Abstract

In this thesis, different aspects concerning how to make synthetic talking faces more expressive have been studied. How can we collect data for the studies, how is the lip articulation affected by expressive speech, can the recorded data be used interchangeably in different face models, can we use eye movements in the agent for communicative purposes? The work of this thesis includes studies of these questions and also an experiment using a talking head as a complement to a targeted audio device, in order to increase the intelligibility of the speech.

The data collection described in the first paper resulted in two multimodal speech corpora. In the following analysis of the recorded data it could be stated that expressive modes strongly affect the speech articulation, although further studies are needed in order to acquire more quantitative results and to cover more phonemes and expressions as well as to be able to generalise the results to more than one individual.

When switching the files containing facial animation parameters (FAPs) between different face models (as well as research sites), some problematic issues were encountered despite the fact that both face models were created according to the MPEG-4 standard. The evaluation test of the implemented emotional expressions showed that best recognition results were obtained when the face model and FAP-file originated from the same site.

The perception experiment where a synthetic talking head was combined with a targeted audio, parametric loudspeaker showed that the virtual face augmented the intelligibility of speech, especially when the sound beam was directed slightly to the side of the listener i. e. at lower sound intesities.

In the experiment with eye gaze in a virtual talking head, the possibility of achieving mutual gaze with the observer was assessed. The results indicated that it is possible, but also pointed at some design features in the face model that need to be altered in order to achieve a better control of the perceived gaze direction.

Sammanfattning

Den här avhandlingen innehåller studier av hur syntetiska talande ansikten kan göras mer uttrycksfulla. Hur ska vi samla in data för våra studier, hur påverkas artikulationen av uttrycksfullt tal, kan inspelade data användas på olika ansiktsmodeller, kan vi använda ögonrörelser i agenten för att kommunicera? Arbetet som beskrivs i den här avhandlingen inkluderar dessa frågeställningar, men också ett experiment där ett talande ansikte använts som komplement till riktat ljud, för att underlätta förståelsen av tal.

Datainsamlingen som beskrivs i den första artikeln resulterade i två multimodala corpora. I den följande analysen av inspelat data kunde det konstateras att

uttrycksfullt tal påverkar artikulationen, även om det behövs fler studier för att förvärva mer kvantitativa resultat och för att dels täcka in fler fonem och emotionella uttryck, dels för att kunna generalisera resultaten till fler individer.

När ansiktsanimeringsparametrarna (FAPs) byttes mellan två olika ansiktsmodeller, och också mellan två olika forskningsinstitutioner, kunde vissa problematiska områden identifieras trots att båda modellerna var av standarden MPEG-4. Den utvärdering som gjordes visade att det bästa igenkänningsresultatet erhölls när ansiktsmodellen och animeringsfilen var skapade på samma institution.

Perceptionsexperimentet där ett syntetiskt talande ansikte kombinerades med parametriskt riktat ljud visade att det virtuella ansiktet förbättrade talets uppfattbarhet, särskilt när ljudstrålen var riktad vid sidan av försökspersonen och därför var av lägre ljudstyrka.

I ett experiment med ögonrörelser i ett virtuellt talande ansikte studerades möjligheten att uppnå en upplevd ömsesidig ögonkontakt mellan det virtuella ansiktet och försökspersonen. Resultaten visade att det är möjligt, men pekade också på att viss design i ansiktsmodellen behöver ändras för att kunna uppnå en bättre kontroll av den upplevda blickriktningen.

Acknowlegdements

I would like to thank my supervisor Björn Granström and co-supervisor Jonas Beskow for not giving up on me during these years and for guiding me through the academic jungle.

I wish to direct a special thanks to my office-mate, Loredana Cerrato Sundberg. You have helped me with constructive ideas and inspiration, you always cheer me up and as if that was not enough you also feed me with fantastic Italian food. I couldn't wish for more.

Many others at the department have been a great help and good company during my time there. I especially want to thank David House and Rebecca Hincks for all your support and lovely smiles.

And of course, without my family and friends, none of this would have been possible.

Contents

List of papers

The papers that are marked with (*) are included in this thesis.

*Jonas Beskow, Loredana Cerrato, Björn Granström, David House, Magnus Nordstrand, Gunilla Svanfeldt (alphabetical order). The Swedish PF-Star Multimodal Corpora. *Proceedings of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces* (2004), pp.34-37.

*Magnus Nordstrand, Gunilla Svanfeldt, Björn Granström, David House. Measurements of articulatory variation in expressive speech for a set of Swedish vowels. In *Journal of Speech Communication 44* (2004), pp.187-196.

*Jonas Beskow, Loredana Cerrato, Piero Cosi, Erica Costantini, Magnus Nordstrand, Fabio Pianesi, Michaela Prete, Gunilla Svanfeldt (alphabetic order). Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces. *Proceedings of Tutorial and Research Workshop on Affective Dialogue Systems (ADS'04)*, pp.301-304.

