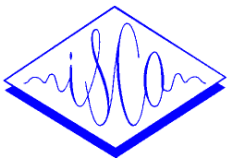

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Measurements of Articulatory Variation and Communicative Signals in Expressive Speech

Magnus Nordstrand, Gunilla Svanfeldt, Björn Granström, and David House

Centre for Speech Technology, Department of Speech, Music and Hearing
KTH, Stockholm, Sweden
{magnusn, gunillas, bjorn, davidh}@speech.kth.se

Abstract

This paper describes a method for acquiring data for facial movement analysis and implementation in an animated talking head. We will also show preliminary data on how a number of articulatory and facial parameters for some Swedish vowels vary under the influence of expressiveness in speech and gestures. Primarily we have been concerned in expressive gestures and emotions conveying information that is intended to make the animated agent more ”human-like” as described in the objectives of the PF-Star[1] project.

1. Introduction

Animated talking heads are becoming increasingly popular as a human-computer interface. Improvements in graphical performance, articulatory synthesis [1, 2, 3] as well as visual naturalness increase the areas in which such technologies might be useful, e.g. virtual language tutors, computer game characters or as communication aids for hearing impaired people. As interactive possibilities increase so might expectations of the capabilities of these agents. Having a computer game character not being able to signal intentions and emotions, both in face and voice, might be a rather dull experience.

When people speak they often use both visual and acoustic signals in order to supply information. Normal speech communication is thus multimodal. The visual modality can qualify the auditory information providing segmental cues on place of articulation [4] and prosodic information concerning prominence and phrasing. But we also use facial signals to influence other participant’s behaviour by expressing affectual signals [5] and extralinguistic information such as signals for turn taking [6, 7], emotions and attitude [8]. There are a number of studies showing how expressiveness and emotions affect our facial display, e.g. how we raise our eyebrows, move our eyes or blink, or nod and turn our head [8], and how we use these visual cues to signal emphasis, etc. But less is known about how articulation is affected by expressiveness in speech. By expressive speech we refer to speech performed when affected by an emotion, e.g. smiling while we speak.

When looking at articulatory synthesis based on phonetic input, it is exclusively based on non-expressive speech, i.e. the material is read with a neutral voice and facial expression. However, expressiveness might affect articulation and how we produce speech a great deal and an articulatory parameter might behave differently under the influence of different emotions. For example Fonàgy [9] show how intraoral speech mechanisms, e.g. the tongue was affected by the expression of emotions. Better knowledge about this behaviour will also help us adjust the articulatory rules controlling the articulation of an animated talking head.

There have been attempts to take into consideration how articulation may change depending on speaker or style and make use of that knowledge in audiovisual synthesis. Pelachaud et al. [10] proposed a method where they could define various speaker characteristics such as speech-rate and timing issues for e.g. intonation and punctuation and make adjustments to coarticulation. In their model faster speech-rate led to a decrease in intensity of lip shape and in the case of deformable segments their associated lip shapes lost their characteristic shape. They also described how this model could generate emotions and expressions, but the articulation was not affected by these rules.

The goal of this study is to gain more insight onto how the articulation is affected by the expressed emotion. We have studied two groups of Swedish vowels (rounded and unrounded). Furthermore, we have analysed how communicative signals, like eyebrow movements were affected by expressive speech.

2. Background

We present here the background and the aims of the European project PF-Star in which this study is a part of, and thereafter we give a brief description of our animated agent model.

2.1. PF-Star

The aim of the PF-Star project is to establish future activities in the field of multi-sensorial and multi-lingual communication. This will be achieved by providing technological baselines, comparative evaluations, and assessments of prospects of core technologies, which future research and development efforts can build on. The areas addressed are analysis and synthesis of emotions in speech and faces, speech-to-speech translation and speech technologies for children. The experiment and results in this paper only concern the first field, and more specifically, the synthetic faces.

In the first phase of the project, the work mainly consists of material collection and of defining an annotation format. The issues for this paper is the method of collecting data and a presentation of some preliminary results from our first recording.

