Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 36 (2008) 219–238
Journal of Phonetics
www.elsevier.com/locate/phonetics

Facial expression and prosodic prominence: Effects of modality and facial area

Marc Swerts*, Emiel Krahmer

Communication and Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands

Received 10 October 2006; received in revised form 21 May 2007; accepted 22 May 2007

Abstract

This article addresses two related questions regarding the perception of facial markers of prominence in spoken utterances: (1) how important are visual cues to prominence from the face with respect to auditory cues? and (2) are there differences between different facial areas in their cue value for prosodic prominence? The first perception experiment tackles the relation between auditory and visual cues by means of a reaction-time experiment. For this experiment, recordings of a sentence with three prosodically prominent words were systematically manipulated in such a way that auditory and visual cues to prominence were either congruent (occurring on the same word) or incongruent (in that the auditory and the visual cue were positioned on different words). Participants were instructed to indicate as fast as possible which word they perceived as the most prominent one. Results show that participants can more easily determine prominence when the visual cue occurs on the same word as the auditory cue, while displaced visual cues hinder prominence perception. The second experiment investigates which area of a speaker's face contains the strongest cues to prominence, using stimuli with either the entire face visible or only parts of it. The task of the participants was to indicate for each stimulus which word they perceived as the most prominent one. Results show that 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.

1. Introduction

One important aspect of the human's perceptual mechanism is its remarkable capacity to integrate input from various sensory modalities (e.g. vision, hearing, touch, taste). The way we perceive our environment is essentially multimodal in nature as our brain fuses information from different modalities to produce a coherent percept. This has been shown, for instance, by the 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" (Bertelson, Vroomen, de Gelder, & Driver, 2000; Kohlrausch & van de Par, 1999). That is, when humans are processing incoming sounds, they are not only analysing the auditory signal which enters the perceptual system through

*Corresponding author. Tel.: +31 13 4662922; fax: +31 13 4663110.
E-mail address: m.g.j.swerts@uvt.nl (M. Swerts).

0095-4470/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2007.05.001

the ears, but they also process information in the visual signal, where observers tend to be especially sensitive to visual cues from a speaker's face (e.g. McGurk & MacDonald, 1976). While previous work on how our perceptual system integrates auditory speech information and visual cues from a speaker's face has largely concentrated on effects at the segmental level, our knowledge of audiovisual interactions at the prosodic level is very limited. The current paper addresses the latter problem, in particular dealing with the perception of prominence, which can be characterized as the property of some words to "stand out" with respect to other words in the same utterance. For instance, in response to the English question "Who went to Malta?", the utterance "Amanda went to Malta" would typically be produced with an accent on the first word of the sentence, which would make this word perceptually more salient than the words in the remainder of that sentence.

Most of the research so far has focused on acoustic cues to prominence, where it was found that words can be prosodically highlighted by means of variation in pitch, duration, loudness, and voice quality (Cruttenden, 1986; Ladd, 1996). In more recent years, it has regularly been reported that such prominent words can also be marked by means of facial expressions, such as eyebrow movements, or by more exaggerated movements of the articulators (Cho & McQueen, 2005; Dohen, Lœvenbruck, Cathiard, & Schwartz, 2004; Graf, Cosatto, Strom, & Huang, 2002; Keating et al., 2003). In general, there have been claims that head movements and eyebrow movements are correlated with acoustic features of prosody, such as fundamental frequency and amplitude (e.g. Cavé et al., 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Accordingly, such visual markers have been implemented in animated synthetic characters as markers of important bits of information (Cassell, Vihjálmsson, & Bickmore, 2001; Pelachaud, Badler, & Steedman, 1996). However, while there is a long tradition on acoustic correlates of prominence, we still need a good deal of knowledge on the visual correlates. In particular, not many studies so far have reported on how visual cues to prominence are processed by observers, and how they relate to auditory markers (see, however, Beskow, Granström, & House, 2006; Granström, House, & Lundeberg, 1999; House, Beskow, & Granström, 2001). Therefore, the current study will concentrate on two questions regarding their contribution for the perception of prominence: (1) how important are visual cues to prominence from the face with respect to auditory cues? and (2) are there differences between different facial areas in their cue value for prosodic prominence? Let us elaborate on these two questions in the remainder of this Introduction.