*Gunilla Svanfeldt and Dirk Olszewski. Perception experiment combining a parametric loudspeaker and synthetic talking head. *Proceeding of Interspeech 2005*, pp. 1721-1724

* Mikael Nordenberg, Gunilla Svanfeldt and Preben Wik (alphabetic order). Artificial gaze. Perception experiment of eye gaze in synthetic face. *Proceedings of Second Nordic Conference on Multimodal Communication* (2005), pp. 257-272.

Loredana Cerrato and Gunilla Svanfeldt. A Method for the detection of head nods in expressive speech. *Proceedings of Second Nordic Conference on Multimodal Communication* (2005), pp. 257-272.

Jonas Beskow, Loredana Cerrato, Björn Granström, David House, Mikael Nordenberg, Magnus Nordstrand, Gunilla Svanfeldt. Expressive Animated Agents for Affective Dialogue Systems. *Proceedings of ADS 2004*, pp. 240-243.

Magnus Nordstrand, Gunilla Svanfeldt, Björn Granström, David House. Measurements of Articulatory Variation and Communicative Signals in Expressive Speech. *Proceedings of AVSP 2003*, pp.233-238.

Gunilla Svanfeldt, Magnus Nordstrand, Björn Granström, David House. Measurements of articulary variation in expressive speech. *Proceedings of Fonetik 2003*, pp.53-56.

Gunilla Svanfeldt. Röststyrt gränssnitt för planering inom ortopedisk kirurgi. Master's thesis, Department of Speech, Music and Hearing, KTH, 2003.

1 Introduction

In some cases, what you say is less important than how you say it. And how is not only determined by the tone of your voice, but also by the look on your face, your body posture as well as the situational context. These are all important factors in human communication, and are signals that we have learned to use with great skill. If one of these dimensions is omitted, the communication gets more vulnerable and the risk of misunderstanding increases. For example, many people prefer talking face-to-face rather than on the telephone, since the communication gets better when the visible signals can be read and interpreted. The human face-to-face interaction is a highly efficient way of communicating. In developing synthetic talking agents all possible cues should be used.

Speech with no expressivity (audible or visible) could be compared with written text in what may be conveyed. Only what is verbally expressed is mediated. The wish for expressing emotions and attitudes is evident, nowadays we even add smilies in text to express ourselves, such as ☺ or ☹. Expressiveness that can signal conversational cues is also an asset in managing the interaction.

In this thesis, some issues that have to be taken into consideration when studying expressiveness in artificial talking agents will be brought forward. The issues will be discussed in relation to other researchers work and the results of studies and experiments related to these questions will be presented. Hand gestures and other parts of the body are beyond the scope of this thesis and will not be discussed. The included papers cover issues like the collection of multimodal data and analysis of that data, results from an eye gaze experiment, a cross-cultural study with expressive synthetic faces and finally a perception experiment where targeted audio was combined with a talking head.

1.1 Talking heads

Virtual talking faces are useful in a number of situations. Visible articulation helps intelligibility of speech in unfavourable audio conditions (Beskow et al. 1997) as well as for hearing impaired persons (Agelfors et al. 1998). A new kind of listening situation is when using so called targeted audio, which can be described as a beam of sound. This can be achieved by for example using a technique known as the Parametric Array (Westervelt, 1963). An experiment where targeted audio and a talking head are combined is described in paper 4. It was found that the talking head was of great help in distinguishing between phonemes that sounds alike but are easily separated visually. However, the use of visual cues is not limited to only lip reading support. Non-verbal signals can be used in order to facilitate and regulate the conversational flow. A communicative face can signal turn taking, understanding as

well as misunderstanding, acceptance, questioning and other important aspects of human communication such as attitudes and emotions.

Talking heads can also be useful in learning situations. In second language learning, an agent has several advantages. Besides a possible unlimited practice time (the teacher never gets tired) and the absence of prestige that may be present in front of a human teacher, the agent can provide for augmented reality (Engwall et al. 2004). Massaro and Bosseler (2003) performed a study on how a talking head could influence the language learning for autistic children. It showed that the presence of the face contributed significantly to the language learning.

However, the use of an agent with a human face also puts demands on the behaviour of the system. A human face with no expressivity runs the risk of being perceived as cold and unpleasant, apart from lacking the efficiency that was desired in the first place.

Talking faces were used in paper 3, 4, and 5 in this thesis, and the models were all previously developed at the department of Speech, Music and Hearing. In paper 3 and 5, data-driven MPEG-4 face models were used (Beskow and Nordenberg, 2005). The main advantage with these is that they follow a standard. The MPEG-4 standard defines 68 facial animation parameters (FAPs) and 84 facial feature points (FPs) (Pandzic and Forchheimer, 2002). The same FAP values would produce the same result in all MPEG-4 compatible face models. However, the reality is not that simple, as was found in the cross-cultural experiment in paper 3.

