The importance of different facial areas for signalling visual prominence

Marc Swerts and Emiel Krahmer

Communication and Cognition
Tilburg University, The Netherlands
{m.g.j.swerts/e.j.krahmer}@uvt.nl

10.21437/Interspeech.2006-377

Abstract

This article discusses the processing of facial markers of prominence in spoken utterances. In particular, it investigates which area of a speaker's face contains the strongest cues to prominence, using stimuli with the entire face visible or versions in which participants could only see the upper or lower half, or the right or left part of the face. To compensate for potential ceiling effects, subjects were positioned at a distance of either 50cm, 250cm or 380cm from the screen which displayed the film fragments. The task of the subjects was to indicate for each stimulus which word they perceived as the most prominent one. Results show that, while prominence detection becomes more difficult at longer distances, the upper facial area has stronger cue value for prominence detection than the bottom part, and that the left part of the face is more important than the right part. Results of mirror-images of the original fragments show that this latter result is due both to a speaker and an observer effect.
Index Terms: prominence, facial areas, audiovisual speech

1. Introduction

One important aspect of the human's perceptual mechanism is its capacity to integrate input from various sensory modalities (e.g. vision, audition, touch, taste). The way we perceive our environment is essentially multimodal in nature as our brain fuses modalities to produce a coherent percept. This is very clear from observing various ways in which visual cues have an impact on the way acoustic information is decoded: what people "hear" is affected by what people "see" as well (Kohlrausch & van de Par 1999). Various studies have shown that humans are especially sensitive to visual cues coming from a speaker's face (e.g. McGurk & MacDonald, 1976). The current paper is concerned with the perceptual integration of visual markers of prosodic prominence, i.e., the feature by which some words are perceived as more salient than other words in an utterance. In particular, it investigates which area of a speaker's face contains the strongest cues to prominence. There are reasons to believe that the different parts of a face are not equivalent in their signalling value. The kinds of evidence, both for the vertical and the horizontal axis, are acoustic and perceptual in nature.

If we take a vertical perspective on the face, there is evidence that prominence markers are distributed across the face. Following earlier claims by Ekman (1979), various people have suggested that eyebrow movements can signal prominent words in an utterance (see also Cassell et al. 2001). Important cues may also be located in the mouth area of the face. Keating et al. (2003) found that some of their speakers produce accented words with greater interlip distance and more chin displacement. Similarly, Erickson et al. (1998) showed that the increased articulatory effort for realizing accented words correlates with more pronounced jaw movements. Munhall and Vatikiotis-Bateson (1996) report that the size and velocity of lip movements vary with lexical stress. In addition, there is perceptual evidence that the upper and lower part of a speaker's face do not have equivalent cue value. For instance, it has been shown that practiced observers spend more time looking at and direct more gazes toward the upper facial region when making stress and intonation decisions compared with when making word identity decisions (Lansing & McConkie 1999). Intuitively, one might think that facial distinctions in the horizontal domain may not be that crucial for prominence perception. Nevertheless, there are also indications that the left and right parts of a human's face differ in this respect. There is physiological evidence which shows that faces are a-symmetric in the sense that the left part of a face is not simply the mirror image of the right part. That can most easily be demonstrated with the use of photograph manipulations in which a full image of a face is recreated by combining either the left side of a face with its mirror image, or vice versa with the right side, the endproduct of which differs perceptually from the original complete picture. Directly related to accents, there is empirical evidence from Keating et al. (2003) and Cavé et al. (1996), who report correlations between accented words and eyebrow movements, especially in the left eyebrow. Perceptually, Mertens et al. (1993) showed that subjects looking at faces more often focus their eyes on the left side of the picture, whereas they do not have such a bias when observing an artefact like a vase. Thompson et al (2004) report findings of an experiment in which they had their subjects view faces on which small dots appeared at random positions on the face, and instructed them to react as fast as possible whenever they detected such a spot. This test revealed that the left side of a face was predominant from a perceptual point of view. Given the physiological and perceptual variation in the horizontal domain, it remains to be investigated whether possible left-right distinctions in cue value for prominence are due to a speaker or observer effect.

