Bringing the avatar to life

Studies and developments in facial communication for virtual agents and robots

SAMER AL MOUBAYED

Doctoral Thesis
Stockholm, Sweden 2012
A 3D print showing the Furhat robot head in an attempt to bring the face out of the two-dimensional world, in order to break the Mona Lisa gaze effect. The print is best seen at a slight distance.
Abstract

The work presented in this thesis comes in pursuit of the ultimate goal of building spoken and embodied human-like interfaces that are able to interact with humans under human terms. Such interfaces need to employ the subtle, rich and multidimensional signals of communicative and social value that complement the stream of words – signals humans typically use when interacting with each other.

The studies presented in the thesis concern facial signals used in spoken communication, and can be divided into two connected groups. The first is targeted towards exploring and verifying models of facial signals that come in synchrony with speech and its intonation. We refer to this as visual-prosody, and as part of visual-prosody, we take prominence as a case study. We show that the use of prosodically relevant gestures in animated faces results in a more expressive and human-like behaviour. We also show that animated faces supported with these gestures result in more intelligible speech which in turn can be used to aid communication, for example in noisy environments.

The other group of studies targets facial signals that complement speech. As spoken language is a relatively poor system for the communication of spatial information; since such information is visual in nature. Hence, the use of visual movements of spatial value, such as gaze and head movements, is important for an efficient interaction. The use of such signals is especially important when the interaction between the human and the embodied agent is situated – that is when they share the same physical space, and while this space is taken into account in the interaction.

We study the perception, the modelling, and the interaction effects of gaze and head pose in regulating situated and multiparty spoken dialogues in two conditions. The first is the typical case where the animated face is displayed on flat surfaces, and the second where they are displayed on a physical three-dimensional model of a face. The results from the studies show that projecting the animated face onto a face-shaped mask results in an accurate perception of the direction of gaze that is generated by the avatar, and hence can allow for the use of these movements in multiparty spoken dialogue.

Driven by these findings, the Furhat back-projected robot head is developed. Furhat employs state-of-the-art facial animation that is projected on a 3D printout of that face, and a neck to allow for head movements. Although the mask in Furhat is static, the fact that the animated face matches the design of the mask results in a physical face that is perceived to “move”.

We present studies that show how this technique renders a more intelligible, human-like and expressive face. We further present experiments in
which Furhat is used as a tool to investigate properties of facial signals in situated interaction.

Furhat is built to study, implement, and verify models of situated and multiparty, multimodal Human-Machine spoken dialogue, a study that requires that the face is physically situated in the interaction environment rather than in a two-dimensional screen. It also has received much interest from several communities, and been showcased at several venues, including a robot exhibition at the London Science Museum. We present an evaluation study of Furhat at the exhibition where it interacted with several thousand persons in a multiparty conversation. The analysis of the data from the setup further shows that Furhat can accurately regulate multiparty interaction using gaze and head movements.
ACKNOWLEDGEMENTS

During the time of my PhD studies at TMH, I have been privileged to work with an outstanding group of people that filled research with fun and passion.

First and foremost, I want to thank my supervisors Jonas Beskow, and Björn Granström.

Jonas, you have been a LEGEND... wait for it... DARY supervisor. Without you, this thesis would have been a bag of potatoes, no ifs ands or buts. Thank you for always following my work with great interest and engagement, and for sharing with me the passion for fun-engineering. Your support, excitement, generous feedback and down-to-earth discussions have been central to my studies; without your mind-twisting solutions, Tcl tricks, Wavesurfer plugins, and animated faces, this thesis would not have been the same.

Gratitude belongs also to my supervisor Björn Granström. Your expertise, patience, and your interest in my work, your pragmatism and attention to detail, and problem-solving attitude have been great lessons in soldering the researcher in me, and kept me on a straight track to reach this deadline, against all temptations!

I bow to the giants, Gunnar Fant in Momoriam (1919-2009), Anders Askenfelt, Björn Granström, and Rolf Carlson at TMH for setting up and maintaining the probably greatest speech lab in the world!

I owe thanks to Joakim Gustafson for interest, for lightning-fast thinking and out-of-the-box ideas, and for a unique sense of humour. To Gabriel Skantze, co-author of several of the articles in this thesis, for being such a punctual, enthusiastic, and sharp colleague and collaborator, and for a bag-of-skills on dialogue money can’t buy. To Jens Edlund, co-author of Paper D and several other papers, for being a superb dude, and a collaborator with an immense appetite for discussion, interest, inspiring thoughts and beers at Östra. To David House for reviewing much of my work. To Catharine Oertel for commenting on parts of the thesis and for help with the cover. And of course to Furry for fame and glory! and all the non-sense conversations; your days of world domination are coming.

My gratitude goes to Giampiero Salvi for under-the-hood skills with Linux, HTK tips, successful collaborations, and for comments on the thesis. To Mattias Heldner for discussions and for sharing expertise on Swedish prosody and accents, and statistical analyses. To my old officemate Ananthakrishnan Gopal for tolerance with the debugging of noisy dialogues and for allowing me to maintain a crowded room! and to Kalin Stefanov, my new officemate for your thrill-seeking attitude, and again for tolerance with noisy on-site demos. To Kjell Elenius for photography and being a great ski instructor! To Anna Hjälmarsson,
an example perfectionist, and a daredevil go-kart driver – a day will come when my name on that list is above yours! To Mats Blomberg for ASR expertise and non-stop friendliness. And to Anita Kruckenberq for such a shiny and glittery character.

Thank you to Kjetil and Roberto from the Music group for fun music-demos and hacks. To Susanne, Fatima, Yoko and Roxana for laughter and care. To Kerstin for an extraordinary speed and help in administrative issues.

A lot of fine people have made my time at TMH much more than a research project. I salute, including everyone I mentioned above, Andreas, Anne-Marie, Becky, Bo, Chris, Daniel Elenius, Daniel Neiberg, Gael, Glaucia, Inger, Jana, Kahl, Laura, Marco, Olov, Peter, Preben, Raveesh, Simon, Sofia, Sten and all the others I missed to mention. The PhD beers, the Fredagsfika, Friday-hacks, Christmas Parties, and deadlines, no doubt made TMH a second home to me!

Preben Wik, thank you habibi for great times, for sharing many habits with me, standing out in the cold! for jumping out of an aeroplane, for extra-curricular projects and for music. You’ve been a great friend from day one. And thanks also for feedback and proofreads on the thesis.

I also want to thank GSLT, the national Graduate School for Language School for Language technology, in which I have been an associate and received partial funding. GSLT deserves great credit for bringing researchers on Speech and Language together from around the country, for organising interesting courses, and for the great winter getaway retreats.

Appreciation is due to everyone who contributed to, and showed interest in the work I have done in one way or another. Praise is due to all the many (and repeatedly) patient participants who took part in the studies, and to all the anonymous reviewers (especially those who did accept the papers) for valuable comments and suggestions. Many thanks to Anne-Marie Öster for your help in conducting the experiment in Paper A with the hard-of-hearing subjects; and many thanks to the Phonetics Lab at Stockholm University for lending us their Tobii gaze-tracker.

To my friends from outside the gates of KTH. Thank you Johnny-boy for great days as a flatmate, a priceless friend to share interests, frustrations with, and for preserving my sanity while writing this thesis. To Julia for being so colourful, and radiant, and for your enthusiasm about my “talking-necks”. To Rami and all the wisdom from the Book of Ramone and all the boring-kind-of-fun! To Iyad for a very long friendship, and for sharing the love of food! and shelter, with me. To Katkoot for your support, patience, for kayaks, travels, and world-class risottos. To Kinda, Ira, Alesia, Hala, and Charley. You’ve all been great at making sure I maintain a minimum level of social skill during the last period, for taking my mind off when I needed that most, for making me come to terms with the reality of doctoral education, and for filling life with more perspectives and making winters much more lightful and delightful!
Love to my close family. To Bashar and Noura, my best friends, for your lifelong support and enthusiasm; to Julie for making me feel so old. To Hsnee my young brother for keeping me up-to-date with the rock music scene and festivals, all the new techy stuff, and the next generation slang; and for being a great and bright friend. I am so proud of you!

This thesis is dedicated to my parents Bassam and Nada, for their never-ending love and support throughout my life and for their trust and faith in me. I can never be thankful enough.

The work presented in this thesis has been partially funded by the following projects. The EU PF6 project H@H (Hearing at Home, 045089), the EU FP6 project Mon-AMI (IP-035147), the EU FP7 IURO (Interactive Urban Robot, 248314), and EIT ICT Labs project CaSA (Computers as Social Agents, RIHA 12124).
INCLUDED PAPERS

The included papers will be referred to as Paper A → Paper I, and will be cited in this format throughout the text. These papers, included at the end of this book, are identical in content to the versions that have been published; however, they have been reformatted for consistency throughout.

**Paper A**

**Paper B**

**Paper C**

**Paper D**

**Paper E**
**Paper F**

**Paper G**

**Paper H**

**Paper I**
PUBLICATIONS BY THE AUTHOR

This is a chronologically ordered list of publications by the author. Papers from this list which are not included in the thesis are cited throughout the text as a regular reference and are included in the bibliography of the thesis.


CONTRIBUTIONS TO THE PAPERS

The work presented in the thesis was done in collaboration with colleagues. All the work presented here was done with the Speech group at the Department for Speech, Music and Hearing at KTH, and many members of the group have always generously contributed in discussions, reviews, and software tools. Other authors have contributed, at different times, with ideas, design and methodologies, system building, evaluations, and proofreading.

In the following I try to point out, in general terms, the contributions of the co-authors for the papers included in the thesis.

Paper A

The different ideas in the paper have evolved through discussions from all the authors, and were done in the context of the H@H EU project. The first experiment in the paper, which measures the contribution of the different facial cues of prominence to intelligibility, was designed in close collaboration between SA and JB. The second experiment which looks at the pattern of face reading using gaze tracking was designed by all the authors. The article was written mainly by SA, with significant contributions from JB & BG.
Paper B

The main idea of the paper and the methodology was suggested by SA. GA implemented and applied the machine learning tests on the data. LF added spectral tilt to the set of featured used. Writing was done in close collaboration between SA and GA with input from LE.

Paper C

The idea of the paper has emerged over time through discussions from both authors. The experiment was carried out by SA with input from JB. Writing was done by SA with input from JB.

Paper D

The idea of measuring the perception of gaze direction and quantifying the Mona Lisa gaze effect in different displays came about from discussions by SA and JB. The experiment was also designed in collaboration between SA and JB. The analysis was done by SA. The hypotheses and the model for the Mona Lisa gaze effect were suggested and verified by JE. The article was written in close collaboration between SA and JE, with input from JB.
**Paper E**

*The experiment was designed and carried out in collaboration between SA and GS. The analysis and the writing were mainly carried out by SA with input from GS.*

**Paper F**

*The evolution of Furhat over time, from the basic idea, was done by work and discussions from all the authors, and led by SA. The dialogue framework that is used in Furhat is built by GS. The article was written mainly by SA; GS contributed with content on the London Science Museum setup, with input from JB and BG.*

**Paper G**

*The design of the setup at the museum was made mainly by GS and SA and from discussions and ideas from all the authors. The dialogue authoring toolkit was created by GS. The attention model and face control were made mainly by SA. All authors took part in managing the setup at the exhibition. The analysis and writing was done by GS with input from all the authors.*
Paper H

The idea and the methodology in the experiment came about from discussions with all the authors. The experiment and the analysis were done by SA, and the article was written by SA with input from all authors.

Paper I

The experiment and the writing were done in close collaboration between both SA and GS.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two-Dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-Dimensional</td>
</tr>
<tr>
<td>ANOVA</td>
<td>ANalysis Of VAriance</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recogniser</td>
</tr>
<tr>
<td>AV</td>
<td>Audio-Visual</td>
</tr>
<tr>
<td>ECA</td>
<td>Embodied Conversational Agent</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental Frequency</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HRI</td>
<td>Human Robot Interaction</td>
</tr>
<tr>
<td>IR</td>
<td>Infra-Red</td>
</tr>
<tr>
<td>KTH</td>
<td>Royal Institute of Technology</td>
</tr>
<tr>
<td>LSD</td>
<td>Least Significant Difference</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>SADFISH</td>
<td>Sad Angry Disgust Fear Interest Surprise Happy</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TTS</td>
<td>Text-To-Speech</td>
</tr>
<tr>
<td>WIMP</td>
<td>Windows, Icon, Menus, Pointer</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Mark-up Language</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

Abstract ........................................................................................................................................ iii
Acknowledgements .................................................................................................................... v
Included papers ........................................................................................................................ xi
Publications by the author .......................................................................................................... xi
Contributions to the papers ....................................................................................................... xv
Abbreviations ............................................................................................................................. xix
Table of contents ....................................................................................................................... xx
Terminology ............................................................................................................................... xxiii
Thesis Overview ......................................................................................................................... 1

## PART I – INTRODUCTION ................................................................. 5

1.1. Giving machines a face ................................................................................................. 7
1.2. The human face: What is it all about? .................................................................. 10
1.3. Auditory and facial integration in humans ......................................................... 14
1.4. Multimodal communication .................................................................................... 15

## PART II – AUDIO-VISUAL PROSODY ............................................. 23

2.1. Facial correlates to prosody ....................................................................................... 26
2.2. Audio-visual prominence ......................................................................................... 28
2.3. Modelling acoustic prominence ............................................................................. 39
PART III – GAZE AND THE PERCEPTION OF GAZE DIRECTION .... 43

3.1. Face-to-face communication over the screen ........ 46
3.2. The Mona Lisa gaze effect .................................. 47
3.3. Avoiding Mona Lisa gaze with 3D projections ........ 48
3.4. Gaze direction in 2D and 3D projections .............. 49
3.5. Mona Lisa gaze in multiparty dialogue .............. 50
3.6. Gaze direction for situated interaction ............... 52
3.7. Taming the Mona Lisa gaze to our advantage ........ 55

PART IV – FURHAT – A BACK PROJECTED ROBOT HEAD .......... 57

4.1. A brief history of projected heads ..................... 59
4.2. Furhat – A back-projected robot head............... 60
4.3. Reading Furhat’s lips .................................. 62
4.4. Furhat at the London Science Museum ............. 65
4.5. Current work .......................................... 75

PART V – CHALLENGES AND FUTURE WORK .................... 79

4.6. The nonverbal dilemma .................................. 80
4.7. ECAs Beyond first impression .......................... 81
4.8. ECAs are handicapped .................................. 82
4.9. Final notes .............................................. 83

Bibliography ......................................................... 85
TERMINOLOGY

The study of embodied multimodal human-machine interaction is a highly interdisciplinary one. Because of the nature of this field, the reader can come across a number of terms that, depending on the perspective, are sometimes used interchangeably. It is possible also to see the same term used in different contexts to refer to different concepts. The following short list in no way attempts to redefine these terms, it gives guidance to their intended meaning when used in this thesis.

Avatar

According to the Oxford English Dictionary, the word avatar originates from the Sanskrit word *avatāra*, defined as “a manifestation of a deity or released soul in bodily form on earth; an incarnate divine teacher”. Another definition from the same dictionary that reflects the modern user of the term is “an icon or figure representing a particular person in a computer game, Internet forum, etc.”

In the literature, avatar has often referred to a digital representation of a user that is directly controlled by that user (as they are used in forums and computer games). In this work, however, the term is not limited to a representation that a human user controls. The avatar could either be controlled by a user or a computer program. Furthermore, the term here is extended not only to digital representations but also to physical ones, such as the bodies of robots.

It is important to note that the avatar and the process that controls it (whether a human or a software program) are not independent from each other; on the contrary, the communication strategies employed by the process behind the avatar is shaped by the range of functionalities the avatar supports. This definition meets the claims put forward by the theories of embodied cognition (Pfeifer & Bongard, 2007).
Embodied Conversational Agent

An Embodied Conversational Agent or an ECA is an autonomous and intelligent computer program that communicates with the user through a human-like embodied avatar. The term was originally coined by Justine Cassel in her book “Embodied Conversational Agents”, (Cassel, 2000). An ECA is typically set to support multiple channels of communication including spoken language and nonverbal channels such as facial expressions, emotions, gaze and gestures. The agent could moreover be of a specific background and culture, and it could maintain a set personality and attitude.

Other terms that refer to an ECA but emphasise certain characteristics of the agent are Intelligent Virtual Agents, Artificial Humans, or Animated Interface Agents.

Talking Head

A talking head is an avatar that is a three dimensional model of a head that is capable of communicating with the user using spoken language. A talking head has an animated face that supports speech-synchronised lip movements.

Robot Head

Although the term “robot head” covers a wide range of appearances and capabilities, in this thesis it refers to a physical three dimensional head that is human-like in appearance. This in special cases can be a screen with a talking head supported by a neck for physical head rotation. The concept of a robot head is discussed later in the summary of the thesis and in Paper F.

Nonverbal Messages

Nonverbal behaviour has been, and still is, under focus and of wide interest in many disciplines, including anthropology, psychology, and human–computer interaction. Yet, it seems rather difficult for any definition in the literature to draw a clear distinction between verbal
and nonverbal behaviour. In their book “Nonverbal Behaviour in Interpersonal Relations”, Richmond et al. (2011) present a rather detailed discussion on the different possible distinctions and their pros and cons which we have no space here to detail.

