

Eyebrow movement as a cue to prominence

Björn Granström, David House and Magnus Lundeberg

(names in alphabetical order)

Centre for Speech Technology, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden

INTRODUCTION

Speech communication is inherently multimodal in nature. While the auditory modality often provides the phonetic information necessary to convey a linguistic message, the visual modality can qualify the auditory information providing segmental cues on place of articulation, prosodic information concerning prominence and phrasing and extralinguistic information such as signals for turn-taking, emotions and attitudes. Although these observations are not novel, prosody research has largely ignored the visual modality. One reason is the primary status of auditory speech, another is the relatively more complicated generation of visual speech. Most of the work that has been done in multimodal speech perception has concentrated on segmental cues in the visual modality.

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 information related to e. g. phrasing, stress, intonation and emotion are expressed by for example nodding of the head, raising and shaping of the eyebrows, eye movements and blinks.

These kinds of facial actions should also be taken into account 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.

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 utterance, etc. [4]. However, there have been attempts to apply such rules to facial animation systems [7]. A few such visual prosody rules have been implemented in our multimodal speech synthesis system [2].

This study is concerned with prosodic aspects of visual speech synthesis. A distinction can be made in visual synthesis between cues provided by the lower and the upper face. The lower face (e.g. lip aperture size, lip movement, jaw rotation, tongue position) provide information on place of articulation, vowel-consonant

alternation and syllable timing. Upper face (e.g. gaze and eyebrow movement) cues are more prosodic in nature in the sense that they overlie the segmental phonetic information of the lower face. This study aims at quantifying to what extent upper face movement cues can serve as independent cues for the prosodic functions of prominence.

METHOD

The test sentence used to create the stimuli for the experiments was ambiguous in terms of an internal phrase boundary. The stimuli were created using the KTH audio-visual text-to-speech synthesis [1] with our latest 3D face

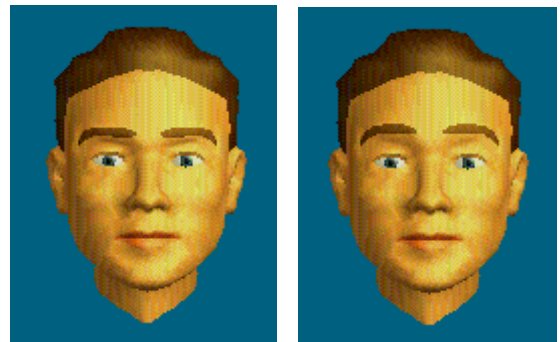


Figure 1. The synthetic face *Alf* with neutral eyebrows (left) and with eyebrows raised (right).

model *Alf*. 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 movements were created by hand editing the eyebrow parameter using the synthesis parameter editor *Veiron* [8]. The degree of eyebrow raising was chosen to create a subtle movement that was distinctive although not too obvious. The synthetic face *Alf* with neutral and with raised eyebrows is shown in Figure 1. Stimuli were recorded on video tape and presented to the subjects using a video projector and a separate loudspeaker.

In a previous study concerned with prominence and phrasing, using acoustic speech only, ambiguous sentences were used [3]. In the present experiment we used one of these sentences:

- (1) När pappa fiskar stör, piper Putte
(When dad is fishing sturgeon, Putte is whimpering)
- (2) När pappa fiskar, stör Piper Putte
(When dad is fishing, Piper disturbs Putte).

Hence, "stör" could be interpreted as either a noun (1) or a verb (2); "piper" (1) is a verb, while "Piper" (2) is a name.

In the stimuli, the acoustic signal is always the same, and synthesized as one phrase, i.e. with no phrasing prosody disambiguating the sentences. Six different versions were included in the experiment: one with no eyebrow movement and five where eyebrow rise was placed on one of the five content words in the test sentence. The audio was presented via loudspeakers and the face image was shown on a projected screen, four times the size of a normal head. 21 subjects were instructed to listen as well as to speech read.

RESULTS

Detailed results from the test are given in [6]. The general picture is that eyebrow movement timed to one word resulted in 20 to 70 % prominence judgement for that word (chance = 20%). The distribution of judgements varies with the word in the sentence. This could be related to phonetic information in the auditory modality since the intonational default synthesis used here put a weak focal accent on the first and the last word in a sentence. This could explain the many votes for the first and the last word, "pappa" and "Putte". However, it may well be related to prominence expectations. In experiments where subjects are asked to rate prominence on words in written sentences, nouns tend to get higher ratings than verbs [5].

DISCUSSION

The results of the experiment indicate that eyebrow raising can function as a perceptual cue to word prominence independent of acoustic cues and lower face visual cues. In the absence of strong acoustic cues to prominence, the eyebrows may serve as an F0 surrogate or they may signal prominence in their own right. While there was no systematic manipulation of the acoustic cues in this experiment, a certain interplay between the acoustic and visual cues can be inferred from the results. A weak acoustic focal accent in the default synthesis falls on the final word "Putte". Eyebrow raising on this word produces

the greatest prominence response. This could be a cumulative effect of both acoustic and visual cues, although compared to the results where the eyebrows were raised on the other nouns, this effect is not great.

In an integrative model of visual speech perception [7], eyebrow raising should signal prominence when there is no direct conflict with acoustic cues. In the case of "fiskar" the lack of specific acoustic cues for focus and the linguistic bias between nouns and verbs could account for the absence of prominence response for "fiskar". Further experimentation where strong acoustic focal accents are coupled to and paired against eyebrow movement could provide more data on this subject.

CONCLUDING REMARKS

This paper presents evidence that eyebrow movement can serve as an independent cue to prominence. Some interplay between visual and acoustic cues to prominence and between visual cues and word class/prominence expectation are also seen in the results. Further work on the interplay between eyebrow raising as a cue to prominence and eyebrow movement as a visual signal of speaker expression, mood and attitude will benefit the further development of visual synthesis methods for interactive animated agents in e. g. spoken dialogue systems.

REFERENCES

- [1] Beskow, J. 1995. Rule-based Visual Speech Synthesis In *Proceedings of Eurospeech '95*, Madrid, Spain.
- [2] Beskow, J. 1997. Animation of Talking Agents, In *Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing*, Rhodes, Greece.
- [3] Bruce, G., Granström, B. and House, D. 1992. Prosodic phrasing in Swedish speech synthesis. In Bailly, G., C. Benoit, and T.R. Sawallis (eds.), *Talking Machines: Theories, Models, and Designs*, 113-125. Amsterdam: North Holland.
- [4] Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. 1996. About the relationship between eyebrow movements and F0 variations. In Bunnell, H.T. and W. Idsardi (eds.), *Proceedings ICSLP 96*, 2175-2178, Philadelphia, PA, USA.
- [5] Fant, G. & Kruckenberg, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR 2/1989*, 1-80.
- [6] Granström B., House D., and Lundeberg L. 1999. Prosodic cues in multimodal speech perception, *Proceedings of ICPHS '99*, San Fransisco, 655-658.
- [7] Massaro, D. W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

[8] Sjölander, K., Beskow, J., Gustafson, J., Levin, E., Carlson, R. & Granström, B. 1998. Web-based educational tools for speech technology, In *Proceedings of ICSLP'98*, Sydney, Australia.