

Measuring and modeling audiovisual prosody for animated agents

Björn Granström and David House

Department of Speech, Music and Hearing, Centre for Speech Technology (CTT)

KTH, Stockholm, Sweden

{bjorn|davidh}@speech.kth.se

Abstract

Understanding the interactions between visual expressions, dialogue functions and the acoustics of the corresponding speech presents a substantial challenge. The context of much of our work in this area is to create an animated talking agent capable of displaying realistic communicative behavior and suitable for use in conversational spoken language systems, e.g. a virtual language teacher. In this presentation we will give some examples of recent work, primarily at KTH, involving the collection and analysis of a database for audiovisual prosody. We will report on methods for the acquisition and modeling of visual and acoustic data, and provide some examples of analysis of head nods and eyebrow settings.

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 mainly 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 also 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 audiovisual 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 Kraemer et al. [2] [3].

In another study [4] both eyebrow and head movements were tested as potential cues to prominence. The goal of the study was two-fold. First of all we wanted to see if head movement (nodding) is a more powerful cue to prominence than is eyebrow movement by virtue of a larger movement. Secondly, we wanted to test the perceptual sensitivity to the timing of both eyebrow and head movement in relationship to the syllable. 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. Sensitivity to the timing of these movements was on the order of 100 ms., although there was a tendency for integration of the movements to the nearest potentially prominent word.

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 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 distinguishing questions from statements, prosody has a well-established role, especially in cases such as echo questions where there is no syntactic cue to the interrogative mode. Inspired by the results of the positive and negative feedback experiment referred to above, an experiment was carried out to test if similar visual cues 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. 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 illustrate how this data can be used for modeling audiovisual prosody.

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 video signals poses a much more complicated 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 Qualisys 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 i.e. [8]. 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. In the present set-up, the second person cannot be recorded with the Qualisys system but is only video recorded. Audio data was recorded on DAT-tape and visual data was recorded using video and the optical motion tracking system. A synchronisation signal was used to match the audio and visual data. 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 (varying from 30 to 35 markers in different recording sessions) to register lip movements as well as other facial movements such as eyebrows, cheek, chin and eyelids. For several of the recordings, markers attached to a pair of spectacles and on the chest were used as a reference to be able to factor out head

and torso movements when analyzing specific articulator and facial movements.



Figure 2: Test subjects with the IR-reflecting markers

The data corpora described here was collected in the context of the EU project PF-Star [9]. 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. [10].

The corpora can be divided into two basic categories: prompted read speech and semi-spontaneous natural dialogues. The subjects were two non-professional actors.

The prompted read speech comprised VCV, VCCV and CVC nonsense words, sequences of digits, semantically neutral utterances such as names of cities, and short, content neutral sentences such as "*Båten seglade förbi*" (The boat sailed by) and "*Grannen knackade på dörren*" (The neighbor knocked on the door). The utterances were recorded in a variety of expressive modes including certain, confirming, questioning, uncertain, happy, and neutral. These were expressions which are considered to be relevant in a spoken dialogue system scenario. Much of the material was also recorded using Ekman's seven universal prototypes for emotions: happiness, sadness, surprise, disgust, fear, anger and neutral [11]. To elicit more visual prosody in terms of prominence, the short sentences were recorded with varying focal position, usually on the subject, the verb and the object respectively.

The spontaneous speech material consisted of two types of natural dialogue. In the first type an information seeking scenario was used in which one participant had the role of information seeker and one of information giver. The domains of the dialogue were movie information and direction giving. The information giver was the subject who was recorded with the IR markers. In the second scenario, the two participants were instructed to interact with each other pretending to be at a travel agency. They were to follow a script with ten different situations which were expected to lead to different expressions of emotions and production of different communicative expressions.

3. Examples of data analysis

Although the data corpora has been limited to only two subjects, one strength of the material is that it provides information on possible intra-speaker variation, particularly with regard to gestures intended to convey different

expressive states. Modelling such intra-speaker variation is of considerable importance in the development of an animated agent. Simple mean values of point locations reveal considerable intra-speaker differences. In Fig. 3 the mean values of the X and Y for the eyebrow markers are displayed for six different expressive states. For this speaker, neutral and certain are very similar while uncertain is displayed with lower eyebrows. For both happy and questioning a higher mean eyebrow position is used.