In paper 4, a rule-based face model (Beskow 1995, Beskow 1997) was used in the experiment. The face is coupled to the wavesurfer software[1], which was employed for the acoustic output, using formant synthesis (Carlson et al. 1982). The parameter set that is essential for the visual speech articulation consists of: jaw rotation, mouth width, lip rounding, lip protrusion, labiodental occlusion, bilabial occlusion and tongue tip elevation.

1.2 The concept of emotion

When working with expressiveness, there are several almost philosophical questions that arise. Expressivity can be said to be both emotions (i.e. happiness, anger etc) and other nonverbal expressions, e.g. signalling turntaking, emphasis, and understanding.

The definition of emotion is debated within psychology, within the speech community the Ekman definition is widely used. This is a category based definition, and six basic (primary) emotions have been found to be universally recognisable (Ekman and Friesen, 1975). These emotions are defined as happiness, sadness, fear, disgust, surprise and anger. The non-basic emotions can be seen as blends of the basic emotions. The emotions can be expressed with varying intensity.

[1] www.speech.kth.se/wavesurfer

The dimensional approach is another way of describing emotions. Each state of emotion is then defined according to different scales, such as pleasure, arousal and dominance as proposed by Mehrabian and Russell (1974). The issues of debate within this approach are about the number of dimensions that are necessary to cover the emotional space, and what they should consist of.

The relation between the experienced emotion and what is expressed in the face is problematic to assess, which can make the eliciting and recording of facial expression difficult. How do we know what we are recording? Most psychologists agree on that a person cannot report about his inner state of emotion. There are too much going on that we are not aware of ourselves, and there is also to some extent self-censure. This means that some external observer must decide about the experienced emotion. But if an observer is to judge the subject's emotional state, it implies that the external expression is a reflection of the inner state. Which we don't know.

Another way to proceed is to use physical measures. But also with this approach we first need to know about the inner emotional state in order to tune the parameters. Otherwise we don't know what we are measuring.

Another perspective would be to disregard the inner state of emotion and only focus on the outside. What is important in a number of situations is that we succeed in sending the accurate signals and that the observer understands what we wish to convey.

Although Ekman and Friesen (1975) defined the basic emotions as universal, it must be kept in mind that both cultural and individual differences exist in the expression of emotion. Both the intensity and the way of expressing emotions are likely to differ and the conversational signals will be different depending on the culture. The individual differences are a smaller problem for synthesis of emotional expression than for emotion recognition and understanding. The reason for this is that synthesis can be considered successful if as little as one way of expressing emotions and communicative signals is realised, while a recognition system must understand all persons and all kind of expressions.

In order to create an expressive talking head it is desirable to analyse and model spontaneous emotion and conversational signals. But it is difficult to elicit and record spontaneous emotions under controlled conditions. There are of course ethical aspects as to whether to scare someone just to be able to record fright. Even to record positive emotions such as joy can be difficult. At ATR in Japan, a rather radical solution to this has been worked out (Campbell, 2001). One of the co-workers has recorded her speech in many everyday situations during several years. The speech can then be analysed for factors such as who she addresses, among other things. However, in this case audio only was recorded. A video recording would be much more intrusive.

We understand acted emotions, even though they might be different from the spontaneous. Some argue that having an artificial agent showing emotions is ridiculous since the user knows that the agent is not actually experiencing the

emotion, and is therefore not trustworthy. But let us compare with theatre or other artistic performances. The audience is very well aware of the fact that the actors play a role, but accepts and appreciate a good emotional and expressive performance. It will be as if the actors were actually sad, even though they were only pretending. But if the actors were not convincing, the whole play will be experienced as "false". This is important to keep in mind when developing expressive agents.

As can be seen in the included paper 1, we have tried some different strategies for eliciting emotions; acted emotions together with restrained speech in the first recordings, acted emotions with some more interesting phonetical coverage and finally more spontaneous dialogues, where focus was more on conversational signals rather than emotions.

1.3 Combining modalities

When adding modalities, the coherence between them is important. For example, if the voice sounds happy, the happiness should be visible in the face, and the other way around. If the face and the voice do not convey the same expression, the observer may be confused, and this may result in a completely differently perceived expression.

One example where characteristics of some different emotions were mixed in speech synthesis was in an experiment performed by Carlson et al. (1992). Sentences uttered with acted emotions ('happy', 'angry', 'sad' and 'neutral') were recorded and features as duration and pitch contour were measured. These features were then mixed in synthesized utterances and listening tests were carried out. It was found that only changing duration did not affect the perceived emotion, but in combination with changing the pitch more effect could be observed. Especially the 'sad' pitch contour was strong in affecting utterances originally spoken with other emotional expressions. Interestingly, when applying the 'angry' pitch contour to other emotional sentences, the result was also often perceived as 'sad'.

