voice-only interactions fall short. This is where animated agents and social robots have an important niche to fill. In situations that involve long-term or sustained interaction (e.g. a personal tutor or a companion robot) agents that leverage other modalities such as gaze, gesture and facial expression will serve to increase engagement, attention and robustness in the interaction. But we can also use these embodied interaction technologies as vehicels for more basic scientific discovery. Just as early speech synthesis models paved the way for discoveries in human speech perception, talking heads and visual speech models have been used extensively in research on mechanisms on audiovisual speech, both in terms of perception and production. Taking the paradigm further, the availibility of simulated humans that can take part in human face-to-face interactions gives us the ability to test hypotheses of human communicative behaviour by implementing them in artificial simulations. The words of physicist Richard Feynman, *What I cannot create I do not understand* succinctly captures the idea. In particular, the general scientific approach that governs the work covered in this paper can be described as a form of analysis by synthesis, where a certain communicative phenomenon is first codified as a software algorithm and then evaluated with human subjects (after which the process optionaly iterates). In this paper I describe a series of experiments where talking heads and social robots have been used to gain insight into phenomena within human spoken communicative behaviour.

The remainder of this paper is organized as follows. First we consider talking heads both in virtual and physical form (i.e. robots) and their relative merits and drawbacks. Then we look at how talking heads and robots may be used as tools in studying different aspects of human communicative behaviour in a variety of settings such as audio-visual speech perception, interaction and dialogue. Finally we discuss the role of embodiment for a virtual agent or robot: how does a physical robot differ from a virtual one, or a full-bodied humanoid vs a talking head; how may we measure these differences and how may we compensate for them?

2. VIRTUAL AND PHYSICAL TALKING HEADS

In 1972, Parke created the first parametric 3D model of the human face for computer animation [11], which was an enabler and an inspiration for a generation of reserachers to start building virtual talking heads; 3D facial models controlled by rules or data-driven techniques in order to generate realistic visual speech animation based on linguistic input. These types of models have applications in a large number of settings such as education, training, simulation, entertainment, telepresence and virtual/mixed reality. There are however inherent limitations with pure animation based solutions when it comes to real-world applications, in that the character is confined to the screen it is being displayed on. On the other hand, physical talking heads, in the form of robots, makes it possible for the interaction to be *situated in the same physical space* as the user. This has some very important implications, such as the ability to convey spatial gaze-information[5], cruical for efficient multi-party interaction or join attention. Furthermore, it has been shown that users tend to show more engagement with robots than with virtual characters [7]. There are however some apparent drawbacks with mechatronic faces. These include motor noise, mechanical complexity (leading to high costs and high maintanence), limited motion capacity and limited customizability (not possible to change face identity). We have developed a hybrid solution, the Furhat robot [10], that is a robotic head with movable neck that uses projected animation to display the face. Furhat was concieved out of frustration with the limitations of current social robotics technologies, and it combines the benefit of animation (highly flexible, silent and low-cost) with those of robotics (physically situated, socially engaging), which makes it a suitable platform for investigations in social face-to-face interaction.

3. SPEECH INTELLIGIBILITY AND CO-VERBAL MOTION

When we speak, our faces are in constant motion. Not only the articulators, but also eyebrows, eyelids, eyes, head and neck are involved in highly orchestrated movement.

The contribution of lip movements to speech perception has been thoroughly studied and quantified, e.g. Summerfield [14] showed that the information carried in the face compared to only the acoustic signal can equal up to 11 dB 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. There are practical applications that capitalize on this effect, e.g. in the domain of accessibility and inclusion: in two projects [12], [4], we developed a *virtual speech reading support system* based on animated talking heads driven by speech in real-time. In these projects we carried out intelligibility studies with speech in noise, revealing gains in word-level intelligiblity from an animated talking head close to that obtained with a video of the real talker.