2.2. The animated agent

Our animated head [1] gives us great freedom when it comes to making it expressive and having it to perform gestures. However, it is a tedious task to manually tailor every expression it is to perform. Moreover, this may not always yield the desired expression. By collecting data on how a real person actually

[1] **P**reparing **f**uture multi**s**ensorial in**t**er**a**ction **r**esearch

performs different gestures we might become better at controlling the model and also find characteristics in various gestures that have not yet been thought of. The animated head used in this study is also capable of lip-synchronised speech with rule-driven articulation. This rule-driven articulation uses specified target values for each segment based on its viseme classification when performing the coarticulation. However, these target values are optimised for non-expressive or neutral speech, so when having it perform an expressive face (e.g. a happy smiling face) while it is speaking and letting these target values control the articulation might result in an inconsistent articulation.

2.3. Methods for dynamic data collection

To obtain dynamical 3D-data of facial movements a number of methods have been developed. Some techniques are based on video image processing. For example, Reveret and Benoît [11] used a technique where they automatically could fit the video image to a model of the face or a part thereof, i.e. the lips. Benoît et al. [12] used a method where they, by applying blue makeup to the subjects lips, automatically could measure different articulatory parameters performed by the subject. However, we wanted to be able to obtain both articulatory data as well as other facial movements at the same time and it was crucial that the accuracy in the measurements was good enough for re-synthesis of the animated head.

Optical motion tracking systems are gaining popularity for being able to handle the tracking automatically and having good accuracy as well as good temporal resolution. The Qualisys system that we used in this experiment has an accuracy below 1 mm with a temporal resolution of 60 Hz and had already successfully been used to acquire articulatory data for our animated head. Similar systems also exists such as OPTOTRAK[2] and ELITE[3]. However, having to affix markers to the face is unpractical if you want to measure i.e. the inner contours of the lips since the cameras need to have optical contact with the beads attached to the face. It would for example be difficult to track the markers when performing a bilabial occlusion since the markers would be hidden to the camera.

3. Data collection and processing

The data acquisition and processing was very similar to earlier facial measurements carried out at CTT by i.e Beskow, Engwall and Granström [13]. The audio and visual data was collected by having an actor read prompted sentences. The recorded subject was a male native speaker of Swedish. Audio data was recorded on DAT-tape and visual data was recorded using the optical motion tracking system Qualisys[4] (see Fig. 1). The sentences to be read and acted were shown on a screen and recorded in one-minute chunks, due to a limitation in the motion tracking system. A synchronisation signal produced by the Qualisys system was recorded on one channel of the DAT-tape enabling audio and visual data to be matched.

By attaching infrared reflecting markers to the subject's face (see Fig. 2), the system is able to register the 3D-coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms. We used 30 markers to register lip movements as well as other facial movements such as eyebrows, cheek, chin and eyelids. Additionally we placed three markers on the chest to register head movements with respect to the torso. A pair of spectacles with four markers attached were used as a reference to be able to factor out head and body movements when looking at the facial movements specifically.

[2] http://www.bts.it
[3] http://www.ndigital.com
[4] http://www.qualisys.se

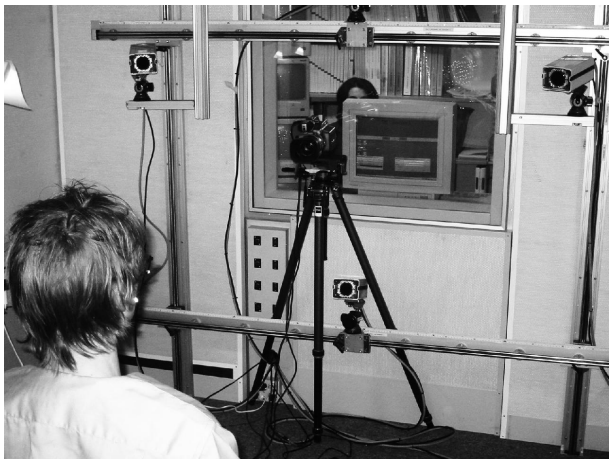

Figure 1: *Data collection setup with video and IR-cameras, microphone and a screen for prompts.*

