PROSODIC CUES IN MULTIMODAL SPEECH PERCEPTION

Björn Granström, David House and Magnus Lundeberg
Centre for Speech Technology, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden

ABSTRACT

Potential visual prosodic cues for prominence and phrasing comprising eyebrow movements were manipulated using a system for audio-visual text-to-speech synthesis which has been implemented based on the KTH rule-based synthesis. Two functions of prosody (prominence and phrasing) were tested in two separate experiments. A test sentence, ambiguous in terms of an internal phrase boundary, was used for both experiments. Acoustic cues and lower face visual cues were held constant for all stimuli. Upper face cues were eyebrow movement where the eyebrows were raised on successive words in the sentence. The results indicate a general coupling between eyebrow raising and perceived prominence while suggesting a more complicated relationship between eyebrow movement and phrasing. The results also point against a tight perceptual connection between F0 and eyebrow movement indicating that selective eyebrow movement can be effective as an independent prosodic cue to prominence.

1. 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 and phrasing.

2. METHOD

2.1. Stimuli

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 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* [9]. The degree of eyebrow raising was chosen to create a subtle movement that was distinctive although not too obvious. The total duration of movement was 500 ms and comprised a 100 ms dynamic raising part, a 200 ms static raised portion and a 200 ms dynamic lowering part. The synthetic face *Alf* with neutral and with raised eyebrows is shown in Figure 1. Stimuli were

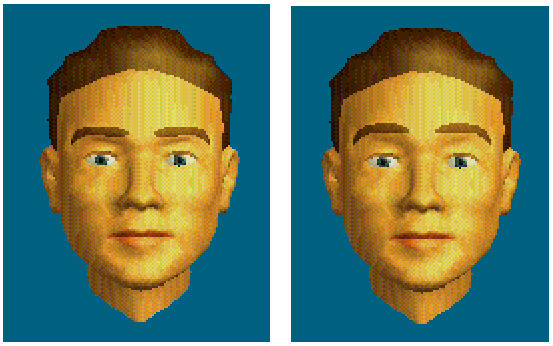

Figure 1. The synthetic face *Alf* with neutral eyebrows (left) and with eyebrows raised (right).

recorded on video tape and presented to the subjects using a video projector and a separate loudspeaker.

2.2. Subjects

In the two experiments described below the same 21 subjects participated. All were part of a speech technology class taught at KTH. No one reported any hearing loss or visual impairment. 14 subjects had Swedish as their mother tongue. All except one of these reported that they had a central Swedish (Stockholm) dialect.

7 subjects had other mother tongues than Swedish (1 Finnish, 2 French, 2 Italian and 2 Spanish), but all had working competence in Swedish (attending a masters level class given in Swedish at KTH). In the result section, the results of the total group as well as for these subgroups are presented.

3. EXPERIMENT 1 – PHRASING

3.1 Method

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. In [3] different segmental and prosodic disambiguation strategies are discussed. In the present series of experiments the possibility of visual disambiguation was investigated. 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. In the test list of 20 stimuli, each stimulus was presented three times in random order. The first and the last item of the list were dummies and not part of the data analysis.

All subjects participated in the same session. The audio was presented via loudspeakers and the face image was shown on a projected screen, four times the size of a normal head. The viewing distance was 3 to 6 meters, simulating a normal face-to-face conversation distance of 0.75 to 1.5 meters. In this range of distances the visual intelligibility is judged to be close to constant [8].

The subjects were instructed to listen as well as to speech read. Two seconds before each sentence an audio beep was played to give subjects time to look up and focus on the face. No mention was made of eyebrows. The subjects were made aware of the ambiguity in the test sentence and were asked to mark the perceived interpretation for each sentence.

3.2 Results

In Figure 2 the results from experiment 1 can be seen. It is obvious that there is a bias for all the stimuli to more often (about 60%) be perceived with a phrase boundary after ”stör”, i.e. interpretation (1). This is possibly also the default interpretation of the sentence without speech for most subjects, since Piper is a rather uncommon name. On the whole very little difference is seen between the different stimulus conditions.

The non-Swedish subjects behaved very much like the Swedes, perhaps with one exception. For the Swedish subjects there was a small increase in the (1) interpretation when there was an eyebrow rise on p/Piper. One possible explanation could be that an eyebrow movement could be associated with a phrase onset, but on the whole there is rather limited evidence in this experiment that eyebrow movements contributed to phrasing information.