The relative importance of facial cues with respect to auditory cues for signalling communicatively relevant information has been a research topic for a few decades, but most of that work has been limited to either McGurk effects or the relative contribution for signalling attitudinal or emotional correlates of speaker utterances. Data from the latter type of studies in particular have been used as evidence that visual information is far more important for communicative purposes than acoustic information (Dijkstra, Krahmer, & Swerts, 2006; Mehrabian & Ferris, 1967). However, these results do not necessarily imply that visual information is predominant for signalling other kinds of functionally relevant information as well, such as prominence. In particular, preliminary evidence so far suggests that observers extract more cue value from auditory features when it comes to marking prominent information in an utterance (Keating et al., 2003). This was confirmed by our own results from an earlier set of pilot studies, in which participants were presented with audiovisual versions of simple Dutch utterances like "blauw vierkant" (blue square), produced by a synthetic head. The utterances were varied such that they contained a pitch accent or a visual eyebrow marker on either the first or the second word. In a first functional study (Krahmer, Ruttkay, Swerts, & Wesselink, 2002), we found that people pay much more attention to auditory than to the eyebrow information when they have to determine which word in an utterance represented new information. Other follow-up studies confirmed the relatively weak cue value of these visual features, yet at the same time provided evidence that the visual cues do have some perceptual relevance (Swerts & Krahmer, 2004). A first perception study investigated the naturalness of various combinations of visual and auditory markers of prominence and revealed that observers tend to prefer these two to co-occur on the same word (congruent condition) rather than to be displaced on different words (incongruent). A second perception study brought to light that observers find that the prominence of a word is boosted if a pitch accent is additionally marked with a visual eyebrow movement, whereas the prominence of that same accent is downscaled if the visual marker occurs on a neighbouring word. In research using data coming from real speakers (Krahmer & Swerts, in press), participants were presented with utterances having a pitch accent and a facial prominence marker on one of its

words. These utterances were presented to observers either in an audio-only or audio-visual condition, which revealed that an accented word is rated to be more prominent when an observer could actually "see" a visual marker as well, compared to a condition where the observer could only hear the accented word.

So while all these studies on audiovisual cues to prominence perception show that visual markers do have some import for signalling prosodic prominence, it is still not clear how important these markers are compared to auditory cues. One drawback is that much evidence is based on the outcome of experiments with a synthetic Talking Head: to gain more insight into the cue value of eyebrow movements for the perception of prominence, many studies made use of an analysis-by-synthesis technique, creating stimuli whose visual properties were systematically varied to learn more about the relative effect of this parameter on focus perception (e.g. Beskow et al., 2006; Granström et al., 1999; House et al., 2001; Krahmer et al., 2002; Krahmer & Swerts, 2004; Swerts & Krahmer, 2004). While the implementations of the visual cues were inspired by claims in the literature, it would seem important to supplement such results with findings of observations on real speakers to see whether they indeed use visual markers for the determination of prominence. Moreover, most of the tasks used in the experiments discussed above on prominence perception were offline and consisted of elicited metalinguistic judgments of participants on the naturalness, prominence level or semantics of an utterance. This is different from many experimental studies in which speech processing is studied in a more online manner. For explorations of the cognitive effect of pitch accents, use is often made of a reaction time paradigm (Terken & Nooteboom, 1987) or eyetracking (Dahan, Tanenhaus, & Chambers, 2002) which allows for more direct measurement of the import of accents on speech processing. Terken and Nooteboom (1987) found that people's reaction times are longer when given information is accented or when new information is deaccented. So far, this experimental technique has not been used for studying facial correlates of prominent information. If eyebrow movements or other visual markers can perform a similar function as pitch accents, it is a reasonable hypothesis that a correct placement will enhance the listeners' interpretation, while incorrect placements may hinder it.