The work presented in this thesis does not discuss any linguistic systems aside from the vocal one (spoken language). Because of that, we will commit to the linguistic distinction between verbal and nonverbal messages, defined simply as “communication beyond words” – that is, nonverbal communication is one that cannot be written down in words. In other words, information encoded nonverbally cannot be simply deducted from the lexical channel.

Human-like

Human-like in this text will describe an entity that looks human or a process that generates human-like signals or behaviour. Human-like interaction might also be referred to in the text as natural interaction, meaning interaction that is similar to the one that takes place when humans interact with each other.

Visual Prosody

Although the term visual prosody has been recently used in several studies (reviewed in Part II of this thesis), and there is an apparent increase in interest to study and describe it, there is no clear and traceable definition of what exactly is meant by visual prosody. Clearly however, the term is strongly connected to visual events that resemble these in acoustic prosody. We define visual prosody here as visual movements that correlate with, substitute or complement acoustic prosody. Visual cues in this thesis are restricted to head and facial movements.

Face-To-Face Communication

Face-to-face communication refers to a situation where humans, or humans and machines communicate using spoken language while being visible to each other. This communication includes - in addition to
spoken language - other visual channels such as gestures and gaze. In the case of a human-machine face-to-face communication, the machine communicates with the human using an anthropomorphic embodiment.

Face-to-face communication can further be situated. This means that the parties involved in the communication are part of a shared environment in which the communication takes place.

**Multimodal Interaction**

There is a great variety of definitions about what multimodality is. These vary from the perspective of theoretical models of human information exchange to definitions based on particular applications.

In principle, multimodal interaction provides the user with multiple “modes” of interaction with the system (or with other humans). Oviatt (2002) presents a discussion on multimodality and the distinctions between multimodality and multimedia. Bernsen (2002) introduced a Modality Theory, emphasising the importance of a taxonomical structuring of types of information for the design of multimodal interfaces.

In our work, we will refer to multimodal interaction as an “interaction that involves the use of two or more of the five senses for the exchange of information” (Granström et al. 2002). This definition is synonymic to *multisensory interaction*.

This is not universally agreed on, and we intend not to argue the suitability of the different definitions. Brensen (2002) and Oviatt (2002), for example, consider a system that can interact with gesture and gaze a bimodal system (considering gesture and gaze as two different modes of interaction, in spite of that both utilise only the visual channel).

In our definition, two different types of information that are exchanged using one modality (e.g. the visual modality) are considered as different communication *channels* (such as the lexical channel, prosody, gaze, and gesture). Throughout this text, multimodal interaction is referred to as an interaction that involves only the visual and the auditory channels. In our studies, the visual modality is used to communicate facial signals, and the auditory modality is used to communicate vocal information.
THESIS OVERVIEW

The human face is a communication channel that can carry significant amount of information to facilitate verbal and non-verbal interaction between humans. An understanding of the structure and functions of this channel and along with its intimate relation to speech and discourse is a fundamental requirement for utilising it in human-machine communication.

The umbrella that covers the work in this thesis is how we can design and control artificial animated faces (avatars) so that they support the speech signal and provide two things: 1) a powerful tool for studying human situated multimodal interaction; and 2) develop embodied conversational agent interfaces that employ the pool of rich human signals when interacting with humans. This thesis presents a selection of studies that contribute to the experimental design, exploration, system building, and evaluation of virtually and physically situated avatars.

Part II starts by describing a set of experiments and studies on audio-visual prominence. Prominence is a nonverbal phenomenon that is strongly connected to the speech signal. Acoustic prominence has been previously found to exhibit strong facial correlates such as head and eyebrow movements. Prominence provides an ideal case for studying the perception and synthesis of prosodic facial signals, and how these can contribute to the development of talking heads, while enriching their human-like behaviour.

In Paper A we found that visual cues of prominence represented by eyebrow and head movements, not only provide information about the prosody of the underlying speech signal, but can also increase speech comprehension and intelligibility if incorporated into animated faces. Using these findings, an audio-visual synthesis experiment was done to study gaze patterns of subjects who looked at animated faces that displayed prominence gestures. These gaze patterns were compared with the patterns the same subjects used when looking at animated faces that do not exhibit prominence gestures. The study found that when prominence gestures were present, the gaze patterns observed
when reading the synthesised face were more similar to typical patterns that take place when looking at a human face.

Motivated by those findings, studies in Paper B & Paper C aimed at building data driven models to classify and estimate prominence from the acoustic signal to drive prominence-related gestures automatically from a speech stream.

Gaze behaviour is important - not only from the human side, but also from the avatar’s side. The subtle movements of the eyes can reveal a significant amount of information about the state of attention and focus, interest, and affect, and are used to regulate interaction in dyadic and multiparty conversations. This information is more efficiently transferred using gaze than speech, making gaze an indispensable complementary facial signal in spoken communication.

Part III covers a set of experiments on the human perception of gaze movements generated by an avatar. An accurate human perception of the direction of the avatar and the movements of its eyes is a fundamental requirement before any computational gaze models exhibited by the avatar can be functional when this avatar interacts with a human partner. This issue becomes even more central in situated and multiparty interaction where the attention of the agent might be directed to other humans or objects that are located in the physical space of the human, rather than in the space of the virtual character.

Typically, animated faces are displayed on flat screens. A main limitation of using animated faces visualised on flat displays, which is investigated in Paper D, is the Mona Lisa gaze effect. This effect arises from the fact that flat displays lack direction, and the direction of anything displayed on them does not relate to the dimensions of the real world where the human is situated. This study proposes a solution for this by bringing the animated face out of the screen and projecting it onto a static physical human-shaped head model. In this way, the animated face is taken outside of the virtual world and brought into the real situated world where the human is located. The study reveals that this solution eliminates the Mona Lisa gaze limitation, and shows that gaze direction produced by the projected animated face is perceived accurately in the space of the interaction.

From this finding, an interaction experiment to study the effects the the perception of gaze direction has on multiparty interaction, is carried out and described in Paper E. The experiment supports the findings from the previous study, and shows that flat displays are indeed
problematic to an effective turn-taking behaviour using gaze. In addition to this, it shows that the approach of projecting the face onto a physical 3D head model speeds up turn-taking time and significantly improves turn-taking accuracy. From an additional questionnaire study, the projection approach is perceived as rendering a more human-like face, with eyes perceived to be easier to read.

This positive outcome drove the effort to build a back-projected robot head that utilises animated faces. In this approach, the animated face is back-projected onto a 3D-printed plastic mask that matches, in its design, the face model that is projected onto it. Part IV describes Furhat, the outcome of this effort. Furhat is a back-projected, high-resolution human-like, situated animated character that is a hybrid solution between an animated face and a mechatronic robot head, which harvest the advantages of both and avoids their disadvantages. Furhat is described in detail in Paper F. The approach used in Furhat allows for rendering natural lip movements along with subtle and accurate movement of social and nonverbal facial signals in a robot head, while still giving the illusion that the robot is equipped with a face that “moves”.

Furhat was exhibited as part of a robot festival at the London Science Museum in December 2011. The exhibition put Furhat to the real-world test and showcased a multimodal multiparty dialogue that allowed interaction between Furhat and two simultaneous users. Furhat, in this setup, played the role of an “information seeking” robot, that is explained in Paper G. The exhibition resulted in a large interaction corpus that can be used in studies on multiparty human-robot interaction in an exhibition-like environment. The system furthermore tested the use of head-pose, eye-gaze and different facial feedback gestures for dialogue regulation. 86 questionnaires were collected from the visitors and analysed, revealing several positive aspects on the users’ interactive experience with Furhat.

To evaluate the visualisation of Furhat’s face, a lip reading experiment was done in Paper H. The experiment aimed at measuring the contribution of Furhat’s face movements to speech intelligibility, by comparing it to an on-screen avatar. The results show that people benefit from reading Furhat’s lips as much as and even slightly more than reading the same lip movements on an avatar in a flat display. This verifies that although the jaw in Furhat does not physically move, the perception of the synchronised lip animation is not hindered.
Furhat can be used as research tool for multiparty situated interaction, where models of head-pose and gaze movements can be studied and evaluated in analysis by synthesis experiments. The developments in Furhat were majorly motivated by its ability to use gaze and other orientation signals in situated human-avatar interaction. An example of such a study is in *Paper I*. The study quantifies the human perceptual granularities of gaze when using head-pose, dynamic eyelids and eye movements for the perception of the location of situated objects on a table. Such situated interaction studies would not have been possible without the use of a physical 3D model because of the limitations of flat displays. The results from the experiment are highly relevant to the design and control of Furhat in situated face-to-face interaction.

Furhat, supported by its lips, gestures and gaze, offers an interesting paradigm for both the virtual agents and the robotics communities, allowing for a highly flexible and simple animation of social and human-like behaviour that result in avatars that are perceived more alive.
"I believe that robots should only have faces if they truly need them."

Donald Norman

Sonny the robot: What does this action signify? [winks]
Sonny the robot: As you walked in the room, when you looked at the other human. What does it mean? [winks]

Detective Del Spooner: It's a sign of trust. It's a human thing. You wouldn't understand.

— (I, Robot. 2004)
1. Introduction

Science fiction has been a driving force behind the visions and developments of many of the technologies of today. Perhaps one of the more pronounced products of science fiction is in artificial intelligence. The ability to create machines that exhibit human intelligence is often represented by artificial creatures, or robots, which share certain behavioural characteristics with humans. Science fiction writers often envisioned robots to have levels of intelligence that match and in many cases surpass that of humans.

Science fiction writers distinguished in their depictions robots from humans in certain aspects, emphasising some human characteristics that go beyond the limits of what science can advance in a robot. Examples of such characteristics that are human-specific are empathy, emotions, and morality. For science fiction to show the limitations of robots, they depict them as lacking human qualities and human-like behaviours. This is done often not by showing an inability to speak and understand human languages, but rather by their inability to communicate socially relevant behaviours nonverbally.

Robots, although sometimes looking very human, are commonly portrayed as being unable to understand and produce signals of social awareness that are present in human communication; they are repeatedly portrayed as having stiff muscles, dull faces, static eyebrows and sluggish eyes, with metallic voices that lack melody.

It is through these visible representations that robots are displayed to the viewer as being non-humans. Thus, robots have been defined and separated from humans by their inability to mimic human behaviour.

This notion has been prominent in society to such a degree that the word “robot” has evolved to denote certain human personalities that lack emotional and social behaviour. For example, according to the Oxford English Dictionary, a definition of the word “robotic” is “resembling characteristic of a robot, especially in being stiff or unemotional”; the Longman Dictionary of Contemporary English defined this as “someone who does things in a very quick and effective way but never shows their emotions”.

Science fiction has given researchers a vision of how technologies in artificial intelligence are expected to evolve, giving machines human capabilities of vision and language; however, showing that these
capabilities will not suffice if we aim to build machines that are perceived having human qualities. It also has indicated what to do to humanise machines’ behaviour, machines that we are able to ascribe human qualities to, and to share empathy with. That is by giving them nonverbal behaviour.

This thesis does not target the hard problem of strong artificial intelligence to build machines that experience emotions or have social awareness - concepts usually related to consciousness, sentience, and self-awareness. It rather addresses questions about design strategies and signals that humans use in the inter-communication and extend beyond the stream of words and, when implemented in machines, result in a behaviour that is characterised as being human-like while having a communicative value that supports a more efficient, fluent, and rich interaction.

1.1. GIVING MACHINES A FACE

The quest towards building machines that communicate with humans in a human-like fashion is not only driven by the fascination of science-fiction authors and fans, but also because of the value they add to the interaction.

Today, the overwhelming majority of interacting with machines is done using traditional input devices (e.g. keyboard, mouse and touch), with machines providing traditional graphical manipulation using WMIPs (Windows, menus, icons, pointer) as well as multimedia output of text, images, and videos. Although these interfaces are highly efficient for specific tasks (such as image manipulation or web-browsing), they are very different from how humans regularly interact with each other and exchange information between themselves, resulting in a relationship and an interaction experience that is very different from ones humans exercise and master during their life, and one that lacks several dimensions. It also involves a learning curve for the interface with which not everyone is comfortable.

Humans, in most of their daily communication, use spoken language. When humans are visible to each other during interaction, they use diverse visual movements to communicate different types of information; these include gestures, eye movements, facial expressions, and body movements. Most humans are masters of this communication – it is a skill that comes natural from early infancy and much experience.
The importance of giving machines interfaces that understand and speak “human” has, for a long time, been highlighted. Research has invested considerable efforts into building models and engineering solutions to encode and decode information in a human-like manner.

At least two high mountains need to be climbed to reach this goal. The first is to develop input devices that are accurate enough to capture representations of the information the user supplies (speech recognition, gesture recognition, gaze tracking, etc.) along with interfaces that give information to the user using the same representations (speech output, animated gestures, eye movements, etc.). The second is to build models that process the input information over the course of the interaction, but also encode and communicate the information in similar ways as humans would do during a face-to-face conversation. This needs to be done not as the typical Ping-Pong exchange similar to traditional graphical user interfaces of today (question-answer), but rather as a constant exchange of information that accounts for phenomena such as grounding, feedback, error handling, and turn-taking. Such models need to take into account the use of nonverbal representations such as head movement, facial movements, eye movements, and the intonation in speech, and decide how all these different representations are orchestrated to transfer a message depending on its complexity, on the context and on the individual characteristics of the user.

Many research findings support this effort. Moving from a WIMP interface to a spoken interface enhances the communication in several situations. An obvious advantage is that the user can give hands-free input to the machine when the hands are not available (e.g. they are involved in other tasks or are at a distance from the interface, or for users with limited motor-skills (Damper 1984; Cohen 1992)). Another advantage is that the user can receive output from the interface when the eyes are not available (e.g. they are involved in other tasks such as driving or are at a distance from the machine, or when the user suffers vision related problems (Carlson, Granström & Larson, 1976)).

Interaction can also be faster when involving complex messages without going into a hierarchy of menus and pages in traditional GUIs, but rather in one single utterance: for example “Why is this apartment more expensive than the one downtown that you showed me before?” (Gustafson, 2002).
Spoken human-machine interfaces bring several new dimensions and possibilities to the interaction; technologies being developed in this direction have shown much potential to help a large number of persons and to build applications that are more efficient and spontaneous to control. However, there is a multitude of signals that are missing in the human-machine interaction loop when these machines lack a face. Giving machines a face complements in several ways the acoustic speech signal that is communicated by the machine to the user (and vice versa). The human face can be regarded as a communication channel that carries much information that is utilised in human-human interaction; replicating it and its behaviour has the potential to create more human-like, robust, and efficient human-machine interaction.

Humans engaging in a face-to-face conversation use a set of nonverbal facial signals to transfer information on different levels between each other – for example, phonemic, intonational, and emotional (Ekman, 1979). This information can be utilised in human-machine interaction, by providing an embodiment to the interface. Studies have shown that users tend to spend more time with systems that use embodied talking agents, while better enjoying the experience (Walker et al. 1994, Lester et al. 1999).

The first thing a face can give a dialogue system is the identity of the speaker (personality, age, gender, status, etc.). Humans are experts at recognising faces (Donath, 2001) and this fact can be used to make different applications more memorable and recognisable using their unique embodiment.

In addition to giving the system a recognisable interface character, the character can use a large number of messages of social value that are available in the face to enhance the interaction. This is crucial in several types of applications where social competence is beneficiary, such as in teaching, commerce, and interpersonal relations (Reeves, 2000). Reeves notes how although people understand that interactive characters are not real, the characters still cause social responses in the users as if they are real “The willing suspension of disbelief” (similar to how fictional characters in motion pictures affect the audience). Reeves also found that communicating information using an interactive character increases the trust in the source of the information. In addition, this results in a better recall of the presented information (Beun et al. 2003).
Nass et al. (1994) claim that the interaction with embodied agents follows the social conventions exhibited in human interaction; and hence, the users utilise behavioural and communicative signals in their interaction similarly to as they do in human-human interaction. Marsi & Rooden (2007) found that users prefer nonverbal indications of the embodied system state (in their case uncertainty) to a verbal one.

Much empirical evidence indicates that giving machines a face has the potential to enhance the user-machine interaction experience and providing it with interpersonal social aspects. In addition, the human face can encode and transmit information units using a large set of signals that have an important communicative function, which can be exploited in face-to-face interaction. The next section explores some of the signals the face has the power to use during interaction.

1.2. The Human Face: What Is It All About?

The human face is arguably man’s best verbal and nonverbal communication device. The main reason the face is so important in communication is that it is usually visible during interaction. The study of the human face has a long history. Charles Darwin published the first account of a scientific work on facial expressions, in The Expression of Emotion in Man and Animal, in 1872 (Darwin, 1972). But the physiognomy of the face has intrigued many pseudo-scientific minds before that. Several people tried to show how the appearance of the face (commonly referred to as physiognomy or the art of face reading) can indicate personality traits such as criminality, emotional stability, and intelligence. An example of such cultural superstition is the ancient Chinese art of Siang Mien (Keuei, 1999). This describes what man’s face says about him, and how the structure and appearance of the different facial parts (e.g. the shape and size of the lips, nose, and eyebrows) can help to predict a person’s fortune, happiness, luck and destiny. (For other examples, check: Brown, 2000 and Haner, 2008.)