In an experiment by Beskow and Cerrato (2004), mismatches between voice and face of the emotions happy, sad, angry and neutral were studied. In this case the visual cues were found to be dominant.

When we communicate, we have the possibility of expressing the same thing in different ways – verbally (e.g. by explicitly saying "now it is your turn to speak"), vocally (e.g. final lengthening or intonation) and visually (e.g. mutual gaze or perhaps raised eyebrow). Sometimes we may use all of these, and sometimes only one or two. Besides, there may be a number of different signals that sometimes mean the same thing. Therefore, the study of certain phenomena is difficult since they do not always take place in a certain situation. The different modalities can compensate each other, but depending on how they are combined, they do not always convey the same meaning.

1.4 Expressive facial movements

The facial gestures that we use during conversation can be of various types. There are for example the different eyebrow gestures (e.g. frown, rise), mouth gestures (e.g. smile, protrusion), and eye gestures (e.g. gaze away, blink, stare). These gestures very seldom occur as separate gestures, but are more often displayed in combination with or following another gesture.

The study of expressive talking faces should take into account all movements in the face, including the lip articulation as described in paper 2 in this thesis. Fonàgy (1976) showed how the expression of emotions affected also other speech movements, e.g. tongue articulation. In Pelachaud et al. (1996) speech-rate and timing for intonation and co-articulation is modelled and some attention was also given to emotions. Magno Caldognetto et al. (2004) found that in the emotive speech some articulatory targets varied significantly with respect to the expressed emotions.

In a recent study the correlation between facial movements and focal accent was examined. It was found that facial dynamics were reinforced during a word in focal position for all the measured expressive modes (Beskow et al. 2006). The results also indicate differences in the facial dynamics between the different expressive modes.

1.5 Eye gaze

In human-human conversation, eye gaze is an important signal for regulating the flow of conversation (Argyle & Cook, 1976). For example, the speaker often looks away during the planning and while talking in order to keep the floor and then looks at the listener again when finishing talking, as a signal for handing over the turn to the interlocutor.

The amount of gaze also plays a role in how a person is perceived. Someone that seldom meets the gaze is seen as cold, unfriendly or shy, while a person that looks more at the conversational partner is perceived as friendly, credible and attentive. Of course, the latter is in most cases how we want our agents to be perceived.

However, eye gaze in synthetic faces does not only send social signals or serve as a conversational regulator; it may also be used as a deictic gesture, to point at specific objects on the screen. (Bertenstam et al. 1995, Heylen et al. 2005).

The work on developing the synthetic face in order to achieve a more efficient way of communicating, has been mostly concentrating on the speech, which means the lip articulation, and the rest of the area around the mouth. Some concern has also been given to eyebrows, for example to show emphasis (House et al. 2001). Eye gaze was also used to signal that the dialogue system was busy in e.g. the Adapt project (Gustafson et al. 2000), but the movements were adjusted manually and according to intuition, i.e. not based on data. An experiment in the Adapt project compared eye gaze gestures of the talking head with the use of an hourglass symbol. Although the results of the study did not show any increase in effectiveness of the conversation

using gestures, the evaluation showed that the users were happier with gaze gestures than with the hourglass symbol (Edlund and Nordstrand, 2002).

Several research teams have realised the importance of gaze, and are implementing gaze behaviour in various embodied conversational agents and talking heads (Park Lee et al. 2002, Garau et al. 2001). None have studied whether the interlocutor actually experience a mutual gaze with the synthetic face. Since mutual gaze is the basis for eye gaze behaviour, this aspect was in focus in paper 5.

1.6 Recording techniques

The recordings of data that was made as part of this thesis were carried out using an opto-electronic motion-tracking system – Qualisys' MacReflex. Reflective markers were glued to the subject's face and four infrared cameras were used to determine the position in three dimensions with a 60Hz frequency. The system has a sub-millimetre accuracy. In order to factor out head movements, a spectacle frame, with markers attached, is used.

Recording systems, based on a similar type of technology as Qualisys, have been used at other speech research sites, for example Elite (Cosi and Magno Caldognetto, 1996) and Optotrak (Cohen et al. 2002, Vatikiotis-Bateson et al. 1996).

Another recording technique is video image processing, although this was not used in the studies included in this thesis. Revéret and Benoît (1998) used video image processing to measure French visemes. A 3D lip model approach made the tracking of the lip contour easier. An advantage of the technique is the possibility of detecting lip closure with high accuracy. A disadvantage is that the lip corners often are in shadow which makes the measurement of the mouth width somewhat uncertain. Protrusion can not be measured accurately by such a technique, unless two cameras are used, one from the front and one from the side (face and profile). The advantage is that it is possibly less intrusive than an opto-electronic system.

