Visual correlates to prominence in several expressive modes

Jonas Beskow, Björn Granström and David House

Department of Speech, Music and Hearing, Centre for Speech Technology (CTT)
KTH, Stockholm, Sweden
{beskow|bjorn|davidh}@speech.kth.se

Abstract

In this paper, we present measurements of visual, facial parameters obtained from a speech corpus consisting of short, read utterances in which focal accent was systematically varied. The utterances were recorded in a variety of expressive modes including certain, confirming, questioning, uncertain, happy, angry and neutral. Results showed that in all expressive modes, words with focal accent are accompanied by a greater variation of the facial parameters than are words in non-focal positions. Moreover, interesting differences between the expressions in terms of different parameters were found.
Index terms: visual prosody, focal accent, expressive speech

1. Introduction

As we interact with others, we routinely make use of several of our sensory modalities in the process of communicating and exchanging information. A full account of the speech communication process must therefore include multiple modalities. The visible articulatory movements are mainly those of the lips, jaw and tongue. However, these are not the only visual information carriers in the face during speech. Much prosodic information related to prominence and phrasing, as well as communicative information such as signals for feedback, turn-taking, emotions and attitudes can be conveyed by, for example, nodding of the head, raising and shaping of the eyebrows, eye movements and blinks. We have been attempting to model such gestures in a visual speech synthesis system, not only because they may transmit important non-verbal information, but also because they make the face look alive. However, these movements are more difficult to model in a general way than the articulatory movements, since they are optional and highly dependent on the speaker's personality, mood, purpose of the interaction, etc.

In earlier work, we have concentrated on introducing eyebrow movement (raising and lowering) and head movement (nodding) to an animated talking agent. Lip configuration and eye aperture are two additional parameters that we have experimented with. Much of this work has been done by hand-manipulation of parametric synthesis and evaluated using perception test paradigms. We have explored three functions of prosody, namely prominence, feedback and interrogative mode.

In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish [1], a test sentence was created using our audio-visual text-to-speech synthesis in which the acoustic cues and lower face visual cues were the same for all stimuli. Articulatory movements were created by using the text-to-speech rule system. The upper face cues were eyebrow movement where the eyebrows were raised on successive words in the sentence. The words with concomitant eyebrow movement were generally perceived as more prominent than words without the movement. This tendency was even greater for a subgroup of non-native (L2) listeners. Similar results have also been obtained for Dutch by Krahmer et al. [2] [3].

In another study [4] both eyebrow and head movements were tested as potential cues to prominence. Results from this experiment indicated that combined head and eyebrow movements are quite effective cues to prominence when synchronized with the stressed vowel of the potentially prominent word and when no conflicting acoustic cue is present.

The feedback function of prosody was tested in [5] in the context of a travel agent scenario. By varying acoustic and visual features of a talking-head agent the contributions of six different parameters to conveying positive or negative feedback were studied. The features tested were smile, head nod, eyebrow movements, eye closure, intonation contour and delayed response. To convey positive, confirming feedback, the smile was the most important factor followed by declarative intonation. Eyebrow rising and head nodding also contributed significantly to convey positive feedback while eye closure and delay did not show significant results. Features significantly contributing to negative feedback were a neutral mouth configuration, interrogative intonation, a slow upwards movement of the head and eyebrow frowning.

In another experiment, similar visual cues were used to see if they could influence the perception of question and statement intonation in Swedish [6]. Results showed only a marginal influence of the visual cues. While the visual cues for declarative mode reinforced declarative interpretations, the cues for interrogative mode led to more ambiguity in the responses. Similar results have been obtained for English by Srinivasan and Massaro [7].

This type of experimentation and evaluation has established the perceptual importance of eyebrow and head movement cues for prominence and feedback. These experiments do not, however, provide us with quantifiable data on the exact timing or amplitude of such movements used by human speakers such as can be found in e.g. [8][9]. Nor do they give us information on the variability of the movements in communicative situations. This kind of information is important if we are to be able to implement realistic facial gestures and head movements in our animated agents. In this paper we will report on methods for the acquisition of visual and acoustic data, and present measurement results obtained from a speech corpus in which focal accent was systematically varied in a variety of expressive modes.