The human face, including all visible moving parts of the human head, carries a significant amount of information that humans can encode, decode and interpret. Below, we shortly describe some of the important functions of the human face, and their communicative power. Later on, more thorough reviews of some parts of the face and their functions are presented.
1.2.1. The Eyes

The study of the eye and its functions is called *Oculsics*. Of all the features of the face, the eyes are probably the most important means of communication that humans possess after words. Morris (1985) suggests that despite all the talking, listening, moving, and touching we do, we are still visual. The initial contact made between people in a face-to-face setting is usually with the eyes, if no eye contact happens, it is likely that no additional communication will take place.

The basic function of the eyes is vision. The location of where the eyes are directed is read by observers; from this, the objects of focus and interest of the gazer are interpreted by the observer and from this, several cognitive and emotional states can be inferred. This ability to read the direction of someone’s gaze by others allows gaze to play a crucial role in the regulation of situated and multiparty dialogue.

Eye movements have a significant communicative value during conversation. According to Kendon (1967), in two-person (dyadic) conversations, seeking or avoiding to look at the eyes of the conversational partner (i.e. catch their gaze) serves at least four functions: (1) to provide visual feedback; (2) to regulate the flow of conversation; (3) to communicate emotions and relationships; and (4) to improve concentration by restriction of visual input.

Argyle (1976) estimated that when two people are talking, about 60% of conversation involves gaze while about 30% involves mutual gaze (or eye contact). According to Argyle, people look nearly twice as much when listening (75%) as they do when speaking (41%), showing an example of the variability of gaze between listening and speaking in dialogue.

1.2.2. The Vocal Tract

The vocal tract obviously makes vocal communication and spoken language possible. By the movements of the different parts of the vocal tract, different sounds are created. The vocal tract is the source of speech signals (linguistic and paralinguistic) and some of its movements are visible, and hence, by looking at them, certain information about the produced speech signal can be visually inferred. This process is commonly called *speech reading* (or less accurately, *lip reading*).
These visible parts of the mouth - lips, jaw, tongue, and teeth – improve communication especially when the vocal signal is noisy (Sumby & Pollack, 1954). Because of the direct link between speech acoustics and these movements, our brains learn the association between them, and can on some level infer one from the other. Humans take advantage of the appearance of the visible articulators to a degree that speech perception itself is affected by how we see them move. A clear example of that is the McGurk effect, described in the seminal paper of McGurk and McDonald (1976). This shows that the perception of the sound signal can alter if the facial movement seen in synchrony with it is incongruent. This could be explained by the large exposure of audio-facial speech humans experience resulting in the perception of speech being very sensitive to these facial movements. This effect has been shown to matter not only with articulatory movements, but furthermore, for example, also with facial expressions to alter the perception of emotions (Fagel, 2006; Abelin, 2007).

1.2.3. FACIAL SETTINGS

When we speak, we do so in what Trager (1958) calls the “setting” of the act of speech. It is the environment or contextual information that can be inferred from the speaker’s voice, which represent several of the speaker’s own characteristics (age, gender, accent, dialect, etc.). In addition to the act of speech, the term “setting” can be used to describe the face. The facial setting is what the face (be it static or in motion) reveals about its owner. These factors include age, health, gender, race, fatigue, enthusiasm, and emotions, which can even give clues to intellectual background, cultural background and social status. It is through our perception of these features when talking to others that we get help to interpret the verbal message but also to predict much of the interaction and communication patterns that take place.

1.2.4. FACIAL EXPRESSIONS

The subject of facial expressions is of great historical interest for several reasons. Nonverbal theorists equate the study of facial expressions to the study of “emotion itself” (Darwin, 1872; Tomkins, 1962). The human face is a primary channel for transmitting emotional expressions because of the complex repertoire of configurations the many muscles and bones in the face can create.
This thesis does not directly address research questions on facial expressions; however, research and investigations done on facial expressions show how deeply connected facial expressions are to human communication.

Charles Darwin was interested in the study of facial expressions in animals (Ekman, 1973). He believed that facial expressions are essential to survival and that they evolved in much the same way as other physical characteristics did. The debate about whether facial expressions are innate, learnt, or both, is old, but there is increasing evidence that some of them are innate and universal in human beings. For some facial expressions, the meanings conveyed across cultures is the same (Weitz, 1974), which is not the case for many other aspects of nonverbal behaviour. Eibl-Eibesfeldt (1970), a researcher on expressive behaviour, studied the facial expressions of deaf-blind children; he discovered that the expressions of basic emotions (sadness, anger, disgust, fear, interest, surprise and happiness, or SADFISH) are observable in their behaviour and that the probability of them being learnt is practically nil. Such intrinsic and evolutionary encoding of behaviour in the face shows just how important it is as an effective communication device.

1.2.5. EYEBROWS

Eyebrows might have evolved to protect the eyes from sweat, rain, and dust; however, they might have remained because of their function in nonverbal interaction. Eyebrows play a significant role in the configuration of multiple facial expressions (Ekman, 1979). In addition to that, eyebrow movements are one of the facial features that have been shown to be highly co-verbal and in synchrony with different vocal properties such as pitch (Cave et al. 1996); hence, they are an optimal facial cue to information about the prosodic properties of the verbal message.

Part II of this thesis explores in details the different research findings on the communicative function of eyebrow movements. In Paper A, we use an analysis by synthesis setup to investigate whether the movements of the eyebrows can actually help to increase the comprehension of speech when the prosody of that speech signal is degraded.
1.2.6. HEAD MOVEMENTS

Primarily, the main communicative function of the head is to help the person direct attention and scan its environment, which is mainly done in coordination with the eyes to reach a wider visual field of view than the eyes have without head movements. The same function applies to speaking, where the head orientation directs the speech signal towards its intended target. These basic parameters give rise to a more complex system of movements between the speech signal, the eyes, and the head, in order to play different communicative and nonverbal functions in interaction.

The importance of head movements is evident in coordinating situated multiparty communication among humans, where a human might look not only at the face of the human with whom they are communicating, but also at other objects or other humans in the environment.

Head movements also play an important role in nonverbal interaction. However, their functions and dynamics have been studied much less than other facial and nonverbal signals. Some studies have found that co-verbal head movements signal rhythm, accentuation and emphasis, in addition to more structured gestures such as head nods and head shakes (Hadar et al. 1983; Munhall et al. 2004).

1.3. AUDITORY AND FACIAL INTEGRATION IN HUMANS

Human communication is multimodal. All natural human communication systems have evolved among humans sharing the same space and having audio-visual access to each other (except in the dark), providing concurrent auditory and visual signals (Sumby & Pollack, 1954). Until recent human advances in telecommunication (with technologies such as radio broadcasts and the telephone), humans have communicated in face-to-face situations and hence it is natural that communication developed in collaboration between auditory and visual signals. Human sounds can be used for localisation, and as sounds are produced by the vocal tract, directing oneself to the source of the sound results in humans spending large amounts of their interactions looking at each other’s faces.

This strong relationship between auditory-visual processing has been documented in several studies challenging unisensory models of
speech processing in humans. Von Kriegstein et al. (2006) found that even during the processing of auditory-only speech, the brain optimises auditory-only and speaker recognition using speaker-specific predictions from distinct visual face-processing areas of the brain. Van Wassenhove (2005) found that visual speech speeds up the neural processing of auditory speech. Calvert et al. (1997) found that similar activation in the auditory cortex takes place during silent lip-reading. Not only does visual speech help in the processing of auditory-speech, but the opposite is true as well. Schweinberger et al. (2007) found that people can infer visual-facial information about a speaker’s identity by listening to their voices.

Human communication relying on acoustic speech only could even be limited and insufficient. Language is a relatively poor system for the communication of spatial information. Such information is basically visual and its communication is best encoded using a visual system (e.g. gaze and pointing gestures). The ability of humans to communicate the location of food and predators in an efficient way, to interpret the attention state and to infer target objects of interest of other humans is essential for survival. Such signals are most visible when looking at a person’s eyes.

Human infants are born with a predisposition to listen to human speech and to watch the faces of people when they are speaking. An infant learning to communicate and speak in its early days is involved in face-to-face interaction between the care-giver and the infant. Pronunciation instructions and feedback is almost always given with the infant having access to the face of the care-giver. De Villiers & de Villiers (1992) state that from the time an infant is born, they are attentive to the human face. Gopnik et al. (1999) found that infants prefer the sight of a human face to other sights.

1.4. MULTIMODAL COMMUNICATION

A multimodal system is a system that supports the communication between two entities using more than one mode of communication. This by definition requires more than one channel to be involved in the interaction loop.

From the perspective of one multimodal system, these modes of communication can be input modes, output modes, or both. For example, two people talking on the phone is a unimodal system that
involves the auditory channel (refer to the definition of “modality” in the terminology section for more details). Two people using sign language is also a unimodal system but this involves the visual channel. Two people speaking face-to-face while using hand gestures is a multimodal system that involves the auditory and visual channels; for each person involved in this, the communication involves two input and two output modalities (both people use their ears and eyes).

The distinction between input and output modalities in a multimodal system is important for this work: the work in this thesis mainly targets building systems that support two output modalities; in other words, a system, while interacting with a human user, transfers its information using speech and facial movements. We will refer to the process that generates multimodal output in such a system as multimodal speech synthesis.

In general, multimodal communication extends beyond using the auditory and visual channels. However, the topics addressed in this thesis target a subset of multimodal systems, namely when two entities (a human or a machine) communicate using speech and face movements. The rest of this summary will mostly use the term *multimodal*, meaning audio-visual or speech-facial communication. This involves the auditory modality used for speech communication and the visual modality used for the communication of facial signals.

### 1.4.1. MULTIMODAL INTERFACES

Multimodal Interfaces are computer interfaces that support multiple modes of interaction with the user. This traditionally included the use of multiple modes of input such as speech, gesture, pen and touch.

In the beginning of the 1980s, a variety of multimodal interfaces emerged, starting from the “Put That There” demonstration of Bolt’s (1980) which processed speech and pointing simultaneously. This research created a new class of interfaces targeting naturally occurring forms of human communication behaviour (spoken language, body movement, and nonverbal behaviour). The development of these novel interfaces has been increasingly allowed by advances in input and output technologies as they become more available.

Research has documented several advantages that systems supporting multimodal communication have over unimodal systems. These systems, for example, allow for a more efficient, natural, and
robust encoding and communication of the intended messages. A benefit of multimodality is that the user or the system can combine multiple modalities to transfer a single message, which has been shown to decrease communication error rates (Oviatt & VanGent, 1996; Bangalore & Johnston, 2000).

The advantages of multimodal over unimodal interaction derive mainly from two basic properties of multimodal communication: Redundancy and Complementarity. A simple conceptual chart of these is shown in Figure 1.

1.4.2. REDUNDANCY

Redundancy in a multimodal system means that certain information can be transferred using both modalities. In human multimodal communication, redundancy can be used for different goals in different contexts. For example, in noisy environments, humans in a face-to-face conversation can use iconic gestures in addition to the vocal forms of these gestures to guarantee a robust delivery of the message. They can further switch between them in repetitions, grounding or error handling. In quiet environments, these can be used to emphasise, highlight and enhance the message.

As each modality has its inherent characteristics, the choice made to use one modality over the other can in itself have a communicative function. For example, as facial expressions used to express emotions are less invasive than speech, humans in public settings might choose either the auditory or visual form depending on how many people they want to deliver the message to or how much privacy they need.

Redundancy plays a crucial role in multimodal communication. Some of the redundancies from producing speech for example are physically correlated with certain movements of the face. As speech is produced using the vocal tract, which has visible parts that constitute areas in the face (lips, tongue, jaw, teeth), certain information is typically always transferred in redundancy. This redundancy becomes even expected and to a certain degree required for fluent communication. For example, in a human-machine conversation, an avatar producing speech without moving the lips would likely result in the receiver believing that the source of the speech is not that avatar (e.g. The Ventriloquist (Alais & Burr, 2004)), and this would likely disrupt smooth communication.
This redundancy also serves the functions mentioned above such as robustness in interaction. The contribution of lip movement in face-to-face speech communication has been thoroughly investigated in research. In noisy environments, lip movements can substitute some of the loss of information in the verbal speech message at a low segmental level (Erber, 1974; Summerfield, 1992). This is a feature the hard-of-hearing persons exploit to enhance their speech perception, and it is used in noisy environments to support the speech signal.

This redundancy of information between the speech signal and facial movements on the verbal level is the topic of Paper H in this thesis. The study explores and evaluates the contribution of the visible articulators to speech intelligibility using a computer generated talking head.

The lips do not transfer information about the verbal segmental speech sounds (the phonetic content) only. It was found in several studies (e.g. Graf et al. 2002; Cho, 2002; Keating et al. 2003) that certain aspects of the prosodic (nonverbal, supra-segmental) content of the speech signal can be observed in the visible articulatory movements, such as loudness and pitch level, which correlate with jaw opening, lip opening and rounding.
The redundancy in face-to-face speech communication need not be physically causal either (i.e. one is a result of the other because of physical movement). Many non-verbal messages can be transferred using the speech signal, the face, or both. Head gestures (nods and shakes) can be substituted with words or sounds (e.g. yes, no, aha, ehem, etc.). This non-mandated redundancy can be used in encoding prosodic information in human face-to-face communication. Co-verbal head and eyebrow movements can provide prosodic information that can be communicated in synchrony with vocal prosody. These non-verbal vocal/facial movements can serve the same communicative functions (a pitch raise or/and an eyebrow raise at the end of a phrase could signal a question (Cosnier, 1991; Granström & House, 2005)). This redundancy of multimodal prosodic information in the face and speech is explored in *Paper A*, using prominence as a case study on audio-visual prosody. The study investigates and quantifies the effects of facial prominence cues on speech intelligibility.

### 1.4.3. **Complementarity**

The other main advantage of multimodal communication is complementarity. Complementarity in a multimodal system describes certain types of information which are better encoded and communicated using one modality over the other.

In the domain of human face-to-face communication, complementarity is crucial. Communicating information using the modality that is not optimal for that type of information can equally be unnatural. In research by Oviatt, (2002), it was found that disfluency in language during interaction increased by 50% when people had to speak **spatial information**. The communication of spatial information is probably the most important complementary feature of facial movement to speech. Spatial information is hard to encode using speech, while very subtle and efficient to transfer using head orientation and gaze direction. *Paper F, Paper G* and *Paper I* explore the efficiency of using gaze and head direction to describe spatial information, in the context of spatial multiparty human computer interaction. The studies also compare these effects between faces visualised on flat computer screens and via physically situated three-dimensional faces.
In practice, a “black and white” distinction between redundancy and complementarity is simplistic. Obviously, over the course of conversation and depending on the available modalities to communicate over as well as on the complexity of the message, different information in different contexts are easier being transferred using one modality over another. For example, redundancy can be used to exhibit “irony” or “sarcasm” by transferring “contradicting” cues through the auditory-visual channel. In other situations, what is redundant might become complementary. In noisy environments for example, the typically redundant lip movements can complement the loss of intelligibility of the auditory channel, and become required for an intelligible perception of the speech signal.

However, this distinction provides a functional and useful basis for the development of embodied multimodal interfaces, especially in the modelling and control of the interface behaviour.

One can look at the audio-visual interaction between two humans as a two-way information exchange process. At any given point in time, this process involves a transmitter that encodes information, channels that transfer the information, and a receiver that decodes the information.

During an audio-visual interaction between humans, this means that the transmitter of the signal is the speaker’s brain, the encoder is his vocal tract, the channel is air, the decoder is the listener’s ears, and the destination is the listener’s brain (Littlejohn, 1996).

For successful communication, the transmitter ensures that the signals received by the receiver are decoded in the intended way and the message received is identical to the message sent. For this to happen, in dynamically changing environments, the transmitter should be aware first of the capabilities of his own encoders. The signal a speaker produces also relies highly on the properties of the environment. In changing and noisy environments or with large distances between the speaker and the listener, speech communication is less efficient. In such situations, the speaker can adapt to ensure optimal interaction, such as the choice of words, and the volume and the rate of speech. The same applies to the knowledge of the properties of the receiver’s encoders. When speaking to a hearing impaired person, the speaker can adapt his voice to compensate for the loss of the signal that the impairment caused. Finally, the speaker might receive feedback from the listener on whether the message was fully
understood or not, so the speaker can adapt again to compensate for any miscommunication. This process and its properties are adaptive in different contexts and in different environments; speakers and listeners in conversation constantly adapt the form of their messages to optimise communication (e.g. Condon & Ostgon, 1971; Alibali & Heath, 2001).

Looking at communication from this perspective, the different functions of multimodal communication become obvious. When the speaker and listener have two modalities at their disposal instead of only speech, they can further optimise and enrich their interaction by taking advantage of the forms possible to communicate using both modalities. If the audio channel is noisy, the speaker can rely more on the visual channel; when the visual channel is noisy (e.g. partial visibility of the face), the speaker can rely more on the auditory channel, (Lindblom, 1996).