The advantages of the opto-electronic motion-tracking system that was used in this thesis are the accuracy, the 3D positioning and the relatively easy data processing needed after the recording. The disadvantage is the intrusiveness of having markers attached to the face and the constraints that are imposed as to location and physical movement (e.g. the subject has to be careful not to block the infrared rays).

2 Summary of Papers

The papers presented in this thesis cover three different areas: data collection, analysis of the data, and evaluation and perceptual experiments. For each of the papers included in the thesis, a brief summary is provided followed by a description of author contribution and original contribution.

2.1 Paper 1

Beskow, J., Cerrato, L., Granström, B., House, D., Nordstrand, M., Svanfeldt, G. [2] **The Swedish PF-Star Multimodal Corpora**. *Proceedings of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces (2004)*, pp.34-37.

This paper describes the collection of three dimensional multimodal speech data. To capture articulation and other facial movements an opto-electronic motion-tracking system was used. Reflective markers were attached to the subjects face, and markers on a spectacle frame were used to factor out head movements so that facial movements could be traced. The acoustic signal was recorded simultaneously and was synchronised with the spatial data. Video recordings were also made.

Two corpora were collected, recorded with two different actors on separate occasions, and a slightly different placing of the reflective markers. In the first recording, markers were placed on the eyelids (both over and under the eye) in order to capture eye blinks, which could be part of the communicative signals. However, the system had difficulties in capturing the marker on the upper eyelid (the marker was sometimes hidden, especially during the frowning gesture) and did also confuse the upper and lower marker each time a blink occurred. The system requires a certain distance between the markers in order to perceive them as separate markers. The placing of the markers in the second recording was also to some extent adjusted in order to better correspond to the feature points (FP) in the MPEG-4 model (see section "Talking heads").

The first corpus consists partly of short prompted utterances, read with fifteen different emotions or expressions. The Ekman basic emotions as well as expressions more likely to occur in a dialogue system were recorded. The expressions selected were: anger, fear, surprise, sadness, disgust, happiness and neutral, followed by worried, satisfied, insecure, confident, questioning, encouraging, doubtful and confirming. The focus of this recording was to get a rough idea about what emotions or expressions would be useful in future studies, and to examine the possibilities of tracing the recorded facial movements. A pilot study based on this data was performed and is referred to in paper 2.

[2] Names in alphabetical order.

The other part of this first corpus consists of natural dialogues with an information-seeking scenario. The idea was to elicit more spontaneous facial expressions and communicative signals. The signals were produced in a conversational context in a less controlled situation.

The second corpus, where a different actor was recorded, consists of nonsense words and short sentences that are visually well balanced. Six of the emotional expressions in the first corpus were recorded: confident, confirming, questioning, insecure, happy, and neutral. The study described in paper 2 is based on this corpus.

The recognition of the expressions during the short sentences in this corpus was then evaluated in a screening test. The video recordings of the sentences were presented in random order without audio, so only visual cues were available for the subjects. An answering sheet, with seven options for each video file (the six expressions, and an extra category for "other") was filled in by each participant. The results showed that happy was well recognised, whereas the other expressions were more easily confused.

In retrospect, there are some things that could have been done differently. It would had been desirable to have the same actor in both recordings, since individual differences makes it difficult to compare the two recordings – the second recording was in that respect a new start that rendered the first not as useful anymore. Because of this, we would also have needed more material per emotion, for example more repetitions of the same sentence in order to get a more representative material.

Author contribution

The author of this thesis participated in both recordings, including the planning, in the following data preparation, the data evaluation and partly in writing the paper.

Original contribution

This was the first Swedish multimodal corpus for expressive states that was recorded with an optoelectrical motion tracker.

2.2 Paper 2

Nordstrand, M., Svanfeldt, G., Granström, B., House, D. **Measurements of articulatory variation in expressive speech for a set of Swedish vowels.** *Journal of Speech Communication 44 (2004)*, pp.187-196.

Previously, most studies on lip articulation have been based on neutral, read speech. The aim of this study was to investigate how the lip articulation was affected by expression.

The material consisted of three-dimensional data from acted expressive speech, containing utterances presented with six different expressive modes: confident, confirming, questioning, insecure, happy, and neutral.

Rounded and unrounded Swedish vowels were studied using three different measures: lip spreading, lip opening, and lip protrusion. Two subsets of the recorded material were used. The first subset contained four vowels [øː, ɔ, iː, ɛ] with four occurrences each, the second subset consisted of one rounded and one unrounded vowel [ʊ, ɪ] that each occurred 96 times.