To investigate the cue value of different facial areas, we set up a perception test in which we explored how sensitive observers are for different facial areas when they are instructed to rate the prominence of spoken words. In the following, we will present the audiovisual recordings we used for stimulus creation, and the specific experiments. We then present the perceptual results, and end with a discussion of our main findings.

2. Audiovisual recordings

As a basis for our experiment, recordings were made of 6 native speakers of Dutch (4 male, 2 female) between the ages of

20 and 40. In order to remove any visually distracting features, speakers did not wear any remarkable cloths, and were asked to take off their glasses during the data collection procedure. They were instructed to read out different variants of the sentence "Maarten gaat maandag naar Mali" (*Maarten goes Monday to Mali*) and had to produce the utterance in such a way that the first (Maarten), second (maandag) or third content word (Mali) of the sentence would receive an accent. These three target words, which will be referred to as W1, W2 and W3 in the remainder of this paper, were comparable in the sense that they were all bisyllabic words with stress on the first syllable. This stressed syllable began with a labial consonant /m/, which was chosen to increase the visibility of the articulatory movements, i.e., the lips, to produce the sound. In addition, they were asked to utter the sentence in a monotone, so without any auditory or visual markers of an accent. The actual recordings were organised in different blocks of 4 sentence productions, in which a speaker was first asked to utter the sentence in a monotone, and then the 3 realisations with an accentual marking of the first, second or third target word. The audiovisual recordings of all 6 speakers were made in a quite research laboratory at Tilburg university. Speakers were seated on a chair in front of a digital camera that recorded their upper body and face (frontal view) (25 fps). The camera was positioned about 2 meters in front of the speakers. In order to get optimal visual recordings, the speakers were seated against a white background and on a white floor, with 2 spotlights next to the camera focused on the floor in order to mimize reflections. These audiovisual recordings were used as a basis for the stimulus preparations of our perception experiment.

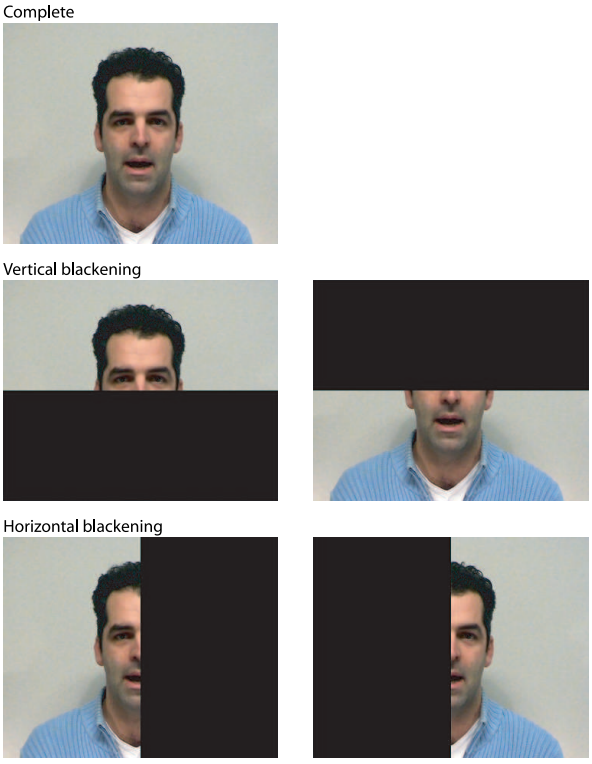

Figure 1: *Different stills which represent different versions of our stimuli, in which the face of our speakers is either completely or partly visible.*