2. Data collection and corpus

For the analysis of acoustic prosodic measurements there exist well-established (semi) automatic techniques operating on the audio signal. Analysis of visual prosodic content is a less

explored problem. To automatically extract important facial movements we have employed a motion capture procedure.

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 resynthesis of an animated head. Optical motion tracking systems are gaining popularity for being able to handle the tracking automatically and for having good accuracy as well as good temporal resolution. The opto-electronic motion tracking system, the Qualysis MacReflex system, that we use has an accuracy better than 1 mm with a temporal resolution of 60 Hz. The data acquisition and processing is very similar to earlier facial measurements carried out at CTT by e.g. [10]. The recording set-up can be seen in Fig. 1.

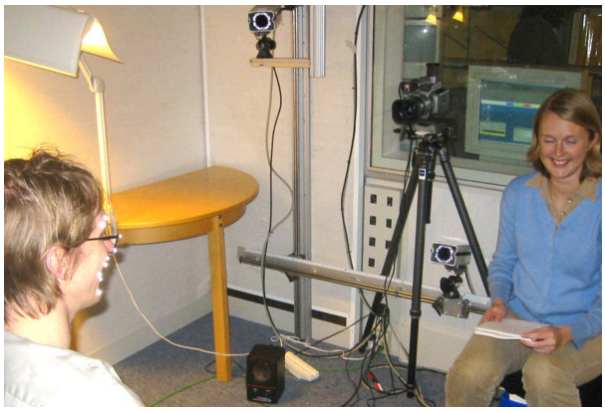

Figure 1: *Data collection setup with video and IR-cameras, microphone and a screen for prompts.*

The subject could either pronounce sentences presented on the screen outside the window or be engaged in a (structured) dialogue with another person as shown in the figure. By attaching infrared (IR) 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 a number of markers to register lip movements as well as other facial movements such as eyebrows, cheek and chin.

The data corpora described here was collected in the context of the EU project PF-Star [11]. The analysis and visual synthesis of emotional expressions was one of the main research areas in the project. The multimodal corpora collected within the project was intended to provide materials for the analysis and modeling of expressive human behavior which could be implemented in animated agents. The data corpora thus reflect these goals. Several different types of corpora were collected. These have been reported on in more detail in Beskow et al. [12].

The speech material used for the present study consisted of 39 short, content neutral sentences such as *"Båten seglade förbi"* (The boat sailed by) and *"Grannen knackade på dörren"* (The neighbor knocked on the door), all with three content words which could each be focally accented. To elicit visual prosody in terms of prominence, these short sentences were recorded with varying focal accent position, usually on the subject, the verb and the object respectively, thus making a total of 117 sentences. The utterances were recorded in a variety of expressive modes including certain, confirming, questioning, uncertain, happy, angry and neutral. With the exception of angry, these were all expressions which are considered to be relevant in a spoken dialogue system scenario. A previous study using this material has been reported in [13].

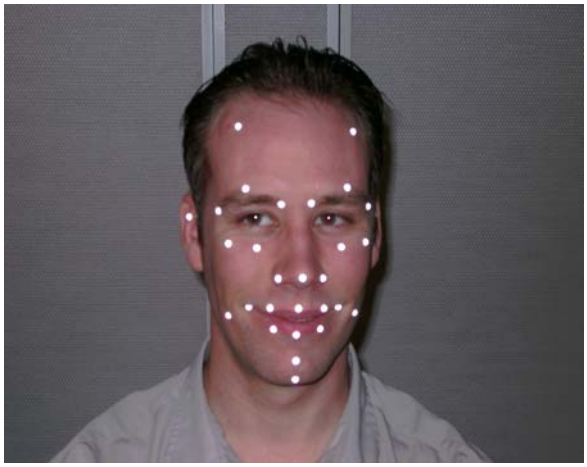

Figure 2: *Test subject with the IR-reflecting markers*

3. Measurement procedure

In the present database a total of 29 IR-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers (on the ears and on the forehead). The marker setup (as shown in figure 2) largely corresponds to MPEG-4 feature point (FP) configuration. Audio data was recorded on DAT-tape, and video was recorded using a mini-DV digital video camera. A synchronisation signal from the Qualisys system was fed into one audio channel of the DAT and DV to facilitate post-synchronisation of the data streams.