The results show that the lip articulation is strongly affected by the expressive mode. In most cases, rounded and unrounded vowels are affected in a similar way. For example, the protrusion is decreased during happy for both groups of vowels. Also the lateral distance, i.e. the spreading of the lips, is increased for both groups during the happy expression. There are, however, some differences between rounded and unrounded vowels during the happy expression. The lip opening is larger for unrounded vowels during the happy expression compared to during the neutral expression, while the rounded vowels instead seem to demand a smaller lip opening during the happy expression. Happy was the expression that affected lip articulation the most.

This study shows the sometimes unexpected flexibility of lip articulation and point out the need of appropriate data when developing talking heads. In order to develop a synthetic talking head, able to express emotions and communicative signals, knowledge about the influence of emotions on lip articulation is required.

Author contribution

The author of this thesis performed the data analysis, and was part of planning the study and writing the paper.

Original contribution

Evidence for how lip articulation is affected during expressive speech.

2.3 Paper 3

Beskow, J., Cerrato, L., Cosi, P., Costantini, E., Nordstrand, M., Pianesi, F., Prete, M., Svanfeldt, G.[3] **Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces.** *Proceedings of Tutorial and Research Workshop on Affective Dialogue Systems (ADS'04)*, pp.301-304.

In the research area of talking heads, there have been various techniques for animating the synthetic face. This makes comparisons and sharing of data difficult. The MPEG-4 standard for synthetic faces is an attempt to resolve this problem, in providing a standard for face animation. The specific data that are used are called Facial Animation Parameters (FAPs) .

This paper describes an experiment that was performed in collaboration between two research sites, one in Italy and one in Sweden4. Each site was in possession of a synthetic talking face of MPEG-4 standard, which at least theoretically enables exchanges of FAPs. Each site had also recorded an emotional multimodal database, using actors from the respective country, from which FAPs were extracted. Each site's FAPs were implemented both in the own and in the other site's synthetic face.

Two parallel tests were run, one at each site, where the subjects were to discriminate between the emotions neutral, happy, and angry for both implementations, as well as for the original videos of both actors. Since no audio was used, only visual cues were available. The participants gave their response on an answer sheet, choosing between the three implemented emotional states.

The results showed that both human faces got higher recognition rates than the synthetic faces. The conditions where the synthetic face used the "home set-up of FAPs" also got higher recognition rates than when the FAPs were switched. This means that the implementation is not completely transparent, that a certain adjustment of the FAPs to the actual model is required to get good results. However, no significant differences were found between the responses made by the Swedish and the Italian subjects.

This experiment identified crucial issues in the process of switching data between the sites, as well as investigating cross-cultural differences in perceiving the expressed emotions. The results indicate that differences in recording conditions and differences in the synthetic face models, despite the MPEG-4 standard, are more important than potential cultural differences.

[3] Names in alphabetical order.

[4] For clarity reasons, the Italian sites are referred to one site, although the Italian researchers that participated in the study are from several Italian universities. They, however, worked together on one talking head, so in this respect they can be thought of as "one site".

Author contribution

The author of this thesis took part in the planning and the recording of the emotional multimodal database and performed the experiment at the Swedish site.

Original contribution

The switching of data between sites in order to explore the possibilities of extended sharing of databases and testing of model independence of the MPEG-4 standard.

2.4 Paper 4

Svanfeldt, G. and Olszewski, D. **Perception experiment combining a parametric loudspeaker and synthetic talking head.** *Proceeding of Interspeech 2005*, pp. 1721-1724.

Using targeted audio, it is the possible to direct and limit the sound to a specific beam. But the targeted audio is very sensitive to reflexions, which lessens the directionality of the beam. In order not to disturb the surrounding, a low volume could be used in order to decrease the audibility of the reflected sound. This study assessed the increased intelligibility that could result from supplementing the targeted audio with a synthetic talking head, visible only for the targeted listener.

This paper investigates how the intelligibility of synthetic speech is affected by the listener's position relative a targeted audio beam. The possible augmentation of the intelligibility that the use of a talking head can provide was also measured.

Two different angles between the loudspeaker and the subjects in the experiment were used: 0° and 45°. Each of these angels was tested both with and without an accompanying synthetic face providing lip articulation. The task consisted of discriminating between seven voiceless consonants [k, p, t, m, n, f, s] uttered in a [a_a] context. The consonants can be divided into three groups: plosives, nasals and fricatives. It was expected that confusion would occur mainly within these three groups, and that the synthetic face would prevent this confusion.

The results of the experiment support the hypotheses. The best recognition results were achieved for the condition with the audio in 0° together with a synthetic face. The condition with the audio in 45° together with a synthetic face was not far behind in recognition rates. Then came the 0° audio only condition and worst results were occurred in the 45° audio only condition.

This experiment also shows some of the difficulties with directed audio. Although the beam itself was well formed and the experiment was performed in a room with few reflective surfaces, the sound volume had to be kept very low in order not to hear the reflections. The usability is thus highly dependent on the interior design. A table, a floor without a carpet, windows and walls are all surfaces that will disturb the final result, regardless of the narrowness and directivity of the original beam.