3.1. Intelligibility and Visual Prominence

Given the intelligibility gains obtained from synthetic *articulatory* movements, we were curious to see if other types of of co-verbal behaviour in a talking head would be able to further increase intelligibility of the speech signal. In [9] we decided to focus on *prominence* which can be generally defined as when a linguistic segment is made salient in its context. Prominence is a prosodic function that (1) is known to be of importance for speech intelligibility, (2) has been shown to be strongly correlated with head- and facial movements and (3) possible to approximate from the raw speech signal without understanding of the semantic content. Point (3) was of particular importance to our application (real-time speech reading support) since we did not want to rely on speech-to-text functionality, because it would incur to much latency in the system for fluent conversation, as well as being error prone (especially at the time of this study).

In order to investigate the effects of facial prominence cues, in terms of gestures synthesized on an animated talking head, we conducted a speech intelligibility, where speech was acoustically degraded and the fundamental frequency cues were removed from the signal by means of a noise-excited vocoder. The speech was presented to 12 subjects through a lip synchronized talking head carrying head-nods and/or eyebrows-raising gestures, which were synchronized with syllables carrying auditory prominence. The experiment showed that presenting prominence as facial gestures significantly increased speech intelligibility compared to when these non-verbal gestures were either absent or are added at randomly selected syllables.

We also conducted a follow-up study examining the *perception* of the behavior of the talking heads when gestures are added over prominent syllables. We used gaze tracking technology and questionnaires with 10 moderately hearing impaired subjects who were exposed to audio book material accompanied by a talking face with and without the additional prominence gestures. The results of the gaze data revealed an interesting pattern: when there was no movement in the face apart from the articulation, the gaze pattern focused almost exclusively on the mouth and on areas outside the face. When non-verbal motion was present, gaze-patterns covered mouth and upper face in roughly equal proportions, which is consistent with patterns observed from studies with natural human talking faces. From the questionnaires, it is evident that the gestures significantly increase the naturalness and the understanding of the talking head.

The results show the importance of non-verbal motion in the face and that even very simple rules (adding gestures on prominent syllables) are an effective means of improving both naturalness and intelligibility.

3.2. Animated Lombard Speech

In the presence of acoustic noise, humans seamlessly adapt their speech production and perception strategies in order to compensate for the acoustic external conditions and maintain communication. According to the theory of Hyper-Hypo articulation [8], speakers economize their speech production with the goal to make themselves understood in a particular communicative situation.

In [1] we were interested in how this effect could be exploited for the purposes of accurate and highly intelligible talking face animation. The basic idea was to expose a talker to various noise levels during speech production and measure his articulation using motion capture. We recorded audio, video and facial motion capture data of a talker uttering a set of 180 short sentences, under three conditions: normal speech (in quiet), Lombard speech (in noise), and whispering. We then created an animated 3D avatar with similar shape and appearance as the original talker and used an error minimization procedure to drive the animated version of the talker in a way that matched the original performance as closely as possible. In a perceptual intelligibility study with degraded audio we then compared the animated talker against the real talker and the audio alone, in terms of audio-visual word recognition rate across the three different production conditions. We found that the visual intelligibility of the animated talker was on par with the real talker for the Lombard and whisper conditions. In addition we created two incongruent conditions where normal speech audio was paired with animated Lombard speech or whispering. When compared to the congruent normal speech condition, Lombard animation yields a significant increase in intelligibility, despite the AV-incongruence. In a separate evaluation, we gathered subjective opinions on the different animations, and found that some degree of incongruence was generally accepted.

The key take-home message from this experiment is that the extended articulatory motion exhibited in Lombard speech indeed translates to increased visual intelligibility, even when paired with non-Lombard audio, and varying the acoustic environment during production appears to be a viable method of eliciting varied and natural visual speech data.