3.1. MPEG-4 Facial Animation Parameters

In the present study, we chose to base our quantitative analysis of facial movement on the MPEG-4 Facial Animation Parameter (FAP) representation, because it is a compact and standardised scheme for describing movements of the human face and head. Specifically, we chose a subset of 31 FAPs out of the 68 FAPs defined in the MPEG-4 standard, including only the ones that we were able to calculate directly from our measured point data (discarding e.g. parameters for inner lip contour, tongue, ears and eyes)

Thus, the first step in the analysis was to convert the data into MPEG-4 FAPs. First the marker movements were decomposed into global and local movement. This was done by expressing all points in a local coordinate system that is fixed to the head, which is consistent with MPEG-4 definition of axes: x-axis pointing left, y-axis pointing upward and z-axis pointing straight forward, in the direction of the nose. This coordinate system was defined using the reference markers on the ears and upper forehead. Global head rotation angles were also calculated, and used to calculate FAP 48-50 – head pitch, yaw and roll.

Given that marker placement in the recording session corresponded to MPEG-4 feature points, calculation of the FAP values is achieved using linear relations and

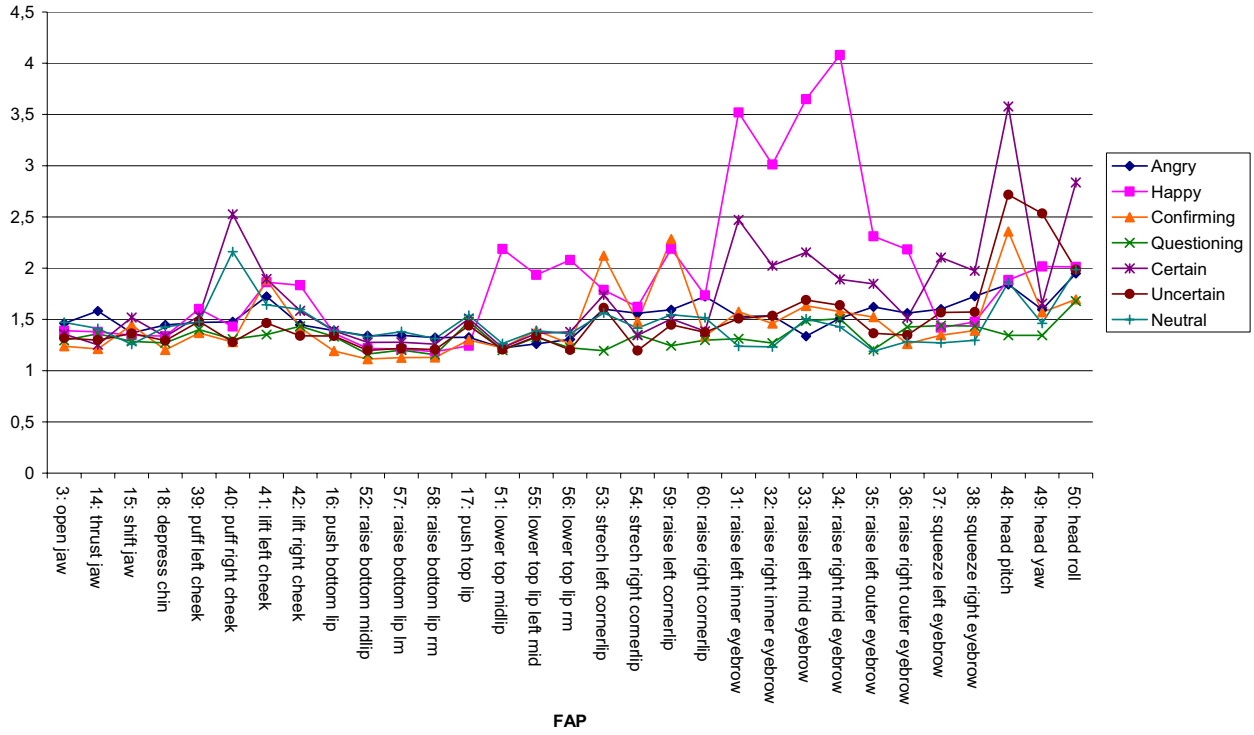

Figure 3: *The focal motion quotient, FMQ, averaged across all sentences, for all measured MPEG-4 FAPs for several expressive modes (see text for definitions and details).*

normalisation factors. If marker coordinates are arranged in a matrix X where each row represents a time frame and each column a coordinate for one of the markers, then a corresponding FAP matrix F can be calculated as

$$F = F_0 + M \cdot U \cdot (X - X_0) \quad (1)$$

Where X_0 represents the markers in resting position, for which corresponding FAP values F_0 have been manually estimated. M is a (manually constructed) matrix that maps marker coordinates (columns in X) to FAPs (columns in F), and U is a diagonal matrix containing the proper FAPU scaling factors for each FAP.