Author contribution

The author of this thesis designed and performed the experiment and the evaluation of the results, did most of the writing, except the targeted audio technical description.

Original contribution

The use of targeted audio together with synthetic speech and a synthetic face.

2.5 Paper 5

Nordenberg, M., Svanfeldt, G. and Wik, P. (alphabetic order). **Artificial gaze. Perception experiment of eye gaze in synthetic face.** *Proceedings of Second Nordic Conference on Multimodal Communication (2005),* pp. 257-272.

In this paper the aim was to investigate whether it was possible to achieve mutual gaze between the synthetic face at our disposal and an observer. The result of the experiment would tell if subjects – for any of the conditions – perceived that the agent looked them in the eyes. We would then learn if the current parameter set-up of the synthetic face that we had access to was sufficient for mutual gaze, and in that case, how to manipulate the parameters to achieve the required effect.

A virtual focalpoint is used as basis for the eye positioning of the agent. The idea is to place the virtual viewpoint where the observer is thought to be [in relation to the virtual camera], so that it appears that the synthetic agent is looking the observer in the eyes. This can be compared with a photograph; if the person on the photo looks straight into the camera when the picture is taken, he or she will appear to look the observer of the photo in the eyes. Thus, a kind of mutual gaze is achieved.

In the experiment, the virtual viewpoint was systematically changed along three dimensions: laterally, vertically, and in the depth dimension. Two different head positions – one front and one slightly from the side – were also tested. 15 subjects

participated in the experiment and were to answer yes or no to the question "Is this man looking you in the eyes?".

The results show that mutual gaze was achieved between the synthetic face and the observer, although not always with the parameter values that were expected. It was more difficult to obtain mutual gaze when the synthetic head was turned slightly to the side, approximately 11 degrees. It was also in general more difficult for the subjects to determine the target of the agents gaze in the depth dimension. This is probably because it is highly unlikely that a dialogue partner would focus on a spot in the air between the dialogue participants, or even on a spot right behind the observer, since it is not visible for someone in front. Glances sideways, however, are more likely to occur, and therefore more easily interpretable as such.

In this experiment, only the synthetic agent was tested. In order to get a real baseline, a similar experiment could be performed using photos of a human, who is told to either look into the camera or a certain distance to the side, behind or in front of the camera. Next step would be to have dynamic pictures (videos) since a lot of information is in the duration and the movements of the gaze. This kind of setup would give more knowledge about how to handle gaze parameters.

Author contribution

The author of this thesis was part of designing the experiment, partly performed the experiment, as well as data collection and evaluation and wrote part of the paper.

Original contribution

The study of whether mutual gaze can be achieved between a synthetic face and a human observer.

3 Conclusions and future work

The papers in this thesis are all related to the goal of making the talking head an efficient and likable partner in communication. Speech and the corresponding facial gestures are so familiar to us that it is in many situations the most convenient way of communicating. However, the complex nature and composition of these cues makes the research a difficult task.

The experiments described in the papers are not exhaustive in their intent, they are rather first attempts at identifying central issues within each area. They gave rise to more questions than they did answer, but these questions are still important for further research. For more quantitative results it is necessary to perform studies with more recorded individuals, with a large amount of speech material and more speaking situations for each of these.

Not only are the results tentative, the methods are still to be further developed. The whole process has been characterised by a difficult balancing between controlled recording situations and spontaneous speech. Almost certainly both should be performed, they will give different results but none of them can probably separately give all the answers that we look for. Controlled recordings gives the researcher some solid ground to start from, but that also means a risk of biasing the interpretation of the spontaneous recordings.

For example, in the eye gaze experiment we tried to establish a set-up where the agent met the eyes of the observer. There are numerous further experiments that could be done to explore the nature of eye gaze in synthetic agents. One is to compare with photographs of humans, where different focus points are used. Another is to try out different turn taking gestures using gaze, and possibly in combination with head movements or facial gestures.

As to the articulation, there are numerous factors to study. Only a few vowels were studied in paper 2 and only with one speaker.

In this work I have looked at several seemingly small fragment of the whole picture – the corner of the mouth and the eye gaze. Hopefully, the results from these studies, or perhaps the questions that emerged, will be yet one step in the process of understanding the composition and interaction of these cues – these cues that constitute the human ability of communicating.

4 References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.E. and Öhman, T. (1998). Synthetic faces as a lipreading support. *Proceedings of ICSLP'98*. Sydney, Australia. Vol. 7, p. 3047.

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press, Great Britain.

Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. (1995): The Waxholm system - a progress report. *Proceedings of Workshop on Spoken Dialogue Systems*, Vigsø, Denmark.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*. Madrid, Spain, pp. 299-302, 1995.

Beskow, J. (1997). Animation of Talking Agents. *Proceedings of International Conference on Auditory-Visual Speech processing (AVSP'97)*, Rhodos, Greece, pp. 149-152, 1997.