This thesis deals with building some aspects of the structure and dynamics of facial synthesis. This is done based on empirical studies by examining the unimodal and multimodal effects of these different variables on the human’s perception of the message.

1.4.4. HUMAN SPEECH COMMUNICATION SUPPORT

Avatars can enhance speech communication between humans when their faces are not used in the communication. By adding a talking head to the speaker’s speech, the listener can get supporting information from the dynamics of the artificial face that enhances the comprehension of the speech signal. An application of this is telephone communication. The talking head, in principle, can generate corresponding facial movements in synchrony with the speech signal. These movements can be used by the listener to enhance speech comprehension (Beskow et al. 1997; Agelfors et al. 1998; Al Moubayed & Beskow, 2011). This same advantage has also been used for the support of the hard-of-hearing (Salvi et al. 2009; Al Moubayed et al. 2009a).

The development of speech-driven facial animation has been a research topic of interest for more than 15 years. Several models have been built to estimate accurate articulatory movements from the human speech signal (Salvi et al. 2009; Al Moubayed & Ananthakrishnan, 2010), and articulatory models have been developed to animate the avatar’s face (Cohen & Massaro, 1993; Beskow, 1995;
Ezzat et al. 2002). These applications require an understanding, quantification and extraction of redundant information from the speech signal, and then encoding this information using the avatar’s face.

**Paper A** in this thesis targets this question. The experiments presented in the paper concern the extraction of prosodic parameters from the human speech signal to drive a talking head using gestures in order to substitute the loss of these auditory parameters. Experiment 2 in **Paper A** describes the positive effects a talking head that generates prosodically relevant gestures has on the communication.

**Paper B** and **Paper C** of the thesis present work on developing linear and nonlinear models to estimate prosodic prominence from the speech signal.
PART II

AUDIO-VISUAL PROSODY
Related Papers

**Paper A**

**Paper B**

**Paper C**
2. Audio-Visual Prosody

Prosody in linguistics is the stress, rhythm and intonation of speech. Prosodic realisations are supra-segmental by definition, and are not restricted to one linguistic segment.

A popular view on prosody is that it is pre-linguistic in origin. The “frequency code” (Ohala, 1994) suggests that there is a vocalic code that is universal, works across species and is independent from language. For example, frequencies produced by small (high) or big (low) species can communicate meanings such as “harmless”, “submissive” and “dangerous” (Mortin, 1977). In vocal languages, prosody primarily involves variations in segmental durations, loudness and pitch of speech sounds. Emotional prosody is considered by Charles Darwin in the *The Descent of Man* (Darwin, 1871, p. 524) to precede human language: “Even monkeys express strong feelings in different tones – anger and impatience by low, – fear and pain by high notes.”

A view on prosody need not assume a mere vocalic constraint to it. All aspects of prosody are produced using muscle motion and this motion can often be perceived directly by the listener (see Section 1.1). From this, hand gestures and facial movements can be considered (in part) prosodic as they carry information that modifies the meaning of the lexical message in a manner similar to acoustic prosody.

From that, it is reasonable to consider an audio-visual account of prosody (audio-visual prosody) – the variations of the acoustic features can also be communicated by visual movements in parallel to the verbal message.

The study of visual prosody is less explored and quantified than its acoustic counterpart. For example, the field of expressive Text-to-Speech synthesis has recently progressed and maturing, advancing models for acoustic prosody which are able to enhance the naturalness of the generated speech, resulting in highly human-like and intelligible output. However, to date, there are only a handful of trials that aimed at building models that generate facial prosody (Albrecht et al. 2002; Pelachaud et al. 1996; Busso et al. 2007).

One way of describing the importance of visual prosody is by comparing it to acoustic prosody. Visual prosody to a synthetic face is what acoustic prosody is to a synthetic voice. A synthetic speech signal that has a constant speech rate, flat pitch, and constant loudness results
in a poor, wooden, and artificial voice that is not suitable for natural and spontaneous interaction with humans. This is analogous to a talking head that moves only the articulators (the mouth) in accordance with the speech sounds while keeping the head, eyes, eyebrows, and facial expressions static.

Variation in the speech signal and its articulatory correlates is not enough to make a natural and expressive talking head. Facial variations, beyond visible articulatory are required. To achieve this, researchers on embodied conversational agents (Badler et al. 1999; Albrecht et al. 2002; Lee et al. 2002; Gratch et al. 2002) employ some form of Perlin noise (Perlin, 1995). Perlin noises, in the case of facial animation, are random movements added to the face (head movements, eye movements, eyebrow movements, etc.) to make it look more “alive”. While Perlin noise contributes to how natural and life-like animations are perceived, in principle the variations it adds have no functional component (i.e. these movements do not have a specific contextual communicative/linguistic function).

Nonverbal audio-visual prosody is important not only to increase the expressivity of the talking head, but also to communicate the different nonverbal functions of prosody largely utilised in human face-to-face conversations.

2.1. FACIAL CORRELATES TO PROSODY

Several studies have successfully identified and measure certain facial movements that communication functions of acoustic prosody. Some prosodic movements have direct physiological correlations to certain visible cues around the mouth area. Segmental duration of linguistic segments (duration of a phoneme, a syllable, or a word), are clearly visible because of the synchrony between speech sounds and lip movement. Some studies have even found certain correlates to F0 movements in the mouth region. In Cho (2002), De Jong (1995) and Krahmer & Swerts (2004), pitch accents were found to be associated with larger, longer, and faster articulations, especially for mouth and jaw opening (see also Erickson et al. (1998) and Keating et al. (2003)).

Outside the mouth area, prosodic cues, that are more visible and less affected by direct muscular impact of articulation, have been observed in the face, mainly in terms of head and eyebrow movements. In an empirical study targeting the relation between eyebrow
movements and pitch accents, Cavé et al. (1996) found that eyebrow movements significantly correlate with pitch accents: when someone speaks, 71% of the rapid raising-falling eyebrow movements happen in synchrony with F0 raises (but the opposite is not true). The authors concluded that although eyebrow movements happen in synchrony with pitch accents, this synchrony might not be physiologically connected (as it does not happen all the time). This synchrony could rather have a functional role in communication, something that we will come to in the next section and has been suggested earlier in several studies (Cosnier, 1991; Birdwhistell, 1970; Pentland & Darell, 1994; Guaitella et al. 2009).

Since not all pitch movements have corresponding eyebrow movements, the question of where and when this correlation takes place is not completely answered yet, although the process is likely to not be fully predictable due to the large variability among individuals, contexts and cultures (Ekman, 1979).

In addition to eyebrow movements, head movements have also been shown in a few studies to play a role in communicating prosody. Munhall et al. (2004), animated the head of a character using head movements copied from a human speaker, along with decreasing the intelligibility of the speech signal by noise. They found that the intelligibility of the sentences increased when head movements were applied to the talking head, compared to the same head that was animated without any head movements. There is no exact quantification of what exactly in the head movements contributed to this increase in intelligibility. However, several ad-hoc hypotheses are discussed - head movements might be indicators of rhythm of speech, or perhaps give clues to the start and end of utterances. A follow up study was carried out by Davis and Kim (2006). They found that by visualising head movements alone, without lip movements (showing the top contour of an animated head), along with the audio signal, increased “slightly but reliably” the intelligibility of the speech signal. However, this increase happened only with sentences produced with large expressivity (yielding larger head movements).

Other analytical studies of head movements have found a correlation between head nods (similar to eyebrow raises) and pitch accents (Hadar et al. 1983; Graf et al. 2002; Swertz & Krahmer, 2006). From the perceptual perspective of these movements in (Granström & House, 2005), a final eyebrow raise in audio-visual sentences has shown
to discriminate between questions and statements, playing the same
typical function of a final pitch raise in question sentences.

All these studies suggest a highly integrated production and
perception of prosody.

Aside from the general findings on the use and benefits of visual
prosody in face-to-face communication, most of the research on audio-
visual prosody has focused on prominence. Prominence is a general
acoustic/prosodic phenomenon that is to a large degree acoustically
understood and documented in different languages and it can be
studied from an audio-visual perspective.

2.2. AUDIO-VISUAL PROMINENCE

Prominence is defined as when a linguistic segment is perceived to
stand out of its context (Terken, 1991). It describes the perceptual
saliency of words (or longer or shorter segments, such as phrases or
syllables) in their own context (Horne, 2000). According to these
definitions, prominence is a qualitative perceptual property of linguistic
segments. Many studies have aimed at quantifying it in different
languages in terms of acoustic properties.

The literature provides several functions of prominence. Segments
can be made prominent to convey information such as contrast, focus
(Gundel, 1999), and information status (Grice & Savino, 1997), in
addition to their inherent stress patterns in some languages (syllables in
words contrasted using lexical stress). Hence, the perception of
prominence in a linguistic message impacts on its interpretation;
consequently, affecting speech comprehension.

From the perspectives of production and of perception, auditory
prominence has a strong relationship with visual and facial movements
outside of the mouth area. This relationship is strong enough to result
from muscular synergy.

An example might best illustrate this audio-visual correspondence. If
we consider the pronunciation of the following phrase; words in
CAPITAL letters are pronounced prominent (emphasised) acoustically,
while underlined words are pronounced in synchrony with a beat
gesture (e.g. eyebrow raise or a head nod):
(1) This is a BLUE car.
(2) This is a blue CAR.
(3) This is a BLUE car.
(4) This is a blue CAR.
(5) This is a BLUE car.

Most people have no problem pronouncing 1 or 2 with an eyebrow raise or a head node on the prominent word; however, most people find it difficult to pronounce phrase 3 or 4 correctly (at least without practicing). People tend to either make both final words acoustically prominent or add a gesture in synchrony with words. This indicates that during speech, it is important that eyebrow raises or head nods are added on prominent words and not on neutrally pronounced neighbouring words. This leads to a proposal that this relationship between acoustic prominence and facial gestures arises from a shared muscular synergy deriving from some sort of shared muscle activation during articulation. Bolinger (1985) previously formulated this as the up-and-down metaphor. The metaphor suggests that when pitch goes up and down, the eyebrows (or the head) tend to follow accordingly. But as shown by Cavé et al. (1996), this is not always the case. Not all prominent words receive gestures, but those that do tend to be prominent. This is the case when pronouncing version 5. People have no problem in adding a gesture or not doing so on the prominent word. This has been also noted by Ekman (1979) who stated that people, on many occasions, do not mark the emphasis in their speech with batons (local facial prominence) or underliners (facial prominence extending beyond one word).

Although this relationship might not be physiologically mandated, research on the production of audio-visual prominence has further described it. From a production point of view (in Beskow et al. 2006), it was found that when subjects make one word highly prominent in a phrase, all facial parameters during the pronunciation of that word exhibit larger variation, this includes not only the articulators (lips, jaw) but also head and eyebrow movements. Krahmer & Swerts (2007) found that when subjects are asked to apply a gesture to a word, the word tends to be produced with higher prominence.

From the perception point of view, seeing an eyebrow raise or a head nod to a word has been shown to enforce the perception of prominence over that word regardless of it was pronounced prominent
or not (House et al. 2001). Also, words with the same level of acoustic prominence are perceived more prominent when seen in synchrony with an eyebrow raise, compared to the same words without a synchronised gesture (Krahmer & Swerts, 2007). This perception is maintained when the acoustic and facial prominence are in synchrony. House et al. (2001) found that de-synchronisation is tolerable up to ~200ms, an average length of a syllable. This provides evidence for the role of facial prominence being additive, and that acoustic and facial prominence play a similar nonverbal emphatic role on linguistic segments.

2.2.1. FACIAL PROMINENCE AND SPEECH INTELLIGIBILITY

We have demonstrated that there is considerable evidence that eyebrow and head movements having a function during speech in synchrony with an underlying linguistic segment. This function is nonverbal and reflects the nonverbal prosodic properties of acoustic prominence.

A clear use of facial prominence gestures in animated talking heads is to enhance the communication of acoustic prominence by playing an emphatic, additive role. But is the function of these gestures only prosodic? Does the perception of facial cues of prominence extend beyond making a segment prominent? These are questions that are investigated in Paper A.

The study in Paper A is motivated by the hypothesis that acoustic prominence could, in fact, impact on the interpretation of the underlying linguistic message by, for example, the correct placement of pitch accents and prosodic boundaries (e.g. Cutler & Otake, 1999; Terken & Hermes, 2000), the question then is whether facial prominence can fulfil that function as an alternative to acoustic prominence. Paper A presents two experiments. The first one is a perception experiment that aims at examining the effects of facial cues of prominence on speech intelligibility. The second experiment utilises these effects in an interaction study to examine, objectively, the effects these gestures have on subjects, when looking at talking heads supported with them.
i. **Audio-visual Intelligibility of Degraded Speech**

To inspect the perceptual contribution of gestures seen in synchrony with acoustic prominence, to speech comprehension, the intelligibility of the acoustics needs to be degraded so that there will be room for enhancement. This is a standard methodology to examine the contribution of different variables to speech intelligibility (Erber, 1974; Summerfield, 1992).

As acoustic prominence is a prosodic property of the auditory signal, and we expect gestures to play a similar role to acoustic prominence, we, in the experiment, aimed to degrade not only the intelligibility of the speech sounds, but also the prosody of the signal. To achieve this, instead of adding noise to the signal, a noise-excited vocoder was used. The vocoder applies band-pass filtering and replaces the spectral details in the specified frequency ranges with noise, and in principle removes the fundamental frequency information from the signal.

One property worth mentioning about such a vocoding is that it has a large training effect. Subjects listening first time to vocoded sentences perceive them with very low intelligibility; however, after a small amount of exposure, the intelligibility of the auditory signal increases considerably, and it converges to a maximum. For this experiment, subjects were introduced to a training session before any experimental condition.

ii. **Gestural conditions**

To examine the differences in the intelligibility of sentences with different gestures added on a talking head, three different gestural conditions were used. In all conditions, a talking head (developed by Beskow, 1995) was used. The lips of the talking head were animated and synchronised with the speech signal using the phonetic labels of the audio files.

- **No gesture**: the baseline used is a talking head that only moves its lips.
- **Head nods**: a fixed design of a head nod is used.
- **Eyebrow raise**: a fixed design of the eyebrow raise gesture is used.
These three choices were taken to compare the contribution of a gesticulated talking head to a static head, and to compare whether there are differences between a head nod and an eyebrow raise in terms of intelligibility.

### iii. Temporal placements

Facial prominence gestures should in principle be synchronised with the prominent segments of the audio signal. According to House et al. (2001) these gestures should not be out of synchrony in more than 200ms at maximum. This synchrony is important for the gesture to enforce the acoustically prominent syllable, rather than its neighbouring ones.

The conditions in terms of alignment, which were used in the experiment, are as follows:

1. **No gesture:** no eyebrow or head gestures were added to the stimuli. The talking head moved its lips only in synchrony with the speech.
2. **Random eyebrow gestures:** eyebrow raise gestures were placed over randomly selected syllable nuclei.
3. **Eyebrow gestures on prominent syllables:** prominent syllables were annotated using a native Swedish speaker, and the eyebrow raise gesture is cantered temporally on the middle of the prominent syllable.
4. **Head nod on prominent syllables:** This is similar to the previous condition, except for using a head nod instead of an eyebrow raise gesture.
5. **Eye brow gestures on pitch accents:** Pitch accents contribute largely to the perceived level of prominence of the syllable they are placed on. The automatic identification of pitch accents (or alternatively here, steep pitch slopes) is relatively easy, and their co-presence with beat gestures has been shown to be strong. This condition placed eyebrow raise gestures on automatically detected steep pitch movements that took place during a syllable nucleus (the vowel of the syllable).

### iv. Contributions to Intelligibility

The results of the experiment show that gestures placed on prominent syllables enhance the intelligibility of the audio-visual sentences
significantly beyond a talking head moving only its lips. Head nods have a relatively higher contribution than eyebrows. Eyebrow movement on pitch slopes do, as well, increase the intelligibility of the sentences compared to a random temporal placement of the gestures or to not using gestures at all.

These results are interesting in several ways: people cognitively are able to infer some aspects of the linguistic message using information given by the movement of the head or the eyebrows. What this study shows is that part of this contribution is prosodic. Through the information about the prosodic structure of the sentence, information about the verbal content of the sentence is retrieved. This can be because gestures synchronised with prominent syllables might give clues about lexically stressed syllables in words, so helping in the segmentation of the sounds into words, or in the selection of word candidates (Cutler & Otake, 1999). There is also some neurophysiological evidence that matching visual information speeds up the neural processing of auditory information (van Wassenhove et al. 2005), which might result in a higher temporal resolution of the signal and in turn increasing intelligibility.

Another interesting finding is that the random placement of eyebrow gestures did not lower the intelligibility compared to a non-gesticulated face but rather slightly increased it (however the difference was not significant). One would have perhaps expected that the misplacement of these gestures might trigger the wrong interpretation of the message, but this does not seem to take place. A possible explanation to this is that a more life-like talking head could lower the cognitive effort needed to read the face. Another explanation is that the random gestures were never added during silence, and hence they could help the listeners to pay attention to the speech signal.