3. Experiment

3.1. *Method*

3.1.1. Stimulus preparations

The stimuli used for this experiment are based on the audiovisual recordings described in the earlier section. Given that the current test is set up to learn more about the relative cue value of different facial areas, we did not include auditory markers of prominence in our design. Our procedure consists of three kinds of manipulations. The first one consists of mixing the monotone realisation of the utterances with the different visual realisations by our 6 speakers. In other words, in the current experiment, the auditory information was always identical for all the stimuli per speaker. After having created all the audiovisual stimuli this way, we produced 5 versions from the video-recording of the full face, by blackening parts of the face, using Adobe Premiere^{TM} as a tool. In the vertical domain, we generated a version with only the upper part of the face visible by blackening the mouth area from the bottom of the video up to roughly the middle of a speaker's nose; the opposite manipulations consisted of versions in which the part from the top of the video down to the middle of the nose was blackened. The left-right manipulations consisted of either blackening the left or right part of the face, from the edge of the video to roughly the middle of a speaker's face. Figure 1 gives some representative stills from one of our speakers (EK). After having created these different versions, we made mirror images of all 5 versions of these stimuli. Figure 2 illustrates an original image together with its mirror.

All the manipulations led to a total of 180 stimuli: 6 speakers $\times$ 3 visual prominences $\times$ 5 facial versions (original, vertically blackened, horizontally blackened) $\times$ 2 displays (original or mirrored). Since only one sentence was used for all recordings, the naturalness of the artificial stimuli was extremely good. An informal inspection of the data did not reveal cases of undesired lipsync effects.

3.1.2. Participants

There were 66 participants (36 male, 30 female) who took part in this experiment on a voluntary basis, students and colleagues from Tilburg University and other academic institutions nearby. The average age of the subjects was 25.5 years old, and they all had normal or corrected to normal vision and good hearing.

3.1.3. Procedure

The task was to indicate which word (W1, W2, or W3) was the more prominent one in a stimulus utterance. The experiment was a paper-and-pencil test and subjects were not requested to react as fast as possible. Subjects were also told that the person with the highest amount of correctly detected accents would receive a book token.

Pilot observations revealed that this task was very easy when participants could see the video clips on a full screen at a normal viewing distance, so that this would lead to ceiling effects, making it difficult to observe any difference between various conditions. Therefore, we decided to manipulate the degree of visibility of our stimuli in a number of respects. First, we made the video recordings smaller, by reducing the size to 185 by 165 pixels, corresponding to 4.8 by 4.3 centimeters. In addition, we added the distance from the screen as a between-subjects factor, in the sense that one third of the subjects had

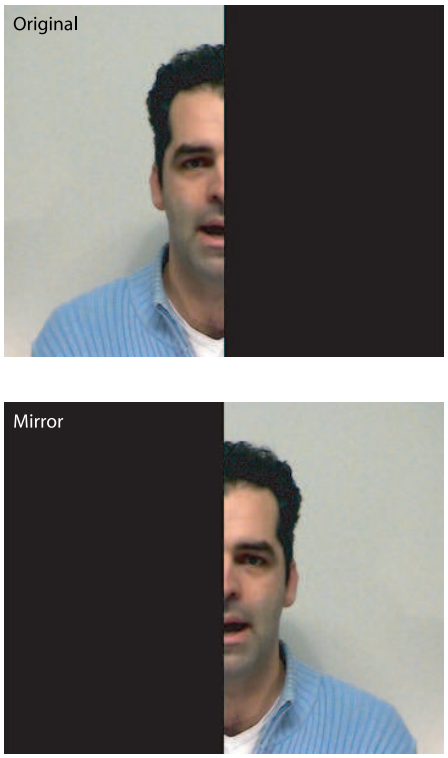

Figure 2: *Two representative stills of a facial expression presented in original or mirrored condition.*

to do the experiment at a "normal" distance from the screen (approximately 50 centimeters from the screen), in the middle condition subjects were positioned at 250 centimeters from the screen, and in the far condition at 380 centimeters from the screen. The middle and far conditions were chosen given some natural conditions of the size of the table on which the screen was positioned, and the size of the room.