3.2. Focal Motion Quotient - FMQ

We wanted to obtain a measure of how (in what FAPs) focus was realised by the recorded speaker for the different expressive modes. In an attempt to quantify this, we introduce the Focal Motion Quotient, FMQ, defined as the standard deviation of a FAP parameter taken over a word in focal position, divided by the average standard deviation of the same FAP in the same word in non-focal position. This quotient was then averaged over all sentence-triplets for each expressive mode separately.

4. Results and Discussion

As a first step in the analysis the FMQs for all the 31 measured FAPs were averaged across the 39 sentences. These data are displayed in Fig. 3 for the analyzed expressive modes, i.e. Angry, Happy, Confirming, Questioning, Certain, Uncertain and Neutral. As can be seen, the FMQ mean is always above one, irrespective of which facial movement, FAP, that is studied. This means that a shift from a non-focal to a focal pronunciation on the average results in greater dynamics in all facial movements for all expressive modes. It should be noted that these are results from only one speaker and averages across the whole database. It is however conceivable that facial movements will at least reinforce the perception of focal accent. The mean FMQ taken over all expressive modes is 1.6. The expressive mode yielding the largest mean FMQ is happy (1.9) followed by confirming (1.7), while questioning has the lowest mean FMQ value of 1.3. If we look at the individual parameters and the different expressive modes, some FMQs are significantly greater, especially for the Happy expression, up to 4 for parameter 34 ”raise right mid eyebrow”.

In order to more clearly see how different kinds of parameters affect the movement pattern, a grouping of the FAPs is made. In Fig 4. The “Articulation” parameters are the ones primarily involved in the realization of speech sounds (the first 20 in Fig 3.) The “Smile” parameters are the 4 FAPs relating to the mouth corners. “Brows” correspond to the eight eyebrow parameters and “Head” are the three head movement parameters. The extent and type of greater facial movement related to focal accent clearly varies with the expressive mode. Especially for Happy, Certain and Uncertain, FMQs above 2 can be observed. The Smile group is clearly exploited in the Happy mode, but also in Confirming, which supports the finding in [5], where Smile was the most prominent cue for confirming, positive feedback, referred to in the introduction. These results are also consistent with [14] which showed that lip corner displacement was more strongly influenced by utterance emotion than by individual vowel features.

The Brow group was also heavily used for the Happy and Certain mode. Otherwise the Certain and Uncertain modes

look surprisingly similar in the present representation. To obtain a more detailed view of the differences between these two expressions, selected individual parameters for both are displayed in Fig. 5, i.e. four eyebrow parameters and the head pitch and head yaw parameters, corresponding to vertical, nod movement and horizontal head turn/shake movement.

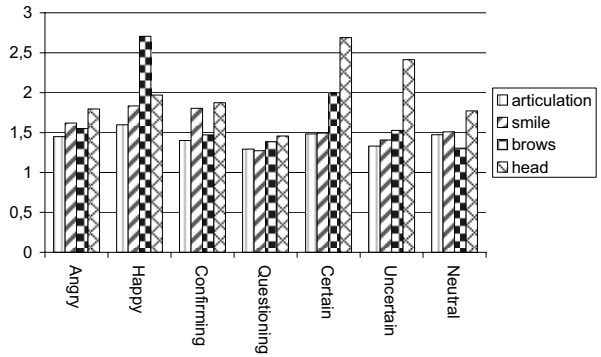

Figure 4: *The effect of focus on the variation of several groups of MPEG-4 FAP parameters, for different expressive modes.*

Now it can be seen that these eyebrow parameters are exploited more for the Certain mode than for Uncertain. The head movement pattern is also clearly different for the two modes. The Head pitch (nod) is used much more for Certain while the Head yaw (shake) is more prevalent in Uncertain.