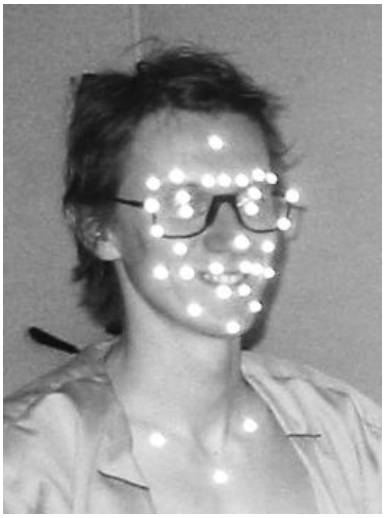

Figure 2: *Test subject with the IR-reflecting markers glued to the face.*

3.1. Material and data processing

We chose to record 15 different emotional expressions. Together with the six universal prototypes for emotions: *anger, fear, surprise, sadness, disgust* and *happiness* [14], we also had the subject to act *worried, satisfied, insecure, confident, questioning, encouraging, doubtful, confirming* and *neutral*. For each emotion, the subject read nine sentences in Swedish. The sentences were kept neutral with respect to content in order not to affect the acted expressions, and consisted mainly of numbers, but also the words ”Linköping” and ”ja”.

The audio signal was used to label the data phonologically by first phonetically transcribing the sentences using the transcription part of a text-to-speech system. The transcriptions were then manually corrected to match what was actually read by the actor. An automatic aligner [15] was used to pair the pho-

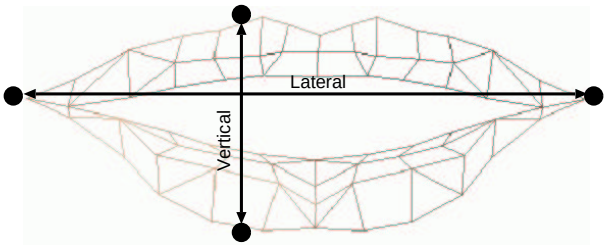

Figure 3: *The positon of the markers used to analyse the vertical and lateral distances for rounded and unrounded vowels.*

netically transcribed speech with the sound signal to retrieve the time for phoneme and word boundaries. This information was then used to match the 3D-data with speech for analysis.

4. Analysis

When analysing the recorded data, we focused on two different areas in the face; the mouth region and the eyebrow-forehead region. The former region can give us clues about how the articulation is affected when emotions are expressed while speaking and the latter is interesting in order to study communicative signals.

4.1. Objectives of the analysis

First of all, a preliminary study was conducted in order to investigate whether articulation was affected by expressive speech and if there might be any difference between the way rounded and unrounded vowels were affected. A more qualitative study was then performed, concentrating only on one vowel of each group.

Previous research [8] has established a number of distinguishable eyebrow actions (AU) depending on the emotion and on the communicative signals that are to be expressed. In this first tentative study, we wanted to investigate the expressiveness and the activity of the eyebrows based on our recording technique.

4.2. Corpus

Since the recorded material for each emotion was limited, we analysed a small sample (only one occurrence of each vowel per emotion) of [øː,ɔ] and [iː,ɛ] for our first study. [iː,øː] were elicited from the word ”Linköping”, [ɔ] from ”484” and [ɛ] from ”961”. Four markers (out of six) around the mouth were studied; they were placed in the corners of the mouth, on the right side of the upper and lower lip. The Euclidean distances between the centres of gravity of these markers gave us the vertical and the lateral distances (see Figure 3).

When narrowing the analysis, five occurrences per emotion of the vowels [iː] (from the word ”nio”) and [ɵ] (from the word ”hundra”) were collected and analysed. The distances were calculated as in the first study.

Finally, for the analysis of the eyebrows, the measurements were performed over nine whole sentences for each emotion by analysing the movements of the three markers affixed to each eyebrow.

Four expressions couldn’t be analysed because the Qualisys system was unable to track some of the markers. This prevented us from getting accurate data to analyse for the expressions *disgust, insecure, confident* and *confirming*.