While it remains a general open question how relevant facial cues are compared to auditory markers, it also is not yet sufficiently clear whether different facial areas differ in their importance for signalling 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 physiological, 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; Pelachaud et al., 1996). Important cues may also be located in the mouth area of the face. Keating et al. (2003) found that some of their speakers produce prominent words with greater interlip distance and more chin displacement. Similarly, Erickson, Fujimura, and Pardo (1998) showed that the increased articulatory effort for realizing emphasized 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, whereas Dohen et al. (2004) report similar results for instances of contrastive focus. In addition, there is perceptual evidence that the upper and lower part of a speaker's face do not have equivalent cue value. It is obvious that observers primarily derive important phonological information from the mouth area (e.g. lipreading), though it has also been reported that people are sensitive to speech related head movements that extend beyond the mouth area, which can increase speech intelligibility (Davis & Kim, 2007; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Prosodic cues tend to be located in the upper part of the face: 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; see also Nicholls, Searle, & Bradshaw, 2004). Similarly, de Gelder, Vroomen, and Bertelson (1999) report that the visual information in the lower part of the face is less important for emotion perception than the visual information in the eye region. In sum, there are various types of evidence, both speaker- and observer-related, that show that the upper and lower part of a speaker's face are not equivalent in their cue value for signalling linguistic or paralinguistic information.

Intuitively, one might think that facial distinctions in the horizontal domain may not be crucial for prominence perception. Nevertheless, there are also indications that the left and right parts of a human's face differ in this respect. It is clear that faces are physiologically asymmetric 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. That there appear to be physiological differences between the left and right side of a speaker's face also appears from studies of orthodontics (Janzen, 1977). Directly related to prominence, there is empirical evidence from Keating et al. (2003) and Cavé et al. (1996), who report correlations between fundamental frequency and eyebrow movements, especially in the left eyebrow. Perceptually, Mertens, Siegmund, and Grüsser (1993) showed that participants 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 object like a vase. Thompson, Malmberg, Goodell, and Boring (2004) report findings of an experiment in which they had their participants 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 dot. This test revealed that the left side of a face was predominant from a perceptual point of view. Left-side dominance has also been reported for lipreading studies (Erber, 1974), gender judgements (Butler et al., 2004) and studies of portraited figures (Kowatari et al., 2004). In sum: given that there are both physiological and perceptual data to show that the left side of a speaker is different from his/her right side, both of these sources of evidence could be responsible for left–right differences in cues to prominence.

The overview of the studies presented above reveals that visual cues are potentially useful as markers of prominent information, yet it is still unclear how important they are compared to auditory cues. In addition, there are reasons to believe that different facial areas, both in the vertical and the horizontal dimension, are different in their possible cue value for marking prominence, but many questions regarding the exact contribution of these different areas are still unanswered. This article wants to give an answer to two related questions regarding the perceptual processing of audiovisual markers of prominence in spoken utterances: (1) how important are visual cues to prominence from the face with respect to auditory cues? and (2) are there differences between different facial areas in their cue value for prosodic prominence? The following sections describe two experiments we conducted to address these questions. The first perception experiment tackles the relation between auditory and visual cues by means of a reaction-time experiment. For this experiment, recordings of a sentence with three prominent words were systematically manipulated in such a way that auditory and visual cues to prominence were either congruent (occurring on the same word) or incongruent (in that the auditory and the visual cue were positioned on different words). The second experiment investigates which area of a speaker's face contains the strongest cues to prominence, using stimuli with either the entire face visible or only parts of it. The task of the participants was again to indicate for each stimulus which word they perceived as the most prominent one. We first present the audiovisual recordings which we used as a basis for creating the stimulus materials for both experiments, after which we discuss the two experiments themselves. We end this article with a general discussion about the implications of the various results for an audiovisual model of prosody perception.

2. Audiovisual recordings

As a basis for the two experiments described below, recordings were made of six native speakers of Dutch (four male, two female) between the ages of 20 and 40. Two of the six speakers were the authors, the other four were students, with no previous experience in audiovisual research. All the speakers were right handed. In order to remove any visually distracting features, speakers did not wear any remarkable clothes, and were asked to take off their glasses during the data collection procedure. They were instructed to utter different variants of the Dutch sentence "Maarten gaat maandag naar Mali" (*Maarten goes Monday to Mali*), which they had to produce in such a way that the first (Maarten), second (maandag) or third content word (Mali) of the sentence would be more prominent. This sentence, or