3.3. Speechreading in 2D and 3D

The Furhat robot [10] was developed as an extension of our talking heads into physical space; we used the same underlying animation tools as described in the previous experiments, but the animation was retro-projected onto a translucent plastic mask instead of being displayed on a screen. One concern we had in this development was that the design would hamper decoding of visual speech when compared to the on-screen face, given that the jaw - which is known to be an importan carrier of visual speech infromation - would be *fixed* on the robot mask. We decided to repeat our earlier speech-in-noise visual intelligibity studies with the robot [10]. In this experiment, 10 normal hearing subjects were presented with acoustically degraded speech paired with one of 6 visual conditions: (1) black screen (audio only), (2) an animated face on flat screen in frontal view (3) back-projected robot face in frontal view, (4) animated face on screen in 45° side view, (5) robot face in 45° side view and (6) a frontal video of the real talker. The results showed (much to our surprise) that the audiovisual intelligibility of the back-projected face was significantly better than that of the animated screen on the face (only surpassed by the video of the real talker). To be noted here is the fact that the face displayed on the screen was *identical* to that being projected onto the plastic mask of the physical robot. The result suggests that the physical presence of the robot face in the same space as the user plays a crucial role, which is in line with findings in literature showing that people have a stronger behavioural and attitudinal response towards a physically embodied agent than a virtual one [7].

4. TURN-TAKING AND GAZE

In this section I will discuss experiments that involve face-to-face interaction. There is a whole range of phenomena related to interaction regulation and turn-taking where gesture, gaze and facial expressions are key signals, and avatars and/or robots can serve as vehicles on the road to deeper understanding of these mechanisms.

4.1. Pushy vs meek

Transformed social interaction (TSI) is a powerful investigative paradigm, that involes mediated interaction where the signal is manipulated on the way [3]. We used TSI to investigate if conversational turn-taking behaviour could be actively influenced by manipulating gaze patterns of an avatar [6]. More precisely, could we make people take the turn more or less often in a conversation with another person (i.e. become more "pushy" or "meek") by changing the gestures of a talking avatar representing the other person? The experiment was set up as follows: Two people were conversing freely over a voice-only skype connecion. On a screen in front of each of the participants was an animated talking head avatar, whose lip movements were synchronized in real time to the speech of the other person [12]. The gaze behaviour of the avatar was controlled by the joint speech activity of the two speakers, according to one of two possible strategies: Strategy *A* was designed to *encourage* the user to take the turn by meeting the gaze of the user at points of potential turn (become more "pushy"), and strategy *B* did the opposite - disencouragement by gaze aversion at turn changes ("meek"). If the avatar on one side employed strategy *A*, the avatar on the other side would use *B* and vice versa. The two strategies were automatically switched at regular intervals (approx. every 10 turns). Six 10-minute interactions (12 different subjects) were recorded with the setup, and the turn-taking statistics were monitored. Results showed that the percentage of times that users took the turn was significantly lower during *B* (meek) conditions than in the *A* (pushy) condition. In a post interview, none of the subjects reported taking any notice to the gaze patterns of the avatar, yet they were *all influenced by it.*

4.2. Multi-Party Gaze in 2D and 3D

One of the initial motivations behind designing the Furhat robot [10] was the limitation of conveying gaze direction faithfully from a head presented on a flat 2D screen, thanks to the Mona Lisa effect [5]. This has implications for multi-party interaction: if multiple users interact with an agent, how do they know who is being addressed by the agent?

An experiment was set up to see if use of a physical talking head (the Furhat robot) would lead to more efficient turn-taking in a multi-party question-answering task, when compared to an avatar on a screen. Five subjects were seated at an equal distance to an animated agent. The agent was either displayed on a 2D screen, or projected onto a 3D mask (Furhat). On each turn, the agent would shift the gaze (without any head rotation) towards a randomly selected subject and pose a question, and await an answer. The experimenter kept track of who responded, as well as the question-response latency. In the 2D condition, the person indicated by the gaze answered in 53% of the cases, while the corresponding number in the 3D condition was 84%. Furthermore, the 2D condition, subjects needed on the average 34% longer to answer. Both effects were statistically significant. Although the ecological validity of the experiment may be questioned, it gives a clear indication of the relevance of physically situated embodiment in multi-party interaction.

5. EXPRESSION AND EMBODIMENT

Unlike their human counterparts, artificial agents such as robots and game characters may be deployed with a large variety of face and body configurations. Some have articulated bodies but lack facial features, and others may be talking heads ending at the neck. Generally, they have many fewer degrees of freedom than humans through which they must express themselves, and there will inevitably be a *filtering effect* when mapping human motion onto the agent. In [2] we study non-verbal expression in three different types of agent embodiments involving facial and/or body motion.