These findings, in addition to showing that visual cues of acoustic prominence can aid speech intelligibility, also quantify this effect through the use of a minimal model of fixed head nods and eyebrow raise movements on well-defined instants in time. Thus, we have shown a clear benefit of generating these gestures either from a speech signal as input (using some automatic prominence detection methods), or using prosodically-rich text in speech synthesis systems.

Empowering an animated talking head with facial gestures aligned with acoustic prominence therefore makes the face more intelligible,
but does it enhance the perceived quality of the talking agent? This was the question investigated in the second experiment of Paper A.

2.2.2. FACIAL PROMINENCE: FROM PERCEPTION TO BEHAVIOUR

From a verbal perception perspective, facial cues of prominence, as found in the previous perception study, do enhance speech intelligibility. The methodology of the study can be looked at as a low-level objective evaluation method on the verbal perceptual differences of the different gestures. But one can also go a step further and ask whether the talking head supported by facial gestures of prominence is perceived as more natural, engaging, believable, and human-like.

The issue of evaluating talking heads, or Embodied Conversation Agents (ECAs), is a central one to the development and deployment of the technology, and has recently been addressed in trials to formulate assessment protocols of the different instantiations, capabilities, and behaviours implemented in them.

In general, two techniques on the evaluation of talking heads in interaction with users can be applied: subjective or objective. Subjective evaluation techniques aim at asking the subjects about their qualitative experience and perception of the different questions under investigation. Objective evaluation techniques measure the effects the talking head has on the users, such as whether users react physiologically and behaviourally in a similar manner when interacting with real humans. Although subjective evaluations can acquire answers to more complex properties of the talking head, they are highly affected by the variability of the users’ understanding of the questions (e.g. questions such as “how natural is the talking head?” depend significantly on the user expectations and understanding of the question, and the target user group). It is also not straightforward to guarantee that all subjects interpret the questions in the same way they were designed to mean. Objective measures on the other hand aim at the target impact on the user during interaction. The problem, typically, is that objective measures might not always be easy to acquire (e.g. the emotional state of the user during interaction). Another problem is that people sometimes do not interact with the system in accordance to the quality of their experience. For example, users can be shy and friendly to a system despite disliking its behaviour (Höök, 2004).
i. Methodological considerations

In the second study described in Paper A, we wanted to address the question of whether people perceive the talking head differently when it visualises beat gestures (as eyebrow raise and head nods) compared to when it does not.

When looking at a talking head that only moves the articulators in synchrony with the speech, one can expect that the user might concentrate visually on the mouth area of the face, since no dynamic information is transmitted anywhere else. This is expected to be the case especially when the audio signal is noisy. Hence, a possible objective measure is to examine the gaze of the user at the talking head. That is, how, where and when people look at the topology of the face in different conditions.

To minimise the interplay among variables in the experiment, a monologue setup is used where a user looks at a talking head that is reading out an audio-book. The audio-book speech signal was automatically annotated with pitch accents.

The experiment included three stimuli presentation conditions. In each one, the subject listened to the audio-book for 5 minutes. One condition used the talking head with only lip movements, the second one with lip movements and facial gestures on pitch accents, and the third one with a completely static talking head (this condition was later disregarded as subjects stopped looking at the screen when realising the face did not move).

The speech signal was degraded using babble noise to reduce the intelligibility and increase the reliance on the visual modality. Tobii state-of-the-art gaze tracker (Tobii T120\(^1\)) was used to monitor how people looked at the face spatially and temporally during the listening in the different conditions. The tracker was embedded in the same screen where the talking head was visualised.

The study included 10 moderately hearing-impaired subjects (m=71.2, std=5.3) who were introduced to a short questionnaire of 5 questions (subjective evaluation) after each condition.

\(^1\) http://www.tobii.com/
**ii. Face-reading patterns**

The gaze fixations and saccades were analysed from the gaze-tracking data for all the subjects and for both conditions “with gestures” and “without gestures”.

From the results, the difference between both conditions is clear: As expected, for a face moving only its lips, the users focus most of the time on the lips (68% of the time), while looking very little at the eyes (7%) and 25% of the time looking outside the mouth-eyes region. However, for the gestural condition, the pattern is relatively different. Subjects look less at the mouth area (60%) and spend significantly longer time looking at the eyes (38%) and hence looking only 2% of the time outside these areas.

Studies in the literature report similar patterns of “face reading” gaze patterns among people in audio-visual monologues (Vatikiotis-Bateson et al. 1998; Raidt et al. 2007). When looking at someone speaking, humans typically look at the eyes-mouth area in a triangular pattern. This pattern is maintained even with high noises, and the amount of time spent looking at the mouth, even with high levels of noise, stays at around 50% of the time.

This gives support to the hypothesis that visualising prominence synchronised gestures in the shape of eyebrow raises and head nods not only delivers the nonverbal and verbal functions of these gestures, but also creates a talking head that is perceived objectively in a more similar manner to a real human face compared to a talking head that only has its lips animated. The subjects further confirmed from the questionnaires that the face that exhibited the gestures was perceived more natural, easier to understand, and more helpful.

**iii. Limitations of the study**

This study examined the face-reading gaze patterns of subjects when they were looking only at two versions of the face: one without gestures and one with eyebrow, and head gestures. The differences are clear between the two versions. But from a modelling point of view, there are several factors that might be interesting to quantify. Here is a list of some of the factors that can be tuned further in future studies in relation to what has been done in the experiment:

1- **Gesture frequency**: in the gestural condition, two thresholds are used to generate the gestures. The first one is the maximum
number of gestures to be added per second, and the second one is the threshold of the steepness of the pitch accent on which there can be a gesture. One might think of these as parameters that describe the personality of the talking head, such as how “gestural” or “expressive” the face is. That perception may be affected by increasing or decreasing the number of gestures synthesised. Another way of synthesising these gestures is in an adaptive manner where the amount of gesture depends on dynamic factors in the interaction, such as the content being spoken, the behaviour of the person interacting with the face, or the noise levels in the environment (discussed in the next section).

2- **Gesture design**: The gestures employed in the study had a fixed duration, and were hand-crafted to look natural. However, the shape of the gestures during natural speech can vary. One can think of dynamically synthesising the length and shape of the gesture depending on other factors. Such factors can be the length of the syllable (or the linguistic segment) it is supposed to be in synchrony with, or with the shape of the pitch accent underlying it. Such strategies can be modelled using data-driven gesture generation techniques.

3- **Gesture choice**: In our study, the type of gesture (i.e. head nod or eyebrow raise) is randomly decided over time. The patterns of face reading observed in the study might change if the face synthesised only one type of gesture rather than a combination of them. Future studies can be designed to investigate the effects of the different gestures on the gaze patterns of the observer. Such conditions can aim at discovering these patterns with faces that only move the eyebrows, the head, or both.

4- **Time of gesture**: The study investigated only the effects when gestures are synthesised in synchrony with pitch accents. It does not examine whether the face reading patterns would change if gestures were to be temporally randomised (similarly to the Perlin noise described earlier), as these gaze patterns might be a result of how life-like the face is perceived and they might not depend specifically on the functions of these gestures. However, even if synthesising temporally randomised gestures would result in similar patterns, the temporal choice of these gestures in this study is functional in the linguistic sense of the word. These beat gestures deliver verbal and nonverbal information, and as seen in the first study, do increase the intelligibility of the speech signal when produced by the talking head.
2.2.3. POTENTIAL EXTENSIONS

The difficult question regarding *when* and *which* beat gestures are to be synthesised (or when people use them) is one that is not completely answered yet. These studies, however, give some additional clues on the use of such gestures.

Aside from the more complicated factors of predicting the shape and timing of a beat gesture, and since these gestures do enhance speech communication and comprehension, one clear and immediate application of these findings is noise-adaptive facial synthesis. In noisy environments, the intelligibility of the speech signal decreases when the noise level increases. As we reviewed and found earlier, there is much evidence that prosodic prominence is multimodal, and prominence can, to a large degree, be transferred via the audio and the visual modality. One possible way to decide on the frequency of the gestures is the level of noise in the environment. The more the auditory noise is, the more prominence cues can be transferred through the face. This would enhance the intelligibility of the communicated message according to the amount of noise that is degrading the message auditorially. This can be used in talking heads, whether they are driven by a human speech signal (in mediated interaction) such as in SynFace (Salvi et al. 2009) by automatically estimating prominence in real time from the speech signal, or when it is driven by an autonomous agent using the prominence labelling of the synthesised speech. Figure 2 shows an illustrative chart of such a possible noise-adaptive signal.

![Diagram of noise-adaptive gestures synthesis setup.]

*Fig. 2. An illustrative scheme of a noise-adaptive gestures synthesis setup.*
As figure 2 illustrates, the talking head can listen to the environment in real time and have some estimate of the auditory noise. Depending on the level and type of the noise, the system can also estimate the loss in the prosodic information in the auditory signal and compensate for that by adapting the facial gesture generation algorithm in real time to compensate for this loss (Fitzpatrick et al. 2011).

2.3. MODELLING ACOUSTIC PROMINENCE

For a system to supported a synchronised gesticulation of prominent segments in an input speech signal (speech-driven gesticulation), models and algorithms should be in place to estimate the prominence of every segment in the speech signal, and depending on that measure, the system should be able to decide when to add a gesture in sync with that segment.

One can also, by using speech corpora, predict which word in a text message should be made acoustically prominent or not, and from that prediction, synthesise that text audio-visually with support for prominence.

The estimation of the level of prominence of linguistic segments is a research area that is receiving increasing interest from different perspectives with different goals in mind. As prominence concerns the question of how the linguistic segments are contrasted and consequently communicated, measures of acoustic prominence are useful. For example, measures of prominence can be used in automatic speech recognition systems, with the syntactic and semantic parsing of speech, in grounding in dialogue, in automatic corpus annotation for expressive speech synthesis, in language learning, and in speech therapy (Hieronymus et al. 1992; Nöth et al. 2000; Tepperman & Narayanan, 2005; Wang & Narayanan, 2007; Al Moubayed et al. 2009b; Ballard et al. 2010; Wik & Granström, 2010).

In the work presented in Paper B and Paper C, we are interested in the task of estimating and classifying prominence from speech. The purpose of this is the choice of appropriate points in time in a speech stream for the synthesis of beat gestures associated with prominent segments, and thus to render a more intelligible and natural facial animation of virtual characters.

Acoustic prominence is a subjective perceptual measure of how salient a linguistic segment is. This is a qualitative measure that needs to
be quantified in order to develop methods that are able to automatically estimate it from speech.

Researchers have been trying to describe acoustic prominence in different languages in terms of its acoustic correlates (see Horne, 2000 for a review), and recently some trials have been relatively successful in building models for the automatic classification of prominence (e.g. Tamburini, 2003; Wang & Narayanan, 2007; Obin et al. 2008). The two studies presented in this thesis are the first trials on automatic prominence estimation in Swedish. However, we believe that the issues they address are applicable and of interest to other languages.

The development of machine learning methods to estimate prominence from speech is a task that is complex in several ways, and it demands special treatment. Below we present a list of some of the issues, when trying to estimating prominence from speech, and when taking the Swedish language as the case study, in relation to two studies presented in Paper B and Paper C.

- The underlying linguistic segment that can be prominent is varying in length: depending on the function of that prominence, segments as short as phonemes and as long as sentences can be emphasised.
- Certain types of prominence are lexical by definition (such as lexical stress), and hence do not carry a nonverbal function.
- In focal accents (the highest level of word prominence in Swedish), the perception of prominence by humans is on the word level, however the acoustic correlates to the realisation is on the syllable level (for details refer to Paper B). This makes it more difficult to define a time window for, first annotating speech data, and second, extracting the relevant acoustic features.
- There is no agreed and inclusive scheme for annotating large speech corpora with prominence, that is not relatively time intensive (partly due to the previous reasons).
- It appears that the perception of the level of prominence over words is highly variable among non-expert annotators. Paper B reports agreement numbers among 4 annotators for word-level prominence, which are similar to the accuracy numbers achieved by non-linear models built from the same dataset.

To avoid several of the complexities described above, while keeping in mind two things: 1) that the purpose of the system is the
synchronised generation of beat gestures on prominent segments, and
2) being able to build such models with a limited amount of resources to
collect annotated speech corpora. To approach that, several measures
and restrictions have been taken into account in the studies in Paper B
and Paper C, in order to arrive at an operational model of prominence
using a small dataset of only 200 sentences:

- Prominence is annotated only on a word level. This was the choice
  in order to give a preference for an estimate that is more
  perceptually relevant over the risk of having feature level
  inconsistency.
- Both studies compare syllable and word-level features. However,
syllable-level features are integrated indirectly into a word-level
measure that allows for an evaluation against the annotated data.
- The selected speech dataset was chosen from a corpus of read-
speech that is recorded for the purpose of speech synthesis. 
Although this set is limited in terms of spontaneity and variability, it
is expected to be more consistent with world level prominence
rather than longer linguistic segments.
- Only three levels of prominence are annotated (prominent, maybe
  prominent, and not prominent). Four annotators were used and a
  democratic vote (the average vote among the annotators) is
  considered as the level of prominence for that word.

2.3.1. NON-LINEAR MODELLING OF WORD PROMINENCE

Paper B explores the use of two machine-learning methods: Support
Vector Machines and Memory Based Learning. The study treats the
three categories of prominence in a ranked fashion (prominence=2,
maybe=1, no prominence=0). The main findings of the study show that
the strongest predictor of word prominence is the length of the word
(the second was the number of syllables per word). These measures are
merely lexical and have nothing to do with how the word is acoustically
realised. Although this result can be directly used in text-to-speech
synthesis systems to predict timings of acoustic prominence, using text
as input, it cannot be used on input speech as it does not rely on the
acoustic properties of the actual speech signal.

The strongest acoustic predictor of prominence was the syllable
duration. Syllable duration has been repeatedly reported as a strong
correlate to prominence (Fry, 1955; Fant et al. 2000; Terken & Hermes, 2000). However, the system using all syllable-level features combined resulted in 62% accuracy, not much higher than the 56% baseline.

### 2.3.2. Linear Modelling of Word Prominence

Aiming at achieving higher accuracy rates using acoustic features, Paper C proposed a simplistic linear model of prominence. The model mainly looks at prominence as an outlier: whenever a segment diverges statistically from its average. The model looks at prominence on a continuous scale (regression), opposed to the categorical ranking used in Paper B (classification), which results in a slightly different measure of accuracy.

As the model is linear, it disregards any non-linear relationships among the acoustic features, but in principle it requires lower amounts of data to converge.

The results show that the linear regression model outperforms the classification model — syllable-level acoustic features combined result in high accuracies (86%), compared to 79% for word-level features.

These findings allow for the use of syllable-level features to estimate word-level features, while also identifying the “most prominent” syllable in that word. This allows for the use of a real-time automatic gesticulation system requiring only phonetic labeling of input speech, a requirement that is typically available in an audio-visual speech synthesis system (such information is usually used for real-time lip synchronisation of input speech). This requirement will also allow the system to run without the need for any word boundary segmentation or lexical features of input words.

It needs to be mentioned however that the assumptions used for the estimation of the linear model (as a weighted average of the individual measures of prominence for each acoustic feature) might not account for more accurate and complex definitions and models of prominence. This needs to rather be looked at as an operational definition that allows for an identification system of prominent segments over time with minimal requirements. That was done to allow for the use of facial cues of prominence in speech-driven facial animation, without the need for any lexical, semantic, or any higher-level linguistic information.
PART III

GAZE AND THE PERCEPTION OF GAZE DIRECTION
Related Papers

**Paper D**

**Paper E**

**Paper I**
3. Gaze and the Perception of Gaze Direction

While facial prosodic movements are to some degree correlated with the speech signal, and so provide relatively redundant information during face-to-face interaction, gaze movements during communication are not. Gaze (where gaze means the combination of eye and head movements) is possibly the strongest complementary facial signal to speech and is a signal of high spatial value.

There is a large body of research studying the different functions of eye and head movements in face-to-face communication, supporting the avenue of evidence for how important the role of gaze is in affective and social communication. Here we give some examples of such findings as motivations for our work, although these will not be elaborated on in the thesis.

Klienke’s review article (Klienke, 1986) on the functions of gaze in social interaction contains the following list: (a) provide information; (b) regulate interaction; (c) express intimacy; (d) exercise social control; and (e) facilitate service and task goals. According to Kahneman (1973) gaze indicates three types of mental processes: spontaneous looking, task-relevant looking and looking as a function of orientation of thought.

Bloom & Erickson (1971) found that infants establish purposeful eye-contact at an age as early as 7 months. Waxer (1977) found that gaze movements correlate with emotions (anxiety levels) in patients. Bente et al. (1998) quantified significant differences in gaze movements relating to attention across sex and familiarity of dyads. Frieschen et al. (2007) provided a comprehensive review of gaze cues of attention in the infant, adult, and clinical population.

Researchers and developers of virtual characters have realised this importance of gaze and gaze movements. Thus, there is an increasing amount of work aiming at exploiting the functions of gaze movements for the modelling of the behaviour of embodied human-machine interfaces (e.g. Poggi & Pelachaud, 2000; Khullar & Badler, 2001; Bilvil & Pelachaud, 2003; Lee et al. 2007; Lance & Marsella, 2010; Andrist et al. 2012; Oertel et al. 2012).