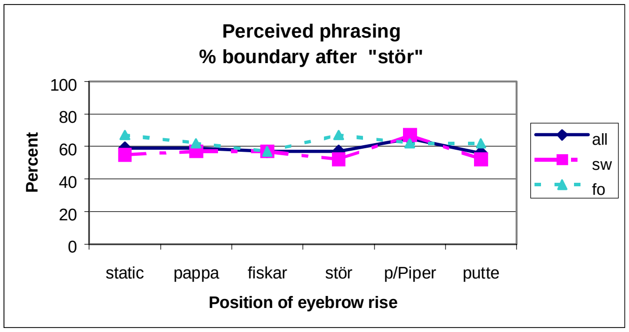

Figure 2. Result of the phrasing/disambiguation experiment. Interpretation (1) i.e. a phrase boundary after “stör” (rather than after “fiskar”). *sw*: 14 subjects with Swedish as their mother tongue. *fo*: 7 non-Swedish subjects

4. EXPERIMENT 2 – PROMINENCE

4.1 Method

In the second experiment we used the same stimulus material as in experiment 1, but the question was now concerned with prominence. The subjects were asked to circle the word that they perceived as most stressed/most prominent in the sentence.

4.2 Results

The results are shown in Figures 3a-f. Figure 3a refers to judgements when there is no eyebrow movement at all. Obviously the distribution of judgements varies with both subject group and 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” in Figure 3a. 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 [6]. This is supported by our data, since ”stör” has the default interpretation of a noun and p/Piper the default interpretation of a verb in experiment 1, while ”fiskar” is always a verb in these contexts. The non-Swedish subjects seem to behave slightly differently in this experiment, since no prominence votes are given to ”fiskar” and ”p/Piper”.

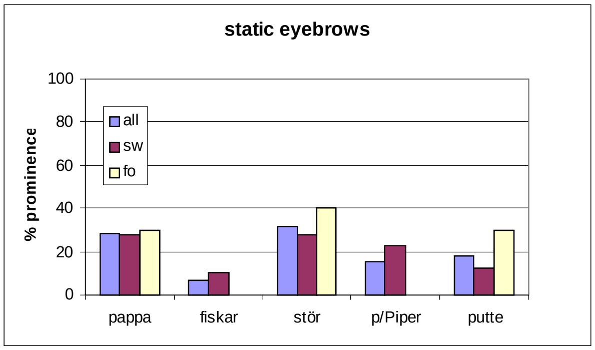

Figure 3a. Prominence responses in percent for each word for the stimulus with no eyebrow movement. Subjects are grouped as all, Swedish (sw) and foreign (fo).

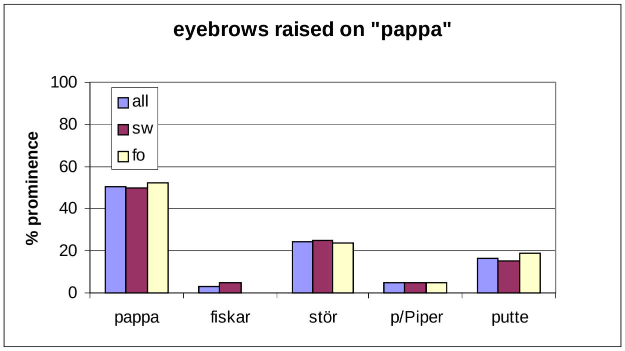

Figure 3b. Prominence responses in percent for each word for the stimulus with eyebrow movement on “pappa”.

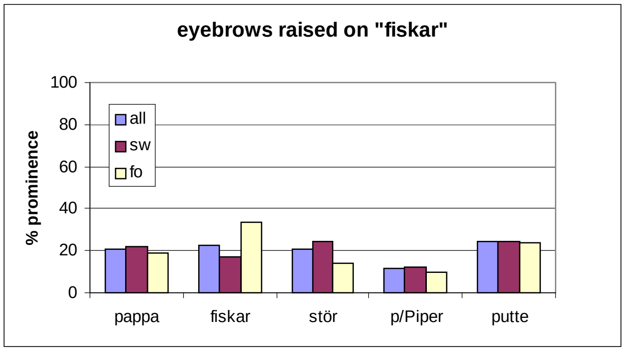

Figure 3c. Prominence responses in percent for each word for the stimulus with eyebrow movement on “fiskar”.

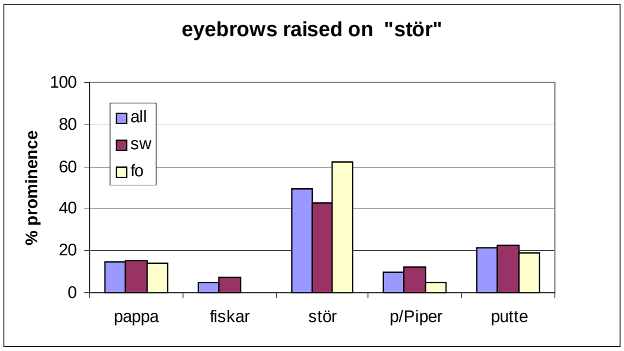

Figure 3d. Prominence responses in percent for each word for the stimulus with eyebrow movement on “stör”. Subjects are grouped as all, Swedish (sw) and foreign (fo).

