Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception

David House, Jonas Beskow and Björn Granström

Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
{davidh|beskow|bjorn}@speech.kth.se

Abstract

The timing of both eyebrow and head movements of a talking face was varied systematically in a test sentence using an audiovisual speech synthesizer. The audio speech signal was unchanged over all sentences. 33 listeners were given the task of identifying the most prominent word in the test sentence. Results indicate that both eyebrow and head movements are powerful visual cues for prominence and that perceptual sensitivity to timing is on the order of a typical syllable duration of 100-200 ms.

1. Introduction

Innovative spoken dialogue systems are beginning to be characterized by designs which strive toward establishing a smooth flow of information modelled on conversational dialogues. In this context, there is considerable interest in developing 3D-animated agents to exploit the inherently multimodal nature of speech communication. As 3D-animation becomes more sophisticated in terms of visual realism, the demand for naturalness in speech and gesture coordination increases. Not only are appropriate and speech-synchronized articulator movements necessary, prosodic signals such as cues for prominence and phrasing, and conversational signals such as turntaking and feedback are also essential. Such signals can be conveyed by both the auditory and visual modality. Verbal (auditory) signals can complement syntax and interact with the prosodic (accentual and phrasal) structure of the utterances. For example, a phrase-final intonation pattern can function as both a cue for prosodic grouping and as a verbal turngiving signal. Gestural (visual) signals such as eyebrow movements and nodding for accentuation can function as parallel signals to intonation (i.e. as linguistic signals) as well as being used as conversational signals (e.g. raised eyebrows to signify an interested, listening agent or nodding to provide encouragement) [1].

There has been, on the one hand, considerable research carried out on the timing and synchronization of articulator movements in audiovisual speech processing (e.g. [2]). On the other hand, much work has also been done on describing spoken and gestural conversational signals in human to human interactions [3]. Work aimed at investigating the coordination of audio and visual prosodic signals in speech perception and the implementation of this knowledge in audio-visual synthesis is not as well represented.

Cassel et al. [4] have modelled speech and gesture in dialogue using two virtual agents, but no user interactivity. Katashi and Akikazu [5] employed animated facial expressions, but no gestures, as a back-channelling mechanism in a spoken dialogue system. Thorisson [6] used a 2D animated character together with input from many sources, including speech and gaze, to model mainly the social aspects of multi-modal dialogue interaction. A good source for relatively recent work in this area is Cassell, Sullivan, Prevost and Churchill [7].

The interaction between acoustic intonational gestures (F0) and eyebrow movements has been studied in production in e.g. [8]. A preliminary hypothesis is that a direct coupling is very unnatural, but that prominence and eyebrow movement may co-occur.

In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish [9], words and syllables with concomitant eyebrow movement were generally perceived as more prominent than syllables without the movement. This tendency was even greater for a subgroup of L2 listeners. For the acoustically neutral test sentence the mean increase in prominence response following an eyebrow movement was 24 percent for the Swedish L1 listeners and 39 percent for the L2 group. One example result is shown in Figure 1.

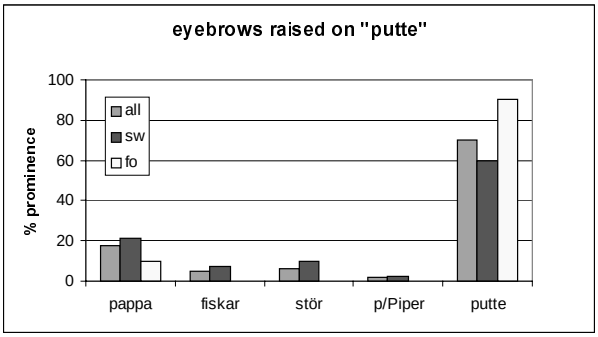

Figure 1. Prominence responses in percent for each content word for the acoustically neutral reading of the stimulus sentence, "När pappa fiskar stör p/Piper Putte," with eyebrow movement on "Putte". Subjects are grouped as all, Swedish (sw) and foreign (fo). (From Granström, House and Lundeberg [9])

This paper presents results from a follow-up study carried out in which 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 the larger surface movement. Secondly, we wanted to test the perceptual sensitivity to the timing of both eyebrow and head movement in relationship to the syllable.

2. Method

2.1. Stimuli

A rule-based audiovisual synthesizer was used for stimuli preparation [10] and [11]. In addition, a control interface that allows fine-grained control over the trajectories for acoustic as well as visual parameters has been developed. The interface is implemented as an extension to the WaveSurfer application [12], which is a tool for recording, playing, editing, viewing, printing, and labelling audio data. The interface makes it possible to start with an utterance synthesised from text, with all the parameters generated by rule, and then interactively edit the parameter tracks for any parameter, including F0, visual (non-articulatory) parameters as well as the durations of individual segments in the utterance to produce specific effects.