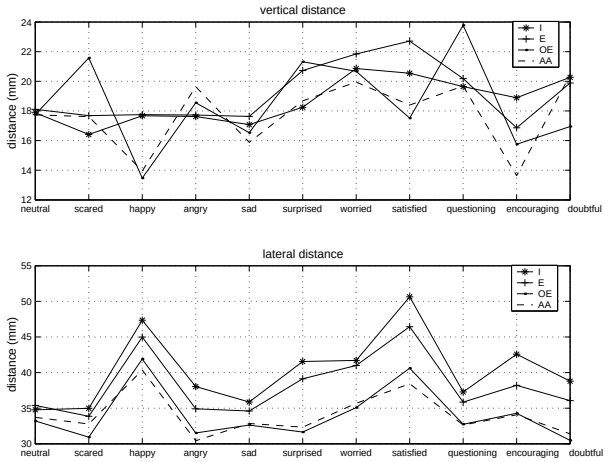

Figure 4: *Vertical and lateral distances (in mm) for rounded and unrounded vowels for different expressions. I=[iː], E=[eː], OE=[øː] and AA=[ɔ]*

5. Results

Here we present results showing how articulation of rounded and unrounded vowels is affected when the speaker is influenced by expressions and emotions. We also present a schematic view of how the eyebrows contribute to the performed expression.

5.1. Articulation – rounded and unrounded vowels

The study of the rounded and unrounded vowels showed that the effect of emotion was significant for both groups, according to a repeated measurement ANOVA test. The effect of vowel was also found to be significant for the lateral distance, but not for the vertical. As shown in Figure 4, there is no appearent grouping of the vowels in the vertical direction, but in the lateral there is a distinct difference between the rounded vowels on one hand, and the unrounded on the other. When analysing the lateral distance in more detail, [iː] was significantly different from the two rounded vowels [øː,ɔ]. For [ɛ], the difference from the other group of vowels was not significant, and, as expected, the rounded vowels were not significantly different from one another. It was also found that the rounded vowels allowed more variation in the vertical direction than the unrounded, whereas the unrounded vowels were more flexible in the lateral direction.

5.2. Articulation – specific phonemes

To examine the differences that we found in the feasibility study in more detail, we concentrated on only two vowels, one of each group ([iː] and [ɵ]). As could be assumed, the differences between the different emotions became more distinct when concentrating on one single phoneme, although the tendencies are the same as in the previous study.

For the lateral distance for the phoneme [iː], the effect of the emotion was significant ($F(10, 44)=23.6$; $p<0.05$) in a repeated measurement ANOVA test. A Bonferroni post-hoc test showed that, *happy* and *satisfied* had a significantly larger lateral distance than all the other emotions. Lowest lateral distance for [iː] occurred for *neutral*, *sad* and *scared*. For the vertical direction the effect of emotion was significant ($F(10,44)=36.1$;

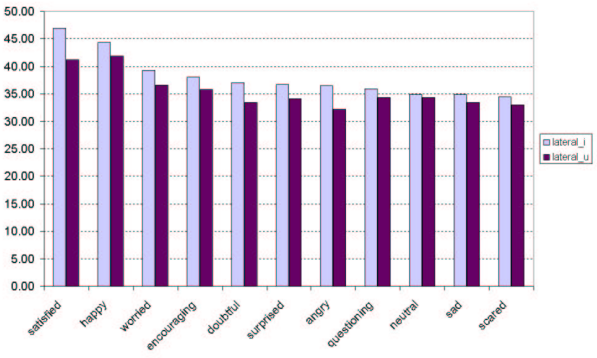

Figure 5: Lateral distances (in mm) for the rounded vowel [iː] and the unrounded vowel [ɵ] for different expressions.

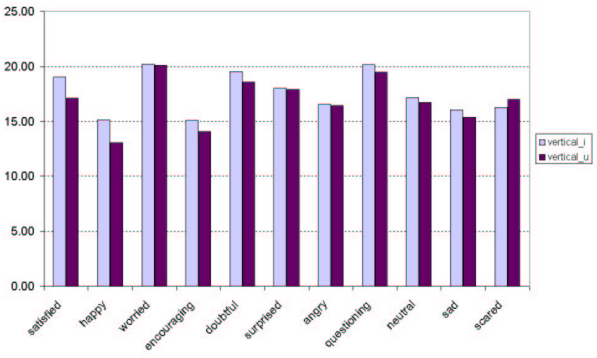

Figure 6: Vertical distances (in mm) for the rounded vowel [iː] and the unrounded vowel [ɵ] for different expressions.

$p<0.05$) for the unrounded vowel [iː]. In this direction, *worried* and *questioning* caused the largest distance, and *happy* and *encouraging* the smallest.