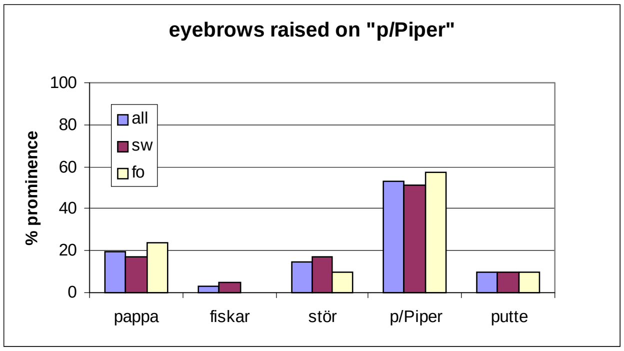

Figure 3e. Prominence responses in percent for each word for the stimulus with eyebrow movement on “p/Piper”.

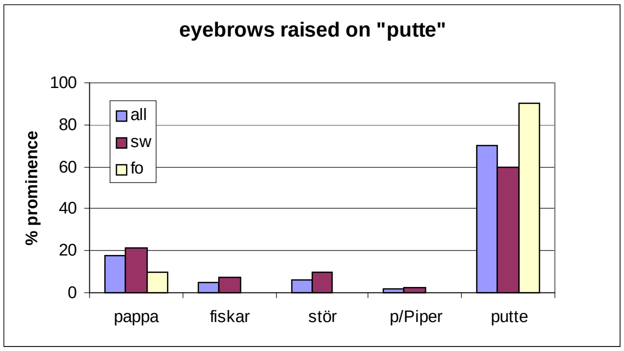

Figure 3f. Prominence responses in percent for each word for the stimulus with eyebrow movement on “Putte”.

5. DISCUSSION

5.1. Prominence

The results of the prominence 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. As mentioned above, a weak acoustic focal accent in the default synthesis falls on the final word ”Putte”. Eyebrow raising on this word (Figure 3f) produces the greatest prominence response in both listener groups. 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” (Figures 3a and 3c) the lack of specific acoustic cues for focus and the linguistic bias between nouns and verbs as mentioned above, 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.

It is interesting to note that the foreign subjects in all cases responded more consistently to the eyebrow cues for prominence. This might be due to the relatively complex stress/tone/focus signaling in terms of F0 in Swedish and the subjects’ non-native competence. It could be speculated that the eyebrow motion is a more universal cue for prominence.

5.2. Phrasing

Given the increased perceived prominence related to eyebrow movements, we could have expected more differentiated results in the phrasing experiment. This, however, would depend on the interrelationship between accentuation and phrasing, where increased prominence on ”fiskar” would serve as a phrase-final signal resulting in an increase in responses ”boundary after fiskar”, while more prominence on ”p/Piper” would serve as a phrase-initial signal and result in an increase in responses ”boundary after stör”. Since eyebrow movement did not result in a high vote for prominence on “fiskar” it is not surprising that boundary after “fiskar” did not increase. There was, however, a slight increase in “boundary after stör” responses among the Swedish listeners where eyebrow movement occurred on “p/Piper” (Figure 2).

The relationship between cues for prominence and phrase boundaries is not unproblematic [3]. The use of eyebrow movement to signal phrasing may involve more complex movement related to coherence within a phrase rather than simply as a phrase delimiter. It may also be the case that eyebrow raising is not an effective independent cue for phrasing, perhaps because of the more complex nature of different phrasing cues.

6. 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. Eyebrow raising as a cue to phrase boundaries was not shown to be effective as an independent cue in the context of the ambiguous sentence. 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 synthsis. In Bailly, G., C. Benoit, and T.R. Sawallis (eds.), *Talking Machines: Theories, Models, and Designs*, 113-125. Amsterdam: North Holland.

[4] Cavé, C., Guaïtella, 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] Cohen, M. M., & Massaro, D. W. 1993. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann & D. Thalmann (eds.) *Models and Techniques in Computer Animation*. 139-156, Tokyo: Springer-Verlag.

[6] Fant, G. & Kruckenberg, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1-80.

[7] Massaro, D. W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

[8] Neely, K. K. 1956. Effects of visual factors on intelligibility of speech. *Journal of the Acoustical Society of America*, 28, 1276-1277.

[9] 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.