The test sentence used to create the stimuli for the experiment was the same as that used in a prior perception experiment designed to test acoustic cues only [13]. The sentence, *Jag vill bara flyga om vädret är perfekt* (I only want to fly if the weather is perfect) was synthesized with focal accent rises on both *flyga* (fly) (Accent 2) and *vädret* (weather) (Accent 1). The rise excursions corresponded to the stimulus in the previous test which elicited nearly equal responses for *flyga* and *vädret* in terms of the most prominent word in the sentence. The voice used was the Infovox 330 Ingmar MBROLA voice. The acoustic stimulus is illustrated in Figure 2.

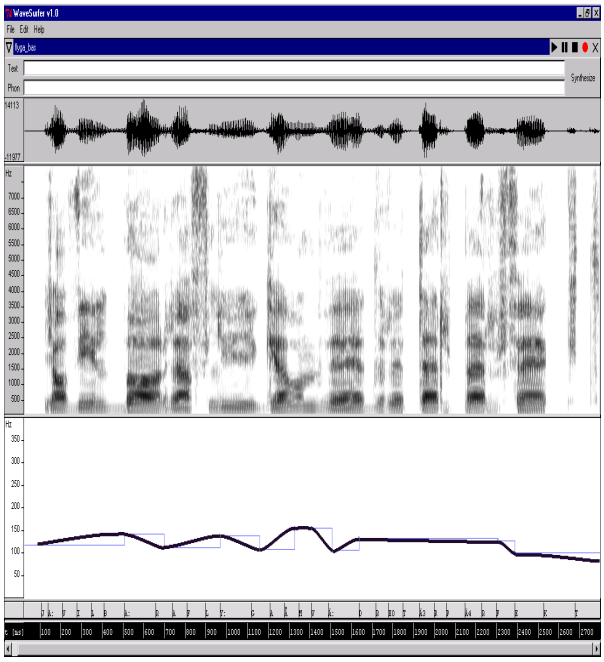

Figure 2. Waveform, spectrogram, F0 contour and transcription of the acoustic stimulus used in the experiment.

Eyebrow and head movements were then created by hand editing the respective parameters. The eyebrows were raised to create a subtle movement that was distinctive although not too obvious. In quantitative terms the movement comprised 4% of the total possible movement. The head movement was a slight vertical lowering comprising 3% of the total possible vertical head rotation. Figure 3 illustrates two conditions: no movement and the maximum simultaneous eyebrow and head displacement in the middle of the [ɛː] vowel of *vädret*. Statically, the displacement is difficult to perceive, while dynamically, the movement is quite distinct.

Figure 3. The synthetic face with neutral eyebrows and no vertical head displacement (left) and with eyebrows raised and head lowered (right).

The total duration of both eyebrow and head movement was 300 ms and comprised a 100 ms dynamic onset, a 100 ms static portion and a 100 ms dynamic offset.

Two sets of stimuli were created: set one in which both eyebrow and head movement occurred simultaneously and set two in which the movements were separated and potentially conflicting with each other. In set one, six stimuli were created by synchronizing the movement in stimulus 1 with the stressed vowel [yː] of *flyga*. This movement was successively shifted in intervals of 100 ms towards *vädret* resulting in the movement in stimulus 6 being synchronized with the stressed vowel [ɛː] of *vädret*. In set two, stimuli 1-3 were created by fixing the head movement to synchronize with the stressed vowel of *vädret* and successively shifting the eyebrow movements from the stressed vowel of *flyga* towards *vädret* in steps of 100 ms. Stimuli 4-6 were created by fixing the eyebrow movement to *vädret* and shifting the head movement from *flyga* towards *vädret*. The acoustic signal and articulatory movements were the same for all stimuli. A schematic illustration of the stimuli is presented in Figure 4.

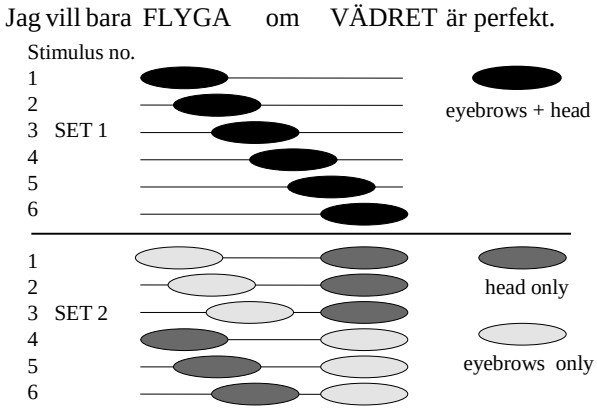

Figure 4. Schematic illustration of face gesture timing

2.2. Subjects and procedure

A total of 33 subjects (18 females and 15 males) participated in the experiment. Most of the subjects were recruited from among students and staff at KTH. No one reported any hearing loss or visual impairment and all were native speakers of Swedish with the central Swedish (Stockholm) dialect predominating.