While much more detailed data on facial movement patterns is available in the database, we wanted to show the strong effects of focal accent on basically all facial movement patterns. Moreover, the results suggest that the specific gestures used for realization of focal accent are related to the intended expressive mode.

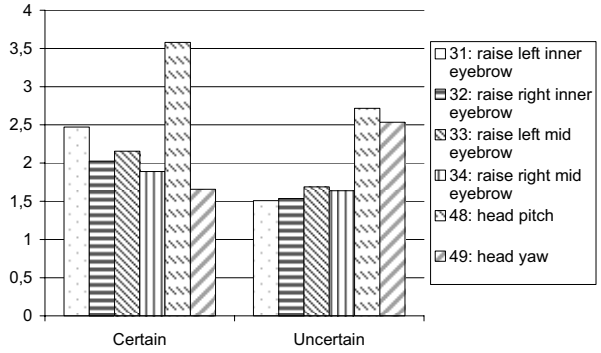

Figure 5: The effect of focal accent on selected parameter variations in Certain and Uncertain readings

5. Acknowledgements

Much of the work presented in this paper has been done by other members of the CTT multimodal communication group including Loredana Cerrato, Mikael Nordenberg, Magnus Nordstrand and Gunilla Svanfeldt which is gratefully acknowledged. Special thanks to Bertil Lyberg for making available the Qualisys Lab at Linköping University. The work has been supported by the EU/IST projects SYNFACE, PF-Star and CHIL, and CTT, the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA, KTH and participating Swedish companies and organizations.

6. References

[1] Granström, B., House, D. and Lundeberg, M., "Prosodic Cues in Multimodal Speech Perception", In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99*). San Francisco, 655-658. 1999.

[2] Krahmer, E., Ruttkay, Z., Swerts, M. and Wesselink, W., "Pitch, eyebrows and the perception of focus", In: *Proceedings of the Speech Prosody 2002 Conference*, B. Bel & I. Marlien (eds.). Aix-en-Provence: Laboratoire Parole et Langage, 443-446. 2002.

[3] Krahmer, E., Ruttkay, Z., Swerts, M. and Wesselink, W. "Perceptual evaluation of audiovisual cues for prominence", In: *Proceedings 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colo., USA. 1933-1936. 2002

[4] House, D., Beskow, J. and Granström, B., "Timing and interaction of visual cues for prominence in audiovisual speech perception", In *Proc of Eurospeech 2001*, 387-390. 2001.

[5] Granström, B., House, D. and Swerts, M., "Multimodal feedback cues in human-machine interactions", In *Proc of the Speech Prosody 2002 Conference,* B. Bel & I. Marlien (eds.). Aix-en-Provence: Laboratoire Parole et Langage, 347-350. 2002.

[6] House, D., "Intonational and visual cues in the perception of interrogative mode in Swedish", In *Proceedings of ICSLP 2002*. Denver, Colorado, 1957-1960. 2002.

[7] Srinivasan, R.J. and Massaro, D.W., "Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English", *Language and Speech* 46(1), 1-22. 2003.

[8] Keating, P., Baroni, M., Mattys, S., Scarborough R., Alwan A., Auer E. and Bernstein L., "Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English", In *Proc. 15th International Congress of Phonetic Sciences*, 2071-2074. 2003.

[9] Dohen, M., *Deixis prosodique multisensorielle: Production et perception audiovisuelle de la focalisation contrastive en Français.* PhD thesis, Institut de la Communication Parlée, Grenoble. 2005.

[10] Beskow, J., Engwall, O. and Granström, B., "Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements", *Proc. of ICPhS 2003*. Barcelona, Spain, 431-434. 2003.

[11] PF-STAR: http://pfstar.itc.it/ (April 2006)

[12] Beskow, J., Cerrato, L., Granström, B., House, D., Nordstrand, M. and Svanfeldt, G., "The Swedish PF-Star Multimodal Corpora", In *Proc LREC Workshop, Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*. Lisbon, Portugal, 34-37. 2004.

[13] Cerrato, L. and Svanfeldt, G., "A method for the detection of communicative head nods in expressive speech", In *Proceedings of The Second Nordic Conference on Multimodal Communication,* Gothenburg University. In press.

[14] Nordstrand, M., Svanfeldt, G., Granström, B. and House, D., "Measurement of articulatory variation in expressive speech for a set of Swedish vowels", *Journal of Speech Communication 44,* 187-196. 2004.