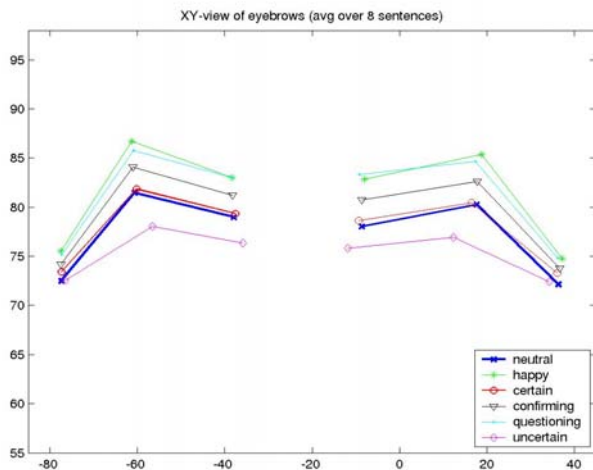


Figure 3: Mean eyebrow positions for one speaker.

3.1. Head movements

In terms of analyzing head movement, we have mainly been looking at the spontaneous material from the travel agent scenario and the short sentences where focus position was varied. The analysis has been made using a video plug-in to the Wavesurfer software [12] enabling synchronization of video and IR marker data. The marker on the nose was used for the detection of head nods.

In an analysis of the function of the head nods in the data from the travel agent scenario, Cerrato [13] categorized the nods as having main functions of feedback, turn managing, batonic (focus and emphasis), politeness, affirmative response and filler. She found that the majority of head nods had a feedback function (about 70%) often accompanying short expressions such as “mm, yes, ok.” Positive feedback predominated with negative feedback signaled by only 5% of the head nods.

In the short sentences, considerable variation was found concerning the presence of head nods accompanying focus. The variation is mostly conditioned by the expression the actor was trying to convey, and he probably employed head movement to differentiate between the expressions. The confirming expression almost always contained systematic head movement on the focused syllable, while the other expressions often lacked such movement. These confirming sentences were therefore used to test an automatic head nod detector described in [14]. Fig. 4 illustrates an example of the results of the automatic detector. The three panels represent the nose marker tracings from the sentence “Båten seglade förbi” (The boat sailed by). Focal accent is marked by capital letters. The tracings represent the vertical location of the nose tip marker as a function of time. The upper line, in each panel, is the actual location with the bold line showing the result of the automatic nod detector. The two lower lines in each panel

are the same measurements, but displaced to be visible. The two lower bold lines represent manual annotation of the nods by two annotators based on the video and tracing representation.

The example shown in Fig. 4 illustrates a series of head movements with the largest in amplitude co-occurring with the stressed syllable in the focused word. It is important to note that the head movements in these sentences probably signal both focal accent and confirmation. Here, the head movements are not simply a visual reinforcement of focal accent, but they also have other communicative functions. In the sentences conveying other expressions such as surprise, sadness or anger, focus was often signaled by means other than head movement such as eyebrow movement, head shaking, or extreme articulation. This is consistent with results reported on in [15][16].

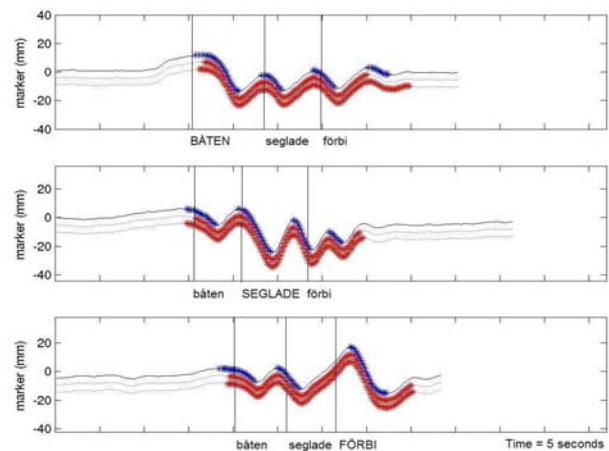


Figure 4: Nose marker traces with automatic and human annotated head nods (see text for details)

4. Experimental applications

The context of much of the work presented here is to create an animated talking agent capable of displaying realistic communicative behavior and suitable for use in conversational spoken language systems. One example of such an application in which the above findings can be implemented is found in our work on a virtual language teacher. The language tutor is an especially challenging application [17]. The effectiveness of language teaching is often contingent upon the ability of the teacher to create and maintain the interest and enthusiasm of the student. Pronunciation training in the context of a dialogue also automatically includes training of individual phonemes, sentence prosody and communication skills. In this context, visual prosody can function as an important aid to prosody training by reinforcing the acoustic prosody, as well as facilitating the flow of the training dialogue [18].