For an intelligent ECA to communicate with its users/interlocutors in a human-like fashion, the design of gaze behaviour in this interface and the perception of this generated behaviour should be well understood.
As gaze is essentially a signal of spatial value, an accurate perception of gaze direction is required for the intended functions of gaze to take place and the signals to be perceived. Although there is a considerable research on the quantification of the perception of human gaze, there is little research targeting the perception of the avatar’s gaze by humans.

One of the main limitations for an accurate perception of the direction of gaze is that the virtual character (the avatar) is usually seen through a two dimensional display. This brings several effects to the interaction compared to a face-to-face human-human situated interaction. We explore and discuss some of these in the following sections.

3.1. FACE-TO-FACE COMMUNICATION OVER THE SCREEN

As example of interacting with people over a two-dimensional display that almost all computer users of today have experienced is video conferencing (e.g. Skype™). The main limitation of a typical video conferencing setup is the absence of a space of shared attention, or as we will call it here spatial co-presence. This limitation ranges from the difficulty in establishing eye-contact (due to the location of the cameras in relation to the eyes of the two parties) to the difficulty to refer to objects in the environment of the other conversant. These limitations led many researchers and companies to investigate and to develop often highly complex solutions to video conferencing, especially when it comes to group conversations (e.g. Gemmel et al. 2000; Vertegaal, 1999). A detailed discussion of these limitations and a proposed setup is included at the end of Paper D.

In addition to the effects 2D displays impose on face-to-face conversation, which can indeed be measured and quantified using perception experiments (such as the ones in Paper D and Paper E), there are other higher-level social and affective changes that arise in the communication over 2D displays. These effects come from the restrictions of movements, orientation and sensory information that they impose on interaction. These effects are more difficult to measure and quantify using perceptual and behavioural analysis.

Kappas and Krämer in their book Face-to-Face Communication over the Internet (Kappas & Krämer, 2011), present a collection of psychological studies that demonstrate some of these effects. Manstead et al. (2011), for example, discuss this from a co-presence
point of view. They state that: “the use of display screens, whether they are on the desk or in the hand, means that interpersonal distance is changed (reduced) compared to normal co-present interactions”. This is due to the miscommunication of gaze direction, mutual gaze, and proximity to the interlocutor.

These changes result in the possibility that the signs of intimacy and affect present in these signals are miscommunicated.

These previous studies provide insights into what can be expected from a “fully capable” virtual agent communicating with its users via 2D screens.

3.2. THE MONA LISA GAZE EFFECT

The Mona Lisa gaze effect is the one in which the direction of the gaze of a portrait is perceived to follow the observer while she moves around the portrait. This is commonly described as an effect that makes it appear as if the Mona Lisa’s gaze rests steadily on the viewer as the viewer moves through the room. This is a result of that objects seen in 2D displays have no direction that is enforced in the 3rd physical dimension (Todorovic, 2006). This effect takes place in the perception of the direction of any object shown in 2D displays.

![Fig. 3. An illustration of the Mona Lisa gaze effect. All three versions of the portrait appear to gaze at the observer. Although that the two sideway images are rotated, the viewer perceives a mutual gaze with the portrait.](image-url)
Looking at the perception of gaze, the Mona Lisa effect in a portrait means that if the gaze is facing forward in the portrait, the observer will perceive a mutual gaze with the portrait no matter where the observer is standing in relation to the portrait. If the portrait is facing forward, and there are multiple observers at the same time looking at the portrait, all observers will perceive a mutual gaze (eye contact) simultaneously, and this perception will be maintained even when the observers move around in relation to the portrait. If the gaze in the portrait is turned sideways, no observer can establish a mutual gaze with the portrait, even if they move around it. An example of this is shown in Figure 3. The two sideway images of the Mona Lisa are rotated in the virtual 3rd dimension, simulating a side view of the portrait. The figure shows that the portrait is perceived to look frontal (mutual gaze with the viewer) even when it is rotated sideways.

This effect has significant implications on the design of avatars that are set to spatially interact with humans. Avatars that are visualised on 2D displays and set to interact with multiple users will be limited in their use of gaze and orientation in the regulation of multiparty interaction. The same would apply to avatars that are developed to point to objects in the environment of the user.

It is important to remember here that one of the highly important advantages of using gaze (and orientation of ECAs in general) is to support speech interaction. Gaze provides a tool to facilitate the integration of spatial information in spoken dialogue. However, 2D surfaces significantly affect the perception of the direction of gaze and hence limit the avatar from exploiting one of its more important assets. 

*Paper D* explores the Mona Lisa gaze effect and further discusses its implications on the interaction with ECAs.

### 3.3. **Avoiding Mona Lisa gaze with 3D projections**

While the Mona Lisa gaze effect occupies faces (or rather any object) visualised on a 2D display, physical 3D objects are Mona Lisa-free. As with sculptures, physically situated objects have an absolute orientation that is independent of the observer’s viewing angle.

To be able to bring the avatar out of the traditional two-dimensional space, and into the three-dimensional interaction environment, we proposed in *Paper D* to project the avatar onto a static 3D face-shaped mask. Projecting the graphics of the avatar’s face onto a 3D face-shaped
mask enforces the orientation of the face thanks to the three-dimensional shape of the mask. To do this, we used a Pico-laser\textsuperscript{2} projector to project frontally the animated face onto a generic mannequin face model.

The technique of projecting a moving picture onto a static object (manipulating static objects with light) is sometimes referred to as \textit{Shader Lamps} (Raskar et al. 2001; Lincoln et al. 2009). This is done in order to change the appearance of static objects by illuminating them using static or moving images.

To test the hypothesis that using a projection of the face on a face-shaped mask would eliminate the Mona Lisa effect and create a perception of gaze direction similar to that of looking at physical objects, we carried out an experiment that targeted quantifying the perceived direction of gaze by observers when looking at the face (using the 2D and 3D displays) from different angles.

### 3.4. Gaze Direction in 2D and 3D Projections

The goal of projecting the animated face on a three-dimensional mask is to allow users to perceive the gaze of the face in absolute terms, and so, no matter where the observer is looking, the gaze of the face would point to the same target point in the space of the observer.

To measure the perception of the gaze direction of the projected face, we introduced an experimental setup that consisted of 5 subjects who were simultaneously seated in front of the projection surface. The subjects were seated at an equal distance from each other and from the projection (shaping an arc around the projection surface) with the middle-seated subject facing the projection surface.

In the study, we manipulated the gaze angle of the face horizontally, and subjects were instructed to note down which subject the face was looking at for each randomised gaze angle. This setup allowed us to collect large numbers of stimuli answers in a short time (i.e. five subjects gave five answers for each gaze stimulus point). This design can in general be used in perception experiments of directional signals with the ability to compare the perception of different sitting angles simultaneously for a single stimulus (as in Edlund et al. 2012a and Edlund et al. 2012b).

\textsuperscript{2} SHOWWX Pico Projector.
We hypothesised in Paper D that when using a 2D display, the following would persist: 1) when the face is looking frontal, all subjects will perceive the face looking at them, and 2) when the gaze is turned sideways, no one will perceive a mutual gaze, but all subjects will perceive the gaze to be averted to the side at an equal rotation. This means that if the face is looking 20 degrees to the right, all subjects would perceive the face to be looking at the subject to their left hand side (assuming subjects are seated at 20 degrees angles from each other). We also hypothesised that in the 3D projection: all subjects will perceive the face looking at the same target no matter where they are seated (e.g. if the face is looking frontal, all 5 subjects will perceive the face to be looking at the middle subject).

To account for possible subjective differences between the subjects, we permutated the seats so that each subject was seated on all the 5 seats.

The results from the experiment conformed to our predictions to a high degree: the projection of the face on a 3D physical mask, almost fully, eliminated the Mona Lisa gaze effect, allowing for an accurate and absolute perception of gaze direction.

From the perception data collected in the setup, we were able to estimate psychometric functions of gaze direction that can map gaze rotation in the avatar’s eye to the perceived rotation in space. These functions make it straightforward to calibrate gaze rotation for optimal use in situated interaction.

3.5. **Mona Lisa gaze in multiparty dialogue**

Despite the perceptual effect of the Mona Lisa gaze, can people involved in multiparty dialogue cognitively compensate for it and continue a regulated interaction? This was the question of Paper E.

One can argue that even if people subjectively perceive the direction of the avatar’s gaze dictated by the Mona Lisa gaze effect, when in multiparty interaction they can still infer the intended gaze target, and maintain a fluent multiparty dialogue. If they indeed do so, then it would be less justifiable to abandon the solution of 2D displays for multiparty setups. For example, a multiparty receptionist ECA that is embedded in a screen is described in (Bohus & Horvitz, 2010). The receptionist was designed to interact with only two interlocutors at a time and it accurately addressed the intended person 86.2% of the time.
However, a setup with only two possible interlocutors is a special one. If the avatar looks to the right, the person standing to the left is likely to answer, even if she perceives the avatar to be looking to her left side, but as no one is standing further to her left side, she would be the likely addressee of that gaze. The same goes for the person standing to the right side. In such a setup, the avatar avoids looking frontal as this would result in an eye-contact with both interlocutors.

Hence, we wanted to investigate this effect with a larger group of interlocutors. The physical setup we chose in Paper E was almost identical to the perception study of the Mona Lisa gaze (presented in Paper D). Five subjects were seated in front of the display surfaces. Instead of asking the subjects to individually note on a sheet the number of the subject that the avatar is perceived to be looking at, the subjects were instead involved in a multiparty dialogue task where the avatar asked the subjects questions, and the only indication of who is addressed by the avatar was gaze direction. The avatar’s verbal behaviour and timing was controlled by a human operator.

The result of the study was that people, to a large degree, cannot infer the intended addressee when seeing the avatar on a 2D surface. Only 50% of the time did the correct addressee pick up the turn, compared to 84% when the 3D display was used.

Although 50% is probably too low for a functional multiparty dialogue, it is certainly higher than a random guess (20%). It seems that people can to some degree learn the intended target of gaze despite their basic perception. However, the study interestingly shows that there is an indication of a cognitive/confusion cost attached to this compensation. On average, the subjects were 0.47 seconds slower in picking up the turn when a 2D display was used (1.85 seconds for the 2D condition compared to 1.38 seconds in the 3D condition).

The study also featured a small questionnaire targeting the quality of the projected face. Subjects indicated that the 3D projected face was perceived significantly more human-like and easy to read compared to the face projected on a flat display.

The results from the studies in Paper D and Paper E were promising in terms of creating an avatar that is suitable for situated/multiparty interaction where directional information can faithfully be perceived by interlocutors.
The idea is relatively simple: using a face-shaped mask and a micro projector, one can build a physically situated animated head that is suitable for situated human-machine interaction.

Compared to the traditional use of robotic heads (as the physical counterpart to animated faces), projecting a face onto a static mask is also a relatively cheap solution to implement. It also requires less maintenance, is light-weight, and comes with low power demand (Paper D shortly reviews some of the benefits of the approach in comparison to virtual animated faces and to mechatronic robotic heads).

The success of this approach, in addition to its ability to render a more believable animated face, has been the driving force behind the development of Furhat: a fully functional prototype of a back-projected animated face. Furhat will be described in Part IV of this thesis.

3.6. Gaze direction for situated interaction

The use of gaze as a cue of direction in situated interaction isn’t restricted to regulating multiparty turn-taking. The previous two studies have targeted the perception and effects of gaze direction between 2D and 3D displays for turn-taking. The two studies used gaze as a cue of the person of attention of the animated face.

When it comes to situated interaction, gaze as a cue to direction is not only limited to regulating turn-taking. As technologies allow systems to be more aware of their environment (using sensing technologies such as scanning, tracking and object detection), embodied agents that are capable of communication using gaze have significant potential in interaction applications with users, where objects in the environment can be included in the interaction. A common setup in which we envision intelligent agents to be used is human-agent collaborative task solving. One can imagine a setup where the robot and the human have a shared space of attention that is part of the interaction (in the form of a table for example), where objects of interest are located.

While being able to regulate turn-taking is important, interlocutors are usually located at a relatively large distance from each other. This might be to some degree forgiving in regard to the granularity of the perceived direction of gaze to infer who is being looked at. It however might not be the case if the objects of interest are located in close vicinity of each other. In the study in Paper I we explore a setup that has
been used to study human-robot interaction using gaze. We hypotthesise that objects-of-interest are located on a table that is shared between the robot head and the human subject. In this setup we aimed to explore the perceptual cues of the direction of gaze over different conditions using an analysis-by-synthesis setup.

The study utilised the Furhat robot head, a back-projected head that relies on the same paradigm of projecting an animated face on a mask, which was used in the previous two studies on gaze. (Refer to Part IV of this text for more information on Furhat.)

For the setup, instead of placing objects on the table surface, and to be able to generate gaze stimuli with a high resolution, we used a printed square-grid paper sheet. This was so that the possible target point of interest (squares on the grid) is sensitive enough to gaze direction that would allow us to quantify the error of the perceived direction with a high level of accuracy.

The test subject and the robot head were seated at different sides of the table. For every time the robot head changed the gaze direction to a different square on the grid; the subject had to indicate which square was perceived to be the target point of gaze of the robot.

3.6.1. Selected Results

The results from the experiment shed light on some of the cues that affect the perceived direction of gaze which are not commonly taken into account in the design and control of generated gaze in robot and animated agents in situated face-to-face interaction. The findings also show that certain cues of gaze need to be calibrated for when used to communicate spatial points of attention. The calibrations can be used either when an agent is used as a front-end to generate gaze movements or when gaze is controlled using, for example, gaze tracking in tele-presence applications.

In the following we summarise some of the findings of the experiment presented in Paper I.

iv. Eyelids

Eyelids are obviously affected by the rotation of eyeballs. By the placement of eyelids, the visibility of the sclera (eye-white) is affected. However, despite the simplicity of implementing natural movements of the eyelids in avatars, it has been mainly ignored in the synthesis of
animated agents (Bailly et al. 2010). Using animated agents also allows for testing the effects of the location of the eyelids using the perception-by-synthesis approach of experiment. It is noteworthy to mention here that it is not possible to test such effects using human subjects since the movement of eyelids is involuntary (i.e. humans cannot normally control for it).

Our results show that using eyelids (that is, when the eyelids are connected with eye movements, and follow dynamically the eyes as they move vertically) affects the perceived direction of gaze compared to when the eyelids are static.

In humans, the eyelids follow the eyes as they rotate vertically and they are located exactly on top of the iris. When used in avatars, it results in a significant shift of 1.47 degrees downwards in the perceived direction of gaze, in relation to when the eyelids are static (fixed for eyes looking frontal).

v. **Head-pose**

The perception direction of gaze is biased by the head orientation (Langton, 2000) and the coordination of head and eyes orientation is a key signal to the attention of their owner. One can focus on an object of interest either by 1) keeping their head static and rotating their eyes towards the object; 2) by keeping their eyes looking frontal while rotating their head towards the object; 3) or by a combination of both. We examined in the experiment whether there are any differences in the perceived direction of gaze when using only the eyes to gaze at the target, or when using only the head. The result shows that keeping the head frontal (providing best visibility of both eyes) while moving the eyes towards the target point results in a higher accuracy (resolution) in the perceived direction of gaze compared to when directing the head towards that point. This difference was found to only affect the perception on the vertical axis. In fact, using the head instead of the eyes to gaze at a point provided the worst cue of gaze compared to all other conditions. We hypothesise this to be a combination of two possible reasons: 1) moving the head vertically might significantly hinder the visibility of the eyes, or 2) the head is very large in size compared to the eyes, and can be less accurate to infer the direction of the head compared to the direction of the eyes.
vi. **Side viewing**

The experiment also found that it is less accurate to read the direction of gaze when looking at the eyes at an angle (compared to looking at them frontally). When sitting at a 45-degrees angle sideways from the head, it was 1.15 degrees less accurate to infer the direction of gaze compared to when sitting frontal. This means that it might be important to calibrate the gaze function of the agent depending on where the interlocutor is sitting to optimise the readability of the direction of the avatar’s gaze.

vii. **Perception of human gaze**

In addition to the synthesised conditions, the experiment examined the accuracy in perceiving the direction of gaze when the gaze is generated by a human rather than by the robot head. The experiment only used one human confederate to generate gaze targets and so does not allow for generalisations from the findings. However, for the specific person that took part in the experiment, it was consistent that subjects had an error in perceiving the direction of gaze that is generated by the confederate. There was a shift in the perceived direction, equalling an average of 3.5 degrees vertically and 2.8 degrees horizontally, when compared to the actual gaze target. Nonetheless, when compared to the accuracy of perceiving the gaze direction of the robot head, it was found that reading the direction of the eyes of the human confederate was the most accurate. One can hypothesise that if the shift in the perception of the direction of the human gaze persists (by additional supportive findings from future studies), it might be possible that by calibrating the gaze of the robot, one can establish a gaze function that results in a more accurate perception of the robot’s than the human eyes.