When analysing the rounded vowels in the previous, more general study no significant differences between the emotions could be established, but when focusing only on the phoneme [ɵ] the results were clearer. The effect of emotion on the lateral and vertical distance, respectively, was significant ($F(10,44)=29.4$; $p<0.05$) and ($F(10,44)=22.5$; $p<0.05$) according to the repeated measurement ANOVA test. The general pattern was similar to the one for [iː], except that *angry* produced the smallest lateral distance and *happy* produced such a small vertical distance that it was significantly different from all other emotions in the analysis.

In general, the two vowels followed the same pattern, which may be interpreted as that emotion is the dominating factor. But, the rounded vowel, [ɵ], consistently had a smaller lateral distance, i.e. the lips were more rounded than for the unrounded vowel [iː]. In figure 5, the emotions are sorted in descending order by the lateral distance. In figure 6, the order is the same as in the previous figure, and when comparing the two images, it is clear that no obvious relation between the impact of emotion on the mouth opening and lip rounding could be stated.

5.3. Communicative signals

The eyebrows have been found to be very informative when expressing emotions [16] in a synthetic face, and are thus very important for the realisation of emotions in an animated head. We were able to detect differences between various emotions but we also found that the position of the eyebrows was rather onstant during the whole sentence. In figure 7 we have plotted the mean value for each parameter over the whole set of sentences for ten expressions. This may be due to the fact that there were no interaction situations during the recording so turn-taking signals were not present. Also, since the sentences were rather neutral to their content, emphasis gestures can not be said to have a strong influence on the behaviour of the eyebrows.

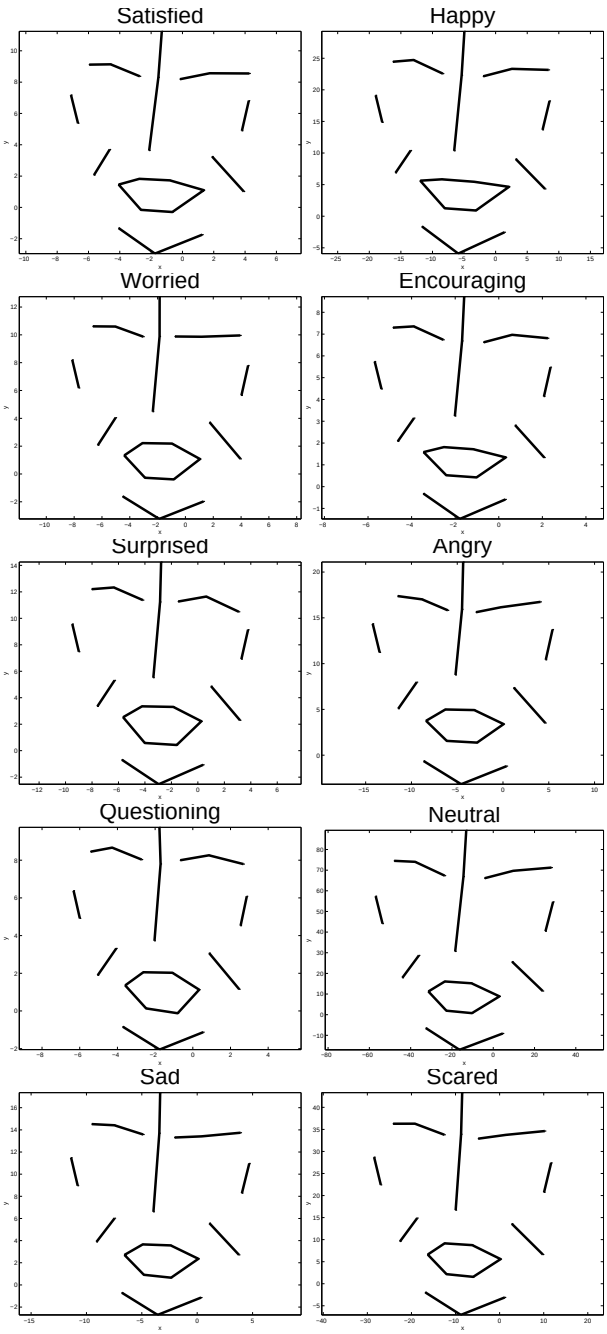

Figure 7: Visualisation of facial parameter mean values.