In a first prototype of the tutor the emphasis was put on vocabulary and quantity training in Estonian. Both perception training and production exercises were implemented. Evaluating phoneme duration is the first task of the pronunciation analyzer implemented in the prototype. The CTT aligner tool measures vowel length by determining and time-marking phone boundaries, based on a transcription of

what is being said and the waveform of the utterance. The time segments are then normalized and compared with a reference. Feedback is supplied both by the tutor and by graphs. Deviations in duration from the reference are signalled both by a remark from the Virtual Language Tutor and by rectangular bars below each phone in a transcription window. A database of average phoneme lengths or text-to-speech synthesis rules for phoneme lengths are being evaluated as possible reference instead of pre-recorded words. It is also possible to play a time warped version of the student utterance that conforms to the model/teacher pronunciation, thus supplying a “best pronunciation” example (Fig. 5) [19]. This new way of language training was very well received by the students. However, in addition to the need for normative descriptions of the prosody to be taught, there is a need to reliably estimate the limits of acceptable variation to be able to give useful feedback to the student.

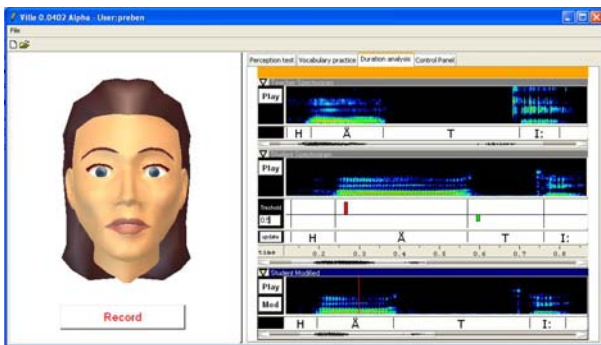


Figure 5: The CTT virtual language tutor set up for learning Estonian quantity

5. Acknowledgements

Much of the work presented in this overview has been done by other members of the CTT multimodal communication group including, Jonas Beskow, Loredana Cerrato, Olov Engwall, Mikael Nordenberg, Magnus Nordstrand, Gunilla Svanfeldt and Preben Wik 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.; Lundeberg, M., 1999. Prosodic Cues in Multimodal Speech Perception. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)*. San Francisco, 655-658.
- [2] Krahmer E.; Ruttkey Z.; Swerts M.; Wesselink W., 2002. 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.
- [3] Krahmer E.; Ruttkey Z.; Swerts M.; Wesselink W., 2002. Perceptual evaluation of audiovisual cues for prominence. In: *Proceedings 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colo., USA. 1933-1936.
- [4] House, D.; Beskow, J.; Granström, B., 2001. Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proc of Eurospeech 2001*, 387-390.
- [5] Granström, B.; House, D.; Swerts, M., 2002. 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.
- [6] House, D., 2002. Intonational and visual cues in the perception of interrogative mode in Swedish. In *Proceedings of ICSLP 2002*. Denver, Colorado, 1957-1960.
- [7] Srinivasan, R.J.; Massaro, D.W., 2003. Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech* 46(1), 1-22.
- [8] Beskow, J.; Engwall, O.; Granström, B., 2003. Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. of ICPHS 2003*. Barcelona, Spain, 431-434.
- [9] PF-STAR: <http://pfstar.itc.it/> (December 2005)
- [10] Beskow J.; Cerrato L.; Granström B.; House D.; Nordstrand M.; Svanfeldt G., 2004. 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.
- [11] Ekman P., 1982. *Emotion in the human face*. Cambridge University Press, New York.
- [12] Sjölander K.; Beskow J., 2000. WaveSurfer - an Open Source Speech Tool, In *Proc of ICSLP 2000, Vol. 4*, Beijing, 464-467.
- [13] Cerrato L., 2006. Linguistic functions of head nods. In *Proceedings of The Second Nordic Conference on Multimodal Communication*, Gothenburg University. In press.
- [14] Cerrato L.; Svanfeldt, G., 2006. 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.
- [15] Keating P.; Baroni M.; Mattys S.; Scarborough R.; Alwan A.; Auer E.; Bernstein L., 2003. Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English, In *Proc. 15th International Congress of Phonetic Sciences*, 2071-2074
- [16] Dohen M., 2005. *Deixis prosodique multisensorielle: Production et perception audiovisuelle de la focalisation contrastive en Français*. PhD thesis, Institut de la Communication Parlée, Grenoble.
- [17] Granström B., 2004. Towards a virtual language tutor. In *Proc. InSTIL/ICALL2004 - NLP and Speech Technologies in Advanced Language Learning Systems*, Unipress, Padova, 1-8.
- [18] Beskow, J.; Granström, B.; House, D.; Lundeberg, M., 2000. Experiments with verbal and visual conversational signals for an automatic language tutor. *Proc of InSTIL 2000*, Dundee, 138-142.
- [19] Engwall O.; Wik P.; Beskow J.; Granström B., 2004. Design strategies for a virtual language tutor, In *Proc. ICSLP 2004, vol 3*, Jeju, Korea, 1693-1696.