3.7. **TAMING THE MONA LISA GAZE TO OUR ADVANTAGE**

Effects such as the Mona Lisa gaze that result from using two dimensional displays have important implications on the design and modelling of interaction which need to be taken into consideration when using gaze (and directional signals in general) with embodied conversational agents. Notwithstanding the Mona Lisa gaze effect does hinder and limit the interaction using gaze and the suitability of
orientation cues in multiparty or situated interaction, its advantages have been harvested, whether intentionally or not, by many. For an avatar set to communicate with a single user via a flat display, the avatar can obtain eye-contact with the user by simply looking forward. The avatar can also avert its gaze by any set rotation angle discarding where the user is standing in relation to the avatar or whether moving or standing still, being assured that the direction of gaze is perceived as intended.

Many studies have taken advantage of this by investigating the cues and effects of gaze control in perception and interaction experiments using avatars in 2D displays with no need to worry about the situatedness of the user in relation to the agent (for examples, Hjalmarsson & Oertel, 2012; Skantze & Gustafson, 2009; Edlund & Beskow, 2009).

**Paper D** further discusses the requirements of different interactive systems in terms of faithfulness of gaze direction depending on the setup and the presentation surface.
“'I wonder what makes us build inefficiently-shaped human robots instead of nice streamlined machines?’”

'Pride Sir.' - Said the robot

Terry Pratchett (The Dark Side of the Sun)
Related Papers

**Paper F**

**Paper G**

**Paper H**
4. Furhat – A Back Projected Robot Head

As shown in the previous studies and discussions, the paradigm of using a static physical head model as a projection surface for animated computer models would not only bring the face outside of the traditional two-dimensional screen, but will also eliminate the Mona Lisa effect and allow for multiparty and situated interaction. From the studies we reviewed earlier, it also appears that people perceive the projected face as significantly more natural than the face shown on the screen.

From a robotic perspective, using the animated computer model as an alternative to a physical robot head solves major difficulties for building naturally looking and moving robot faces; the technology behind facial animation has significantly advanced in recent years, and the control of these faces is very simple and highly flexible. (See Paper F for a review).

Encouraged by the findings from the studies in Part III, and in order to build an avatar face that allows the implementation of situated and multiparty dialogue, in addition to performing research on situated facial signals, we have built Furhat, a back-projected robot head.

In the following section we review a brief history behind the use of back projection for animated faces, and we follow that by a description of Furhat.

4.1. A BRIEF HISTORY OF PROJECTED HEADS

The first description of a face projected onto a face-shape mask that we were able to track is in a patent by Jalbert (1925). Jalbert presents a sketch of a back-projected head used to project faces of real people captured with different facial expressions. The model Jalbert suggested also used mirrors to transfer the images from the projector to the bust. One can trace several other, more recent, patents in relation to Jalbert’s and in one way or another altering certain aspects of the setup (e.g. Liljegren & Foster, 1990; Machtig, 1993; Monroe & Redmann, 1994; Moore, 2002; Jackson, 2010;).

The first traceable implementation of a face projection is the Grim Grinning Ghosts at the Disneyland Haunted Mansion ride, which opened in 1969. The ghosts are displayed simply by projecting a previously
recorded movie of the faces that matched 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 mounted on a pan-tilt gimbal that was driven by the subject’s head movement. The image of the face of that person was then projected onto a mask of that same person (referred to in Naimark, 2005).

Morishima et al. (2002) built the prototype system HyperMask which aims to project a face onto a mask worn by an actor. The mask is tracked in real-time and the projection is adjusted accordingly to guarantee a real-time fit that is seen by the audience.

Recently, more groups have put efforts into creating back-projected interactive heads for human-machine interaction.

Hashimoto & Morooka (2006) use a hemi-spherical surface to back project an image of a humanoid face. The head, although projected on a spherically curved surface, still gives the impression of co-presence in comparison to a flat screen.

Lighthead was built by Delaunay et al. (2009) at the University of Plymouth and is a back projected robot head that is more elaborately designed. It maintains a stylised cartoonish face and is supported by a robotic-arm. Lighthead was built for research on nonverbal interaction.

Mask-bot, created by Kuratate et al. (2011), projects a synthetic photo (photorealistic synthesis) of a real face onto a generic plastic mask, demanding calibration of the image to fit the mask design. Figure 4 shows snapshots of the Lighthead and Maskbot.

4.2. Furhat – A Back-Projected Robot Head

Furhat is a back-projected head that uses a mask that is a 3D printout of the face projected onto it. The face model used in Furhat is relatively human-like in anatomy and size, and it utilises a 3D parameterised animated model (Beskow, 1997) that does not use any textures obtained from real faces (the same model used in studies in Paper A, Paper D and Paper E).
Part IV – Furhat - A Back Projected Robot Head

Fig. 4. Snapshots of Lighthead and Maskbot – two different back-projected robot heads.

Furhat uses a 3D printout of the same face model that is projected onto it so this guaranteed a perfect fit of the different parts of the face to the surface. The mask had several modifications to allow for the animation to be visible with a minimum misalignment with the mask. The lips in the mask were substituted by a curvature that does not enforce the shape of the lips. This was done to allow the projected lips to be perceived as moving. The eyelids of the mask were also flattened to allow for the projected lids to move (in blinks for example). The head was then equipped with a micro projector supported by a fisheye lens. All this fitted to a 2 DoF neck that allowed for head movements.

Furhat is built to study, implement, and validate 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, which is something that cannot be achieved using virtual avatars displayed on flat screens. Paper E reviews the studies behind the motivation to build Furhat and compares it to other alternatives. It also describes in detail how the head was built to allow other researchers to replicate it. Figure 5 shows some photographs of Furhat, for more photographs, please refer to Paper E.
Fig. 5. Photographs of the Furhat back-projected head.

4.3. **Reading Furhat’s Lips**

Creating realistic talking heads requires the visible articulators to appear in synchrony with the speech they produce. This is done not only to enhance the illusion that the talking head is the source of the sound they produce, but also for the crucial role the face plays in speech perception and comprehension.

The speech group at KTH has a long tradition of building talking heads for the hearing impaired, taking advantage of the contribution of the visible articulators to enhance the audio-visual intelligibility of the animated faces. For Furhat to measure up to its on-screen peers, its lips must fit this criterion.

Although Furhat uses an animated face with that same articulatory model, its plastic mask is static (so are its jaw and lips). This fact might introduce a spatial inconsistency and non-alignment between the projected image and the mask.

*Paper H* introduces an experiment that was designed to investigate whether the fact that Furhat’s mask is static might affect the intelligibility of the animated face. This is done by evaluating the audio-visual speech intelligibility of Furhat against the same animated face but visualised on a flat screen.
A main difference between interacting with a face shown on a 2D or 3D surface, as we described earlier, is the fact that 2D displays have no direction and maintain the same visibility of what is shown on them no matter from which angle they are seen. 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 if there is an optimal face-reading angle of the face, the face on a 2D screen can maintain that angle and guarantee optimal intelligibility. This is not the case 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 – if looking at the face with an angle degrades the intelligibility compared to when looking at it straight frontal, this would introduce a negative effect on the intelligibility. To investigate this, the study also included visibility conditions of a 45 degree visibility angle. Figure 6 shows the intelligibility of the test for the different conditions.

Regarding the audio-visual intelligibility of Furhat, the study found that the intelligibility of Furhat’s lips measures up to and even exceeds the intelligibility of the same model shown on a 2D screen. This was the case both when looking straight at the display and at a 45 degree angle.

This verified the suitability of the approach of back-projected animated faces onto a static mask, in terms of the contribution of facial movements. The fact that the intelligibility is even increased by Furhat is indeed positively 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 to avoid enforcing a static shape of the lip. Because of this the size of the lips, when projected, is perceived slightly larger than the lips visualised on the screen. The shape of the lip area of the mask can be seen in Figure 7. 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 was easier and more comfortable to look at Furhat). Furhat’s face is spatially situated and more human-like than a 2D screen, and we speculate that because of the high sensitivity of the test on the cognitive state of the subject,
Fig. 6. The average percentage accuracy rates for the different experimental conditions.

using Furhat might have increased the level of focus and attention of the subjects and this resulted in a more efficient lip reading.

Regarding the question of audio-visual intelligibility in relation to the viewing angle, as mentioned earlier the study found no significant difference between the two displays. This is the case when reading human lips. Erber (1974) found that the lip-reading contribution drops down when looking beyond 45 degrees, although it is not significantly different between 0 and 45 degrees. We confirm this finding in our study. Looking straight frontal or at a 45 degree angle does not introduce a significant difference in the contribution of lip-reading, which means that a 2D display surface effect does not have an advantage over a physically situated 3D head, at least for these two angles. This difference is going to probably change significantly at angles beyond 45 degrees. However, such effects simulate situated face-to-face human-shuman interaction settings more accurately.
4.4. Furhat at the London Science Museum

Furhat was developed as part of the European Commission project IURO (Interactive Urban Robot\(^{3}\)). As part of this project, we were invited to participate in the RobotVille festival at the London Science Museum\(^{4}\), between the 1\(^{st}\) and 4\(^{th}\) of December 2011. The robot festival featured some of the most advanced robots from different research laboratories in Europe.

This invitation provided a perfect opportunity to test Furhat in a real-life setting, where the system could interact with hundreds of members of the public from diverse backgrounds, both genders and different age groups.

For the museum setup, we were interested in firstly testing the head’s ability to carry out spoken multimodal dialogues with multiple people simultaneously, and to explore strategies and signals to keep the multiparty engagement levels high and the dialogue going. Paper G presents the dialogue setup at the exhibition and provides some results on the interaction strategies employed using Furhat. We here summarise and further explore some aspects of the design and evaluation of the system.

Developing and testing ECAs in a museum-like environment is not new. It is rather popular and common to put spoken interactive systems to the test and into the wild with the public, which provides opportunities to showcase futuristic technologies, to collect data, and to

---

\(^{3}\) IURO (Interactive Urban Robot) No. 248314. http://www.iuro-project.eu

\(^{4}\) http://www.sciencemuseum.org.uk/
learn patterns of interaction behaviour, all thanks to the mass exposure such setups attract.

There are several examples of employing ECAs in similar environments in the literature. The multimodal spoken dialogue system August (Gustafson, 2002) was used to collect spoken data for more than half a year in 1998 at the Cultural Centre in Stockholm, Sweden\(^5\). August could answer questions about restaurants in Stockholm. More than 10,000 utterances were collected from 2,500 visitors.

Pixie (Gustafson, 2002) collected data from museum visitors, starting in 2002 and lasting more than two years. Pixie was part of the futuristic exhibition "Tänk om!" ("What if!"), which consisted of a full-scale future apartment, in which Pixie appeared as an assistant and an example of an embodied speech interface. Pixie was introduced to the visitors in a movie portraying a future family living in the apartment. The visitors were also encouraged to ask Pixie general questions about her and about the exhibition. The resulting corpus contains about 100,000 utterances.

In 2004, the life-sized multimodal dialogue system Max was displayed for several years in the Heinz Nixdorf Museums Forum, a public computer museum in Paderborn, Germany (Kopp et al. 2005). Max took written language as input and responded with synthesised speech. In its first seven weeks at the museum, Max recorded over 50,000 inputs.

Finally, Ada and Grace, a multimodal spoken dialogue system designed as virtual twins, first greeted the visitors to the Museum of Science, in Boston, US, in December 2009 (Swartout et al. 2010). The twins acted as museum guides and spoke both to each other and to the visitors. In early 2010, the twins collected over 6,000 utterances in just over a month.

4.4.1. **MANOEUVRING AROUND THE BUMPS**

Getting a conversational interface from the lab and into uncontrolled real-life environments is no easy ride. Capture technologies that appear to work well in a constrained and controlled laboratory are unpredictable and often fail to perform in more natural settings. To get around the limitations of the different technologies involved in the

\(^5\) http://kulturhuset.stockholm.se/
system and arrive at an operational interaction that can target useful research questions, many design tricks and solutions are often employed, and in many cases, while unavoidable, these solutions limit the degree of intuitiveness and naturalness of the systems.

For example, to manage multiparty dialogue in the lab, we use Microsoft Kinect™’s depth camera to track the possible interlocutors with the system, and we use Kinect’s microphone array to capture hands-free directional speech. In the museum, with very high levels of noise, as well as often tens of simultaneous users and changing lighting conditions, these technologies will fail. To solve this, we opted to avoid the use visual input to the system and instead employed two ultrasound proximity sensors that were mounted on two fixed podiums. On the podiums, we placed two microphones, making it clear that for the system to work, the users need to stand by the podium and speak to Furhat into the microphone. Whenever a user approached the podium, the proximity sensor informed the dialogue manager of the presence of a user, and Furhat turned its head towards the podium and engaged the user in a dialogue.

However, the bottle-neck technology in spoken interactive setups is often speech recognition itself. A system that is supposed to interact with the public needs to have an ASR that is age-independent, gender-independent, speaker-independent, accent-independent, and noise robust for the “planned” interactive scenarios to take place. Under these requirements, most systems arrive at design solutions to increase the reliability of the speech recogniser. From the above mentioned examples, Pixie used a push-to-talk button to get input from users, Max used written text as input and speech only as output, and Ada and Grace opted to use a human agent. This agent had experience speaking to the system but also knew what questions are supposed to be addressed to Ada and Grace, and how to ask them. Through this agent, users could direct questions to the system.

Furhat in London needed to talk to multiple people at the same time, while utilising different verbal and nonverbal signals to regulate the interaction between the users and the system. This means that the above solutions are insufficient as many conversational phenomena had to be taken into account, such as interruptions, overlaps, multiparty turn-taking, etc. Indeed, one would expect that for such subtle social signals to even take place in a dialogue, a minimum level of fluency and naturalness needs to be present during the interaction.
To avoid the need for high recognition accuracy and to make the system robust against recognition errors without significantly limiting the fluency and spontaneity of the dialogue, we designed the dialogue to focus on collecting information from the users rather than merely to answer their questions. The dialogue relied on survey questions about the public perception of the future of robotics. The system asked the questions and used keyword spotting to extract certain information from the answers. Whenever successful, the system updated statistics charts of the answers on an LCD monitor that was mounted next to Furhat.

To regulate the interaction, Furhat had a bag of strategies at its disposal. *Paper G* explains the setup at the museum in detail and describes the different strategies employed to regulate a mixed initiative multi-party dialogue with the system. Al Moubayed et al. (2012a) describe in more detail a list of the tricks used in the design of the system.

### 4.4.2. Facial signals in interaction

Furhat used a set of facial signals in the dialogue to regulate the interaction and communicate the status of the system with the user at any point in time. We here list some of the strategies employed in the system. Figure 8 also shows a conceptual categorical list of the functions of the different signals.

- Whenever Furhat was speaking, its lips were automatically animated in synchrony with the speech.
- At all times, Furhat blinked randomly (non-functional blinking) to imitate human biological blinking behaviour.
- When no user was present in front of the system, Furhat randomly alternated his gaze downwards to signal an idle state of the system. Whenever a user walked up to a proximity sensor, Furhat turned its head towards that user, while keeping the eyes looking frontal to maintain a mutual gaze with the user, as an indication that the user has Furhat’s attention. Furhat then initiated the dialogue with a greeting.
When a user left the interaction, Furhat turned its attention to the other user in case a second user was present; otherwise, the system turned back into an idle state.

When Furhat was listening, it kept a mutual gaze with the speaking user. Whenever the user started speaking, Furhat would raise its eyebrows and maintain a gentle smile signalling that the system is listening to the users’ input. During the dialogue, when Furhat did not understand the input speech, it would give feedback to the user with a “yeah” or “okay” that had a positive or a negative prosody. When doing so, Furhat raised its eyebrows when giving positive feedback, and frowned when giving a negative feedback.

When Furhat was speaking, it occasionally alternated its gaze downwards; whenever Furhat finished speaking, it fixated its gaze on the user waiting for input.

**Fig. 8.** A list of the different uses of facial parameters at the London Science Museum setup.
• In the presence of two users at the same time, Furhat would alternate questions to the users by directing its head to an addressee. If the wrong user answers a question or interrupts, Furhat would glance with its eyes at that user and say something like “please wait a second”. Furhat would also occasionally pose open questions to both users. This was done by directing the head to the middle, and alternating the eyes between both users. The system would then attend to one user by measuring the audio levels of the microphones to detect the speaker who grabbed the turn.

• The dialogue was designed so users can pose questions to the system at any point during the interaction (hence mixed initiative). One of the questions users could ask Furhat was to change the colours of its different facial parts (face, eyes, eyebrows, lips). This is a unique feature for a robot head which is possible because Furhat is back projected and changing the colours is software based. The capability to change colours was one of the more popular features for the kids who interacted with the system.

4.4.3. DIALOGUE MANAGEMENT

To orchestrate the dialogue, a newly developed event-driven dialogue authoring framework was used. The framework (now named IrisTK) is based on a notion of state-charts (Harel, 1987). Perhaps one of its most powerful features is the ability to define hierarchical states in the dialogue, minimising the code and the management of events at different levels. The framework is explained in details in Skantze & Al Moubayed (2012).