The stimuli were presented to each subject individually using a computer interface especially designed for the experiment. The audio was presented through headphones and the face was displayed in a frame measuring 13 x 19 cm with the face itself measuring 12 x 18 cm. A 3D graphic accelerator was installed in each test computer to insure sufficient temporal resolution for audio-visual synchronization (average frame rate was at least 80 frames per second).

Subjects were asked to listen to each stimulus while looking carefully at the face and given the task of choosing which of the two words, *flyga* or *vädret*, was most prominently accented. The subjects were also requested to indicate on a scale between 1 to 5 how confident they were of their choice where 5 was certain and 1 was guessing. The presentation order of the stimuli was randomized within each set. The subjects were allowed to listen and look at each stimulus as many times as they wished before making their choice and proceeding to the next stimulus.

3. Results

3.1. Stimulus set 1

The results from stimulus set 1 where eyebrow and head movements occurred simultaneously clearly reflect the timing aspect of these stimuli as can be seen in Figure 5 where percent votes for *vädret* increase successively as movement is shifted in time from *flyga* to *vädret*. Results for stimulus set 1 were consistent between subjects: single factor ANOVA $F(32,165)=1.25$, $p=0.188$, and significant between stimuli: $F(5,192)=21.84$, $p<0.001$.

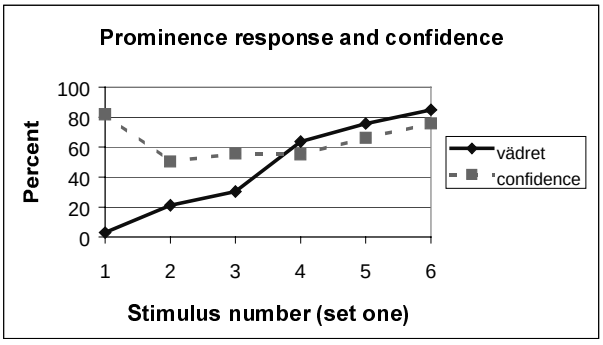

Figure 5. Results for stimulus set one showing prominence response for vädret and confidence in percent.

The endpoint stimuli 1 and 6 where face movement and stressed vowel were in synchrony elicited the greatest number of prominence votes for the respective words. This is also reflected by the high confidence scores for stimuli 1 and 6.

The largest difference between successive stimuli is to be found in the results between the middle stimuli 3 and 4. Only 30% of the responses for stimulus 3 favored *vädret* while for stimulus 4, 64% of the responses favored *vädret*. A single factor ANOVA on adjacent stimuli reflects this difference in that the difference between stimulus 3 and 4 is significant at the 1% level: $F(1,64)=8.033$, $p<0.01$. The only other significant difference was between pair 1 and 2: $F(1,64)=5.383$, $p<0.05$

3.2. Stimulus set 2

The results from stimulus set 2 where eyebrow and head movements were timed separately reflect more ambiguity in the subject responses as can be seen in Figure 6. Results for stimulus set 2 were consistent between subjects: single factor ANOVA $F(32,165)=1.336$, $p=0.124$, and significant between stimuli: $F(5,192)=3.098$, $p<0.05$.

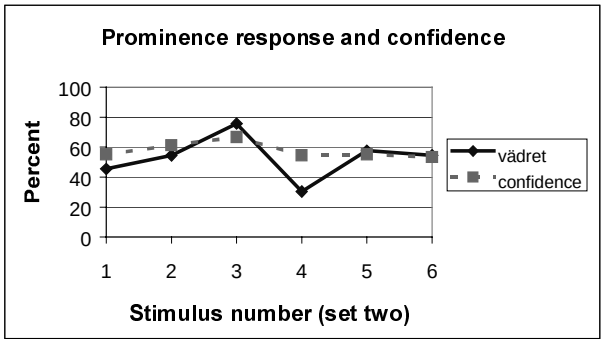

Figure 6. Results for stimulus set two showing prominence response for vädret and confidence in percent.

Only stimuli 3 and 4 received non-ambiguous responses. In stimulus 3, the responses favoring *vädret* (76%) reflect the placement of the head movement. In this stimulus, the head movement is synchronized with the stressed vowel in *vädret* and eyebrow movement is 200 ms delayed from *flyga*. In stimulus 4, the responses also reflect the placement of the head movement and favor *flyga* (60%). In this stimulus, head movement is synchronized with the stressed vowel in *flyga* and eyebrow movement is synchronized with *vädret*. A single factor ANOVA on adjacent stimuli reflects the difference in responses between stimulus 3 and 4: $F(1,64)=16.74$, $p<0.001$. The only other significant difference was between pair 4 and 5: $F(1,64)=5.22$, $p<0.05$.