Beskow, J. and Cerrato, L. (2004). Evaluation of the expressivity of a Swedish talking head in the context of human-machine interaction. *Proceedings of GSCP2004*, Italy.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E and Öhman, T. (1997). The Teleface project - Multimodal Speech Communication for the Hearing Impaired. *Proceedings of Eurospeech '97*.

Beskow, J., Granström, B., House, D. (2006). Focal accent and facial movements in expressive speech. *Proceedings from Fonetik 2006*, Working Papers 52, General Linguistics and Phonetics, Lund University, pp. 9-12.

Beskow, J. and Nordenberg, M.(2005). Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head. *Proceedings of Interspeech 2005* Lisbon, Portugal, pp. 793-796.

Campbell, Nick. The Recording of Emotional Speech (JST/CREST database research). *Proceedings of COCOSDA 2001*, Aalborg, Denmark.

Carlson, R., Granström, B., and Hunnicutt, S. (1982). A multi-language text-to-speech module, *Proceedings of ICASSP '82*, Paris, Vol. 3, pp 1604-1607.

Carlson, R., Granström, B. and Nord, L. (1992). Experiments with emotive speech - Acted utterances and synthesized replicas. *Proceedings of ICSLP*, Banff, Canada, pp. 671-674.

Cohen, M. M., Massaro, D. W., and Clark, R. (2002). Training a talking head. *Proceedings of fourth international conference on multimodal interfaces (ICMI)*, Pittsburgh, Pennsylvania, USA, pp.499-510.

Cosi. P. and Magno Caldognetto, E. (1996). Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications. In *Speechreading by Humans and Machine: Models, Systems and Applications*, D.G. Storke and M.E. Henneke eds., NATO ASI Series, Series F: Computer and Systems Sciences, Vol. 150, Springer-Verlag, 1996, pp. 291-313.

Edlund, J. and Nordstrand, M. (2002). Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. *Proceedings of ISCA Workshop Multi-Modal Dialogue in Mobile Environments*.

Ekman and Friesen (1975). *Unmasking the face: A guide to recognising emotion from facial clues*. Prentice Hall.

Engwall, O., Wik, P., Beskow, J. and Granström, B. (2004). Design strategies for a virtual language tutor. *Proceedings of ICSLP 2004*. Vol 3, pp.1693-1696.

Fonàgy, I. (1976). La mimique buccale. *Phonetica 33*, pp. 31-44.

Garau, M., Slater, M., Bee, S. and Sasse, M., A. (2001). The Impact of Eye Gaze on Communication using Humanoid Avatars. *Proceedings of SIGCHI'01*, Seattle, USA, pp. 309-316.

Gustafson, J, Bell, L, Beskow, J, Boye, J, Carlson, R, Edlund, J, Granström, B, House, D & Wirén M (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. *Proceedings of ICSLP 2000*, Beijing, Vol 2, pp.134-137.

Heylen, D., van Es, I., Nijholt, A. and van Dijk, B. (2005). Controlling the gaze of conversational agents. J.C.J. van Kuppevelt et al. (eds), *Advances in Natural Multimodal Dialogue Systems*, pp.245-262. Springer, Netherlands.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *In Proc of Eurospeech 2001*. Aalborg, Denmark. pp. 387-390.

Magno Caldognetto E., Cosi P., Cavicchio F. (2004). Modification of the Speech Articulatory Characteristics in the Emotive Speech. *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems", Kloster Irsee, Germany* .pp. 233-239.

Massaro, D. W. and Bosseler, A. (2003). Perceiving speech by ear and eye: Multimodal integration by children with autism. *The journal of Developmental and Learning Disorders, 7*:111-146.

Mehrabian, A. and Russell. J. A. (1974). *An approach to environmental psychology.* Cambridge, MA: M.I.T. Press.

Pandzic I. S. and Forchheimer R. (2002). *MPEG-4 Facial Animation – the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons.

Park Lee, S., Badler, J. B. and Badler, N. I. (2002). Eyes Alive. *ACM Transactions Graphics 21 (3) July 2002*, ACM: NY, pp.637-644.

Pelachaud, C., Badler, N. and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science 20*, pp.1-46.

Revéret, L. and Benoît, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. *Proceedings of AVSP'98*, Sydney, Australia, pp. 207-212.

Vatikiotis-Bateson, E., Munhall, K.G., Kasahara, Y., Garcia, F., and Yehia, H. (1996). Characterizing audiovisual information during speech. *Proceedings of International Conference on Spoken language Processing (ICSLP-96)*. Philadelphia PA. Vol.3, pp.1485-1488.

Westervelt, P. J. (1963). Parametric Acoustic Array. *Journal of the Acoustical Society of America,* Vol. 35, p. 535-537.