6. Conclusions

The study of articulation showed that both rounded and unrounded vowels are affected by the emotions expressed. In general, the two groups of vowels follow the same pattern, but some differences could be observed. The unrounded vowels allow more variation in the lateral direction whereas the rounded are more flexible than the other group of vowels in the vertical direction. Another thing worth noticing is that there is no obvious relation between how the two distances in question are affected by the same emotion.

This calls for caution when making assumptions on general tendencies for facial expressions for different emotions. However, we have to keep in mind that in acted speech the emotions are not necessarily expressed in the same way as in spontaneous speech. Also, the articulation is likely to differ between speakers, so these conclusions only apply to the subject in this study. In order to learn more about how emotions change the articulation of phonemes, more extensive recordings followed by a detailed study is required.

According to the results from the eyebrow analysis, we consider our recording technique to be well suited for the capturing of such facial movements.

7. Discussion and future studies

The next step could be to apply these articulatory findings on the rules for articulation of Swedish speech in the facial synthesis to achieve a more correct articulation for the tested emotions and expressions. However, we have to remember that the findings in this study do not cover all rounded and unrounded vowels. Furthermore, in this study the phonemes were always in the same context. It is likely that the result would be different if other coarticulation aspects were introduced. Also we do not claim that the use of these rules will be sufficient to make the face look for example happy or sad. To achieve this we would probably need to make other adjustments to the animated face's expression, but it might be a way of making the articulation more consistent with the performed expression when other deformations are applied to the face in order to get a specific emotion or expression.

8. Acknowledgments

Special thanks to Bertil Lyberg for making available the Qualisys Lab at Linköping University. The PF-Star project[5] is co-ordinated by Instituto Trentino di Cultura Instituto per la Ricerca Scientifica e Technologica. The project is funded by the European Commission, proposal number: IST-2001_37599. This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

9. References

[1] J. Beskow, "Animation of Talking Agents," in *Proceedings of AVSP'97*, Rhodes, Greece, 1997, pp. 149–152.

[2] M. M. Cohen and D. W. Massaro, "Modelling Coarticulation in Synthetic Visual Speech," in *Models and Techniques in Computer Animation*, N. Magnenat-Thalmann and D. Thalmann, Eds., pp. 139–156. Springer Verlag, Tokyo, 1993.

[3] L. Reveret, G. Bailly, and P. Badin, "Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *Proceedings of the sixth International conference on Spoken Language Processing (ICSLP'2000)*, Beijing, China, 2000, pp. 755–758.

[4] G. Collier, *Emotional Expression*, Lawrence Erlbaum Associates, Hillsdale, N.J., 1985.

[5] M. Argyle and M. Cook, *Gaze and mutual gaze*, Cambridge University Press, London, 1976.

[6] S. Jr. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of personality and social psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[7] S. Jr. Duncan, "On the structure of speaker-auditor interaction during speaking turns," *Language in society*, vol. 3, no. 2, pp. 161–180, 1974.

[8] P. Ekman, "About brows: emotional and conversational signals," in *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, Eds., pp. 169–248. Cambridge University Press, Cambridge, England, 1979, New York.

[9] I. Fonàgy, "La mimique buccale," *Phonetica*, vol. 33, pp. 31–44, 1976.

[10] C. Pelachaud, N. Badler, and M. Steedman, "Generating Facial Expressions for Speech," *Cognitive Science*, vol. 20, pp. 1–46, 1996.

[11] L. Reveret and C. Benoît, "A new 3d lip model for analysis and synthesis of lip motion in speech production.," in *Proceedings of the International Conference on Audio-visual Speech Processing (AVSP'98)*, Terrigal, Australia, 1998, pp. 207–212.

[12] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which components of the face do humans and machines best speechread?," in *Speechreading by Humans and Machines: Models, Systems and Applications.*, D. G. Stork and M. E. Hennecke, Eds., pp. 315–330. Springer, Berlin, 1996.

[13] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements," in *Proceedings of the fifteenth international congress of phonetic sciences*, Barcelona, Spain, 2003, (to appear).

[14] P. Ekman, "Emotion in the human face," Cambridge University Press, New York, 1982.

[15] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proceeding of Fonetik 2003*, Umeå, Sweden, 2003.

[16] D. W. Massaro, "Multimodal emotion perception analogous to speech processes," in *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, sep 2000, pp. 114–121.

[5] http://pfstar.itc.it/