Our underlying assumption is that in order to compensate for the lack of degrees of freedom, we need to emphasize clarity and use stylization in the motion. Our approach is to map motion captured from an actor to different (virtual) robot embodiments. By comparing the original and mapped motion, we are able to measure, through perceptual experiments, how much information is retained in the different embodiments. Our study is the first (to our knowledge) to employ *mime acting* for agent and robot animation, which we believe is a valuable source of high quality, stylized motion especially suitable when large amounts of data are needed (e.g. machine learning).

We performed a full performance capture (gaze tracking, face- and body motion capture) of a mime actor enacting short interactions varying the non-verbal expression along five dimensions (e.g. level of frustration and level of certainty) for each of the three embodiments. The recordings were mapped to animated robot configurations representing different embodiments. We then performed a crowd sourced evaluation experiment comparing the video of the actor to the video of an animated robot for the different embodiments and dimensions. Our findings suggest that the face is especially important to pinpoint emotional reactions, but is also most volatile to filtering effects. The body motion on the other hand had more diverse interpretations, but tended to preserve the interpretation after mapping, and thus proved to be more resilient to filtering.

The main contributions of the study regard the way expressive motion is studied across embodiments and the framework in which the motion is applied to the embodiments; instead of recording natural human motion, we record motion specifically tailored to fit the different embodiments. In this sense, we favor believability and appropriateness over realism.

6. CONCLUSIONS

Human spoken communication is inherently multimodal and to study it, we need tools that take his into account. Facial animation and robotics technology can be important assets in this endavour. In addition, as robotics and virtual agents become part of our daily lives, it is important that we understand how interaction with artificial characters works, and how it relates to human-human interaction: what are the limitations, how can we compensate for them, and what research is needed in order to increase our understanding of such interactions as well as the utility of virtual agents and social robots.

7. ACKNOWLEDGEMENTS

All of the studies described in this paper were carried out in collarboration with researchers and students in our group, as cited, and their contributions are hereby gratefully acknowledged.

8. REFERENCES

[1] Alexanderson, S., Beskow, J. 2014. Animated lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech & Language* 28(2), 607–618.

[2] Alexanderson, S., O'sullivan, C., Neff, M., Beskow, J. 2017. Mimebot - investigating the expressibility of non-verbal communication across agent embodiments. *ACM Transactions on Applied Perception (TAP)* 14(4), 24.

[3] Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., Turk, M. 2004. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators & Virtual Environments* 13(4), 428–441.

[4] Beskow, J., Granström, B., Nordqvist, P., Al_Moubayed, S., Salvi, G., Herzke, T., Schulz, A. 2008. Hearing at home-communication support in home environments for hearing impaired persons. *Ninth Annual Conference of the International Speech Communication Association.*

[5] Edlund, J., Al Moubayed, S., Beskow, J. 2013. Co-present or not? In: *Eye Gaze in Intelligent User Interfaces.* Springer 185–203.

[6] Edlund, J., Beskow, J. 2009. Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech* 52(2-3), 351–367.

[7] Li, J. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77, 23–37.

[8] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the h&h theory. In: *Speech production and speech modelling*. Springer 403–439.

[9] Moubayed, S., Beskow, J., Granström, B. 2010. Auditory visual prominence: From intelligibility to behavior. *Journal of Multimodal User Interfaces* 3, 299–309.

[10] Moubayed, S. A., Skantze, G., Beskow, J. 2013. The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics* 10(01), 1350005.

[11] Parke, F. I. 1972. Computer generated animation of faces. *Proceedings of the ACM annual conference-Volume 1*. ACM 451–457.

[12] Salvi, G., Beskow, J., Al Moubayed, S., Granström, B. 2009. Synface: speech-driven facial animation for virtual speech-reading support. *EURASIP journal on audio, speech, and music processing* 2009, 3.