The stimulus materials were shown on a Philips True Color PC screen (107 T 17") of 1024 by 768 pixels. The screen was calibrated before experimentation to guarantee that no black edges would be displayed on the screen. The inter-stimulus interval was 3 seconds, in which time frame participants had to circle in a multiple-choice on an answer sheet whether they thought the first, second or third target word was the more prominent one (forced choice). All stimuli were only presented once. Half of the subjects saw the original stimuli, and half of them saw their mirror versions. The actual experiment was preceded by a short test phase to make participants acquainted with the general set-up. The experiment, including instructions and test phase, lasted about 20 minutes per subject.

4. Results

The experiment has a complete $3 \times 6 \times 5 \times 2 \times 3$ design with the following factors: Visual prominence (3 levels: prominence on W1, W2, or W3), Speaker (6 levels), Facial area (5 levels: complete face, upper part visible, bottom part visible, left area visible, right area visible), Display (2 levels: original, mirrored) and Distance (3 levels: close, middle, far). Table 1 gives a first overall impression of how the responses are distributed for various positions of the visual accents. As can be seen from the numbers on the diagonal in the confusion matrix, subjects tend to perceive the word which receives the visual accent as being the more prominent one.

Table 1: Distribution of subjects' chosen prominences for different visual prominences.

	Chosen prominence			
Visual prominence	W1	W2	W3	Total
W1	1416	307	255	1978
W2	425	1361	191	1977
W3	433	210	1337	1980

The data were analysed with a multinomial logistic regression with the aforementioned variables as independent factors, and the subjects' scores as dependent variable. Scores were represented as a binary variable, either as correct (the response is identical to the position of the visual accent) or incorrect. A customized model which only tests main effects revealed significant main effects for Visual Accent ($\chi^2 = 9.537, df = 1, p < .01$), Facial area ($\chi^2 = 319.441, df = 4, p < .001$), Speaker ($\chi^2 = 176.433, df = 5, p < .001$) and Distance ($\chi^2 = 681.051, df = 2, p < .001$), whereas the effect of Display was not significant. This model accounts for 24% of the variance. Table 2 reveals that initial accents are most often detected correctly, whereas detection becomes increasingly poorer for accents in middle and last sentence position. With respect to the effect of facial area, we see that a whole face presentation leads to the best accent detection, whereas a display of the upper and left part of the face leads to better results that the bottom and right part, respectively. Showing a video in its original format or in mirror image does not generate a significant main effect. Table 2 also shows that stimuli from different speakers lead to markedly different results due to differences in expressiveness between speakers, with relatively poor detection for stimuli from speaker MB and best results for speaker PB.

While Display did not have a main effect, it turned out, using a model which included interactions, that there was a significant 2-way interaction between Facial area and Display ($\chi^2 = 36.533, df = 4, p < .001$). To get more insight into this, we also ran split analyses for different Facial areas (three separate analyses for whole face stimuli, for stimuli with manipulations in the vertical domain, and for stimuli with manipulations in the horizontal domain). Interestingly, the split analyses only reveals a significant interaction between Facial area and Display for horizontally blackened stimuli, but not for whole face stimuli, or for stimuli manipulated in the vertical domain. This can be explained using the data given in Table 3 which reveals that the scores for accent detection at different distances is about the same for original and mirrored display, when stimuli are presented as a whole face or with vertical manipulations. However, the data are quite different for the data shown at the bottom part of this table, which relate to variation in the horizontal domain. First, if we only focus on the column with data for stimuli in their original display, we observe that accent detection goes better if viewers can see the left part of the face than if they see the right part of the face. Second, if we compare the scores for original images with the presentation of their mirrors, we observe that scores become worse when the left side is shown as the right side, while the reverse is true for the original right side becoming left side.