The dialogue system was designed on a modular basis. Modules interact with each other using event triggers. The dialogue consisted of two main modules, namely Dialogue Management and Interaction Control. Our choice of this division was to allow the dialogue management to abstract its processing from the surface structure of the input/output events, while being able to define the processing of input and output behaviour independently from the dialogue content. The multiparty interaction control’s job was to transfer input events from surface structures (depending on what sensory devices are used) into dialogue-meaningful events (such as the translation from a Proximity-On event, into a User-Enter event).
The second main job of the multiparty interaction control is to manage the behaviour of the face, by translating action-events into surface behaviour (such as the transfer from an Attend-All that is responsible to attend both users, into a facial behaviour that decides how the head and gaze should be controlled).

Figure 9 shows a simplified flow chart of the system’s components. Figure 10 shows two snapshots of Furhat and the audience during interaction at the London Science Museum.

### 4.4.4. Major findings

The exhibition at the London Science Museum enjoyed a very large visibility. Furhat was seen by almost 8,000 users during the 4 days of the robot festival. During the 4 days, the system interacted with hundreds of users. All interactions were audio-visually recorded, and all system events were time-logged.

During the exhibition, a selected set of users were chosen to fill out a questionnaire that focused on their experience with the system, targeting more specific aspects of the quality of the interaction and the system design.
Fig. 10. Snapshots of two different interactions with users at the London Science Museum.

i. **Turn-taking strategies**

From the analysis of the logged data, 92% of the questions were correctly picked up by the user Furhat was directing its gaze towards, showing very high accuracy of head orientation and gaze for the regulation of multiparty dialogue. Open questions that were addressed to both users (head directed towards the middle and gaze alternated between both users) also resulted in very high neutrality. 54% of these questions were picked up by the user that was previously speaking, showing no bias towards either of the users.

ii. **Multiparty engagement**

When Furhat did not understand the user’s input, it occasionally passed the question for *elaboration* to the other user saying something like “what do you think about that?” or “can you elaborate on that?”, or it passed the question for *agreement* to the other user, saying something like “do you agree with that?”. The analysis shows that these strategies were relatively successful in keeping the dialogue going and encouraging the subjects to provide more information to the system. However, repetitive elaboration requests exhibited the lowest efficiency. Only 40% of double-elaboration requests were answered by the users.

iii. **Corpus collection**

The exhibition resulted in a multimodal corpus of about 10,000 utterances, out of which were 3200 question-answer pairs. The data that was collected varied in terms of the age of the participants
(estimates of age groups were made in Blomberg et al. 2012), and in gender.

A Word-Error-Rate (WER) calculation was made for the different groups of users, showing very high recognition error rates for children and women (80%-94% error rate), while utterances provided by the staff gave the lowest error rates (~30% error rate). These numbers show how biased speech recognisers are to certain speaker characteristics. In addition they show the big improvement in system accuracy when users are trained to talk and address the system.

The corpus now is being used for research on building models that are female and child friendly through the use of spectral transformation techniques. For more details, see Blomberg et al. (2012).

**iv. User experience**

During the exhibition, 86 questionnaires targeting the quality of the conversational capabilities and the interaction experience were collected. The questionnaire ranked different questions on a 5 point (1→5) Likert-type scale. 46 of the questionnaires were completed by male users and 39 by females.

The results show that people experienced a very positive overall impression of the system. People liked Furhat (mean = 4.08, SD .76). They enjoyed talking to it (mean = 4.13, SD .84), and they liked its response and feedback behaviour (mean = 3.80, SD .71). They also rated the conversation to be rather easy and fluent (mean = 3.17, SD .99). Despite the limited understanding of the system, the score for “did Furhat understand what you said?” was 2.99 (SD 1.05).

**4.4.5. Conclusions**

Furhat at RobotVille has received large and positive press coverage (including from BBC, CNN, Gizmag, and New Scientist). It has been frequently described as “sarcastic”, “witty”, and “intelligent”, giving some evidence on the success of the system design and interaction.

Furhat at the exhibition presents a unique study using a novel embodied conversational agent tested in an authentic environment, giving increased attention to situated aspects of dialogue with a rich use of facial communicative signals and multiparty control.
4.4.6. **Furhat: The Robot Head vs. The Virtual Face**

Is Furhat a robot head, or a virtual face? The answer to this question is very much a matter of context, perspective and definition. One can argue that Furhat is not an intelligent virtual agent (IVA) and is of more interest to research on robotics. This is to some degree true when an IVA is a character that lives in a virtual world where its environment and objects of interest are located and where its actions are performed (e.g. in gaming). However, that argument does not hold as much for IVAs made to communicate with physically situated humans and set to interact with their physical environment. On the contrary, such agents are at a disadvantage if they communicate with humans while living in a virtual space visible to the user through a flat display, which in fact was the primary motivation behind the development of Furhat.

Is Furhat a robot head? That depends on what is meant by a robot head and the purpose for which it is made. Furhat is an animated face that is projected onto a static face-shaped display. For example, if the robot head is designed to produce speech mechanically (i.e. a mechanical vocal tract) then Furhat is of no interest. However, the majority of robotic heads are made to look anthropomorphic and to communicate functions of facial movements, something that animated faces excel at. This is indeed why several robotics studies have stacked flat displays on top of their robots’ bodies.

So when it comes to situated embodied face-to-face human-machine interaction, whether Furhat is a robot head or an animated virtual head becomes a matter of definition, and perhaps what is more important than the definition is the design of behavioural models that lie behind the appearance, a research problem that is of great interest to both communities.
4.5. CURRENT WORK

4.5.1. FURHAT AT THE TÄLLBERG FORUM

As a result of the media coverage and the success of the setup at the London Science Museum, we were invited to present Furhat at the Tällberg Forum 2012, 14th-17th of June, 2012.

The Tällberg Forum is an annual meeting organised by the Tällberg Foundation\textsuperscript{6}, and it brings together about 300 world leaders and experts from different fields (scientists, politicians, debaters, etc.) to discuss pressing issues of world dynamics. This year, the theme of the forum was “how on earth can we live together?”. Furhat was invited by the organisers to take part in presenting a panel discussion on “Future beyond our imagination” and was meant to showcase an example technology that brings together science and art in a new form of interaction. Figure 11 shows a snapshot of Furhat during the panel discussions.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{furhat_panel_discussion}
\caption{A photograph of Furhat introducing a panel session during the Tällberg Forum 2012.}
\end{figure}

\textsuperscript{6} www.tallbergfoundation.org/
As part of the forum, we also showcased an interactive system with Furhat that is capable of having social dialogue with people, and which was set to interact with the visitors at one of the bars during the event.

As part of the setup, we wanted to bridge some of the limitations the system presented at London had.

One of the challenges we faced with setting up the system in London was how to practically detect and track users in real-time in order to carry out a fluent multiparty dialogue by looking at people even when they move. This had to be done in a very dynamic environment with an unknown number of users standing in front of the system. The partial solution we implemented in London was to use proximity sensors instead of cameras because a reliable face tracker that could operate on any number of faces and under unpredictable room lighting seemed out of reach. To carry out multimodal dialogue, one needs also to detect the speaker when input comes into the microphones. In London, we placed microphones that were attached to the proximity sensors so that whenever speech came from the microphone, the head turned always to the same angle.

This solution was partial and introduced limitations on carrying fluent interaction. Users who wanted to talk to the system had to stand at specific places and use specific microphones. Furthermore, the system did not know where exactly in space the user was standing, so if the user moved slightly the head did not follow and eye-contact was not sustained. As the system did not “see” the users, the system could not compensate for the height of the users either, but it was set to look at the same place and, as a result, eye-contact was not always established (although users were mostly able to infer when the system did address them).

We wanted to overcome these limitations at the Tällberg setup by enriching the visual input of the system to be able to track in real-time the speaker among all possible users standing in front of the system. To do this, we integrated the SHORE face tracking and analysis system developed by Fraunhofer (Kueblibeck & Ernst, 2006). Testing in the lab gave indications that SHORE offers high robustness and reliability in tracking multiple faces in real-time. The SHORE system also provides several nonverbal input signals from the faces it tracks such as estimates of age, gender and facial expressions.
In the current system, employing wireless microphone and face tracking has several advantages over the previous version of the system, mainly in its ability to allow free movement of the users in front of the system.

The dialogue in the setup has focused on the ability to carry social small-talk with visitors to the bar in an attempt to explore less task-oriented interaction with a robot head. A more detailed description of the setup at the Tällberg Forum is presented in Al Moubayed et al. (2012b). Figure 12 shows two snapshots of the setup at the Tällberg Forum.

Fig. 12. A photograph of the speaker-tracking system showing the real-time face-tracking in combination with the infrared-based microphone tracker.

Fig. 13. Two photographs of interactions with the system at the Tällberg Forum 2012.
The system that was showcased at the Tällberg Forum has received large national press coverage and been highlighted in a documentary produced by the Swedish national education channel. The system was also demonstrated (as described in Al Moubayed et al. (2012b)) at the International Conference on Multimodal Interaction (ICMI’2012) where it was awarded the “Outstanding Demo Award”.

The development of a rich multimodal multiparty social interaction with Furhat for the Tällberg exhibition is a step towards building more sociable and expressive companions. Such companions utilise state-of-the-art technology to enable the system to initiate and carry out a situated dialogue with rich multimodal input and output signals. We are constantly trying to reach that goal in attempts to make the system more fluent, natural and spontaneous, aiming at triggering behaviours from users that are more similar to the ones they exhibit when talking to other humans.

4.5.2. Furhats for all

The Furhat robot head is a state-of-the-art prototype of the technology of back-projected animated faces. As we discussed in this thesis, it has significant advantages over traditional mechatronic robot heads, in terms of implementation, cost and maintenance, giving it potential accessibility to the human-robot interaction HRI research community. It also solves several limitations of on-screen animated faces for the study on situated embodied interaction. The prototype has received much attention and interest from the research community as a research tool and as a novel human-machine interface.

Since we started working on Furhat, the prototype has been under constant development and optimisation, and it still is. Compared to the first version of the prototype, the current head is more solid, featuring a fish-eye lens making the size of the rig smaller. Also, the head is recently supported by a 3 DOF neck, allowing future development of more natural head movements and gestures.

We aim at building a more robust, versatile, compact in design, and flexible physical head that is cost efficient to produce and customise, as well as a set of software tools that give direct accessibility and control over the behaviour of the head and face.
“Some unspoken human communication is taking place on a hidden channel. I did not realize they communicated this much without words. I note that we machines are not the only species who share information silently, wreathed in codes.”

— Daniel H. Wilson, Robopocalypse

“If we know what it was we were doing, it wouldn’t be called research, would it?”

— Einstein
5. Challenges and Future Work

4.6. The Nonverbal Dilemma

One of the main characteristics of nonverbal behaviour that makes it very different from verbal behaviour is what is called the continuity distinction (Richmond et al. 2011). In human face-to-face communication, nonverbal communication is always present. As Richmond et al. (2011, p5) put it in a grammatically imperfect but rather meaningful way: “When you are in the presence of another human being, you cannot not communicate.” For example, silence or a stare gaze in a conversation might signal meanings that are more powerful than words. This fact is highly problematic and complex when it comes to developing and testing full-fledged human-machine embodied interactive systems. Many studies concern the function and behaviour of different nonverbal signals in human communication in certain setups and under certain conditions (as in this thesis). However, during interaction with users, the multitude of multimodal input variables humans take into account for shaping their behaviour is highly complex. The output behaviour in terms of nonverbal signals is also highly complex, and it contains several output signals that can be triggered in combinations. If multimodal systems ignore some of these signals or some of their functionalities (e.g. due to limitations in the technology and the state-of-knowledge on human communication), the effects of the multimodal signals the system generates might not reflect their intended functions, due to the large interplay among these interdependent signals (a clear example of this is sarcasm where facial expressions could reverse the meaning of the verbal message).

Taking into account all possible contextual and situated information during human-machine interaction is not possible yet. This is partly because of the limitations in capture and synthesis technologies and the difficulty in modelling certain contextual information such as personality, culture, background, situation, and so on. Until systems are able to model all the signals that take place during human-human interaction, ECAs developers need in principle to implement all the acquired knowledge from human-human communication in their systems when testing new signals, as testing these new signals
independently will not guarantee their exact function when used in combination with other behaviours.

Verbal communication follows a more-or-less clear coding system (language). Although researchers have been trying to construct such a system for nonverbal behaviour (as in some of the work included in this thesis), some believe that a main property of nonverbal behaviour is its ambiguity and unpredictability, and constructing a deterministic coding system for nonverbal communication (i.e. a nonverbal language) strips it from its very important characteristics. This is frequently observed by us in real-life interactions between users and ECAs. As long as users start to observe regular patterns in the behaviour of the system, the system is perceived as “programmed” and their interest in it drops.

It is clear that nonverbal messages follow a meaning and a coding system that underlies when and how they are communicated, but perhaps the inherent unpredictability and ambiguity of nonverbal behaviour is what turns a dialogue from a strict information exchange process into a socially intelligent conversation.

4.7. ECAs BEYOND FIRST IMPRESSION

Users do not always take social ECAs seriously. From our experiments with Furhat at the London exhibition, and from many trials in the lab, users communicate with Furhat in patterns that are not usual for human-human communication. For example, 14% of the questions addressed to Furhat were to challenge its knowledge, which included “what is the speed of light?” and “what is the meaning of life?”, while 17% were about his anthropomorphic characteristics such as “do you have a girlfriend?”, “can you see me?” and “are you happy?”. Most of this behaviour, of course, can be explained by the setup and function of the installation, and by factors of novelty and attractiveness of the head. However, it is not uncommon for conversations with ECAs to saturate quickly, and what first appeared to the user as an interesting intelligent entity becomes predictable and “programmed”. However, building meaningful and personal relations with ECAs is a requirement for turning them from an artistic and novel product for exhibitions and demonstrations into social companions.

There is little research on relations with ECAs beyond first encounter. This is understandable as experiments are usually done to measure differences in specific controlled variables in the system. This is
probably also due to the difficulty of carrying long-term interactions with ECAs that are purposeful.

An important research orientation then is how to build interaction models that hold over time, while maintaining the intended social and human-like characteristics. One possible direction to this is to put the social relationship at stake in the interaction, in a way similar to human-human interaction. This means that ECAs can adapt their behaviour and the shape of their social relationship depending on how humans interact with them. Humans are required to invest in their interaction to establish a positive relationship with the ECA. Nonetheless, there has been recent effort in the community to build adaptive interaction models that align to their interlocutor (Kopp, 2010; Bergmann & Kopp, 2012).

4.8. ECAs are handicapped

One interesting perspective for dealing with the development of different ECAs is by acknowledging their limitations in, perhaps, a medical fashion and in different implicit and explicit ways communicate this to the users. One can also take these limitations into account during developments and the design of the system.

For example, all up-to-date ECAs are to a large degree autistic, with symptoms of impairments in social interaction; impairments in communication; and restricted interests and repetitive behaviour (Rapin & Tuchman, 2008). Some ECAs that do not have cameras and cannot see their environment are blind. And because of the limitations of ASR systems they are hearing and language impaired. ECAs that do not build an adaptive model of their interactions and their interlocutors have some degree of amnesia. When looking at ECAs from this perspective, one can:

- Communicate the different limitations to the user (e.g. giving blind ECAs black shades on their eyes).
- Learn from persons with disabilities the strategies they employ to compensate for their disabilities, and allow them to establish some sort of healthy, constructive and social behaviour with other humans, and then implement these techniques in the systems.
- Find applications and uses for ECAs that are suitable despite their disability, from learning how different persons with disabilities function in society.
4.9. **Final Notes**

In many ways, the problem of facial synthesis and facial control is similar to the problem of speech synthesis. The task of generating a stream of lip movements (corresponding to the verbal message) has advanced and relatively matured during the last two decades. However, the synthesis of expressive-less articulatory movements is not enough to render an animated face that is suitable for interaction (as in synthesised speech with flat prosody). These also do not employ any signals that are relevant for the use in spontaneous dialogue and all the phenomena that take place during it.

Similarly to any other visible body movement during spoken interaction, facial movements are deeply connected to the spoken content, as spoken content is to the corresponding facial movements. The coordination of the spoken signal and the facial movement determines how the message is interpreted by human conversation partners and how the signals will affect this interaction.

The research presented in this thesis is a step towards building synthesised faces that communicate information beyond the stream of lip movements. The work contributes to the understanding of the design requirements of artificial faces and the modelling of facial signals and some of their functions, so that when implemented in multimodal embodied interfaces, result in more expressive, human-like faces which enable a more efficient and natural interaction with humans.

The area of multimodal embodied interaction is a relatively young, applied, and highly interdisciplinary research field. Arriving at its ultimate goal of creating socially capable and aware artificial humans is likely to be a long and bumpy road, and requires solutions to a myriad of difficult problems. It is also likely that many basic-research findings and useful technologies and applications will spin off and emerge along the way, similar to other fields such as computer vision and natural language processing. We believe that aiming at creating applications that exploit the technology despite its limitations, in addition to using it as a research tool to model human multimodal communication, will be a necessity in keeping the field active and prospering.


Jalbert, G., (May 18, 1925). *Lay Figure*. Patent, 1653180


Liljegren, G., & Foster, E. (Dec 18, 1990). *Figure with back projected image using fiber optics*. Patent, US 4978216