4. Discussion

It is clear from the results that combined head and eyebrow movements of the scope used in the experiment are powerful 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 seems to be on the order of 100 ms. However, there is a tendency for integration of the movements to the nearest potentially prominent word, thus accounting for the jump in prominence response between stimulus 3 and 4 in set 1. This integration is consistent with the results of similar experiments using visual and auditory segmental cues [14].

The results from set 2 where eyebrow and head movement conflict are not surprising and demonstrate that both head movement and eyebrow movement can function as

independent cues to prominence. Head movement shows a slight advantage revealed by differences in results for stimulus 3 and 4 where head movement synchronized with the stressed vowel determines prominence responses overruling eyebrow movement. The relative salience of head movement is also apparent in the results for stimulus 6 where head movement three positions from *flyga* detracts from eyebrow movement synchronized with *vädret*. If eyebrow movement were equally powerful it would have been expected to prevail over the ambiguous head movement in stimulus 6, as was the case for head movement prevailing in stimulus 3. The advantage of head movement can perhaps be explained by virtue of the larger surface area in motion. The advantage might even be increased if a smaller head were used for example as an agent in a dialog system. In an informal demonstration, where subjects were 2 to 5 meters from the screen using the same head size as in the current experiment, head-movement advantage was quite pronounced.

A number of questions remain to be answered, as a perception experiment of this type is necessarily restricted in scope. Amplitude of movement was not addressed in this investigation. If, for example, eyebrow movement were exaggerated, would this counterbalance the greater power of head movement? A perhaps even more crucial question is the interaction between the acoustic and visual cues. There was a slight bias for *flyga* to be perceived as more prominent (one subject even chose *flyga* in 11 of the 12 stimuli), and indeed the F0 excursion was greater for *flyga* than for *vädret*, even though this was ambiguous in the previous experiment. In practical terms of multimodal synthesis, however, it will probably be sufficient to combine cues, even though it would be helpful to have some form of quantified weighting factor for the different acoustic and visual cues.

Duration of the eyebrow and head movements is another consideration which was not tested here. It seems plausible that similar onset and offset durations (100 ms) combined with substantially longer static displacements would serve as conversational signals rather than as cues to prominence. In this way, non-synchronous eyebrow and head movements can be combined to signal both prominence and e.g. feedback giving or seeking. Some of the subjects also commented that the face seemed to convey a certain degree of irony in some of the stimuli in set 2, most likely in those stimuli with non-synchronous eyebrow movement. Some very preliminary experimentation along these lines has been applied to the simulated use of a talking head as an automatic tutor in [15].

5. Conclusions

The results of this investigation indicate the appropriateness of using both head movement and eyebrow movement as visual markers of prominence in multimodal synthesis. The two parameters can be used in synchrony or as separate and independent cues. Synchronization with the stressed syllable is important, but perhaps not absolutely critical as a large degree of visual integration seems to occur within 100 ms of synchronization with the syllable.

6. Acknowledgements

The research reported here 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 organizations.

7. References

[1] Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepinies & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline: Contributions to the Colloquium,* 169-248. Cambridge: Cambridge University Press.

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

[3] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought,* Chicago: University of Chicago Press.

[4] Cassel, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S. and Achorn B. (1994). Modeling the Interaction between Speech and Gesture, In *Proceedings of 16th Annual Confer-ence of the Cognitive Science Society,* Georgia Institute of Technology, Atlanta, USA.

[5] Katashi, N. and Akikazu, T (1994). Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94), 102-109.

[6] Thórisson, K. R. (1997). Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People, In *Proceedings of First ACM International Conference on Autonomous Agents*, Marina del Rey, California, 536-537.

[7] Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. (eds.) (2000). *Embodied Conversational Agents*, Cambridge MA: The MIT Press.

[8] 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.

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

[10] Beskow, J. (1995). Rule-based Visual Speech Synthesis. In *Proceedings of Eurospeech '95,* 299-302. Madrid, Spain.

[11] Beskow, J. (1997). Animation of Talking Agents. In *Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing,* 149-152. Rhodes, Greece.

[12] Beskow, J. and Sjölander, K. (2000). WaveSurfer - a public domain speech tool. In *Proceedings of ICSLP 2000*, vol. 4, pp. 464-467. Beijing, China.

[13] House, D. (Forthcoming). Focal accent in Swedish: Perception of rise properties for accent 1. In van Dommelen, W. and Fretheim, T. (eds.) *Nordic Prosody 8*, Frankfurt: Peter Lang.

[14] Massaro, D. W., Cohen, M. M. and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am. 100*. 1777-1786.

[15] Beskow, J., Granström, B., House, D. and Lundeberg, M. (2000). Experiments with verbal and visual conversational signals for an automatic language tutor. In *Proceedings of InSTiL 2000,* 138-142. Dundee, Scotland.