Table 2: Percentage correct prominence detection as a function of different parameters: Main effects

Factor	Level	% correct
Visual prominence	W1	71.5
	W2	68.7
	W3	67.5
Facial condition	Complete	77.3
	Only top visible	77.3
	Only bottom visible	51.4
	Only left visible	75.6
	Only right visible	64.7
Distance	Close	86.7
	Middle	70.4
	Far	50.7
Display	Original	69.6
	Mirrored	68.9
Speaker	EK	72.7
	LL	73.2
	MB	54,4
	ME	71.8
	MS	66.1
	PB	77.3

Table 3: Percentage correct prominence detection as a function of combined settings of display, distance, and facial area

Facial area	Distance	Display	
		Original	Mirrored
Complete face	Close	94.9	88.4
	Middle	79.3	81.3
	Far	63.1	56.6
Vertical			
Only top visible	Close	92.9	92.9
	Middle	76.8	76.8
	Far	69.2	55.1
Only bottom visible	Close	72.2	65.7
	Middle	51.5	48.5
	Far	31.8	38.9
Horizontal			
Only left visible	Close	92.9	88.4
	Middle	80.3	78.3
	Far	62.1	51.5
Only right visible	Close	86.9	91.4
	Middle	57.6	73.7
	Far	32.3	46.5

5. Discussion

Our research has shown that observers are sensitive to visual cues from a speaker's face to signal prosodic prominence. However, the cue value differs for different facial areas. In the vertical domain, it turns out that the upper part of a speaker's face is more important than the bottom part, in line with earlier claims by Lansing and McConkie (1999) that subjects tend to focus on the area around the eyes when making prosodic judgments, while the mouth area is more important for word identity decisions (lipreading). In addition, we found that the left area of a speaker's face is perceptually more salient for signalling prominence than his or her right area. Our results both with original videos and videos in mirror format reveal that this preference for the left side is due to a combined speaker and observer effect. That is, a speaker's original left side is always the facial area which gives the more prominent cues, whether it is shown in its original format or in mirror image. This effect could be attributed to a speaker effect. However, that left side becomes less prominent when it is shown as a speaker's right side, which appears to be related to an observer's effect. The reverse effects are true for the speakers' right side of a face, whether shown in original or mirrored display.

6. Acknowledgments

This research is part of the FOAP project (http://foap.uvt.nl), funded by the Netherlands Organisation of Scientific Research (NWO). We thank Marleen Roffel, Gwendolyn Tabak, and Lennard van de Laar for experimental assistance.

7. References

Cassell, J., Vihjálmsson, H., Bickmore, T., (2001), BEAT: the Behavior Expression Animation Toolkit, Proceedings of SIGGRAPH01, pp. 477-486.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. (1996) About the relationship between eyebrow movements and F0 variations, Proceedings ICSLP, Philadelphia, pp. 2175-2179.

Ekman, P. (1979). About brows: Emotional and conversational signals. In: M. von Cranach et al. (Eds.), *Human Ethology* (pp. 169–202). Cambridge: CUP.

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. ICPhS* (pp. 2071–2074), Barcelona.

Kohlrausch, A., & van de Par, S. (1999). Audio-visual interaction: from fundamental research in cognitive psychology to (possible) applications. In *Proc. SPIE*, **3644**: 34–44.

Lansing, C.R. & McConkie, G.W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech and Hearing Research*, **42**, 526–539.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* **264**: 746–748.

Mertens, I., Siegmund H, Grusser O. (1993) Gaze motor asymmetries in the perception of faces during a memory task. *Neuropsychologia*, **31** (9):989-98.

Munhall, K.G. & Vatikiotis-Bateson, E. (1996). The moving face during speech communication. In Campbell, R., Dodd, B. & Burnham, D. (Eds.) *Hearing by Eye II. Advances in the Psychology of Speechreading and Auditory-Visual Speech*, London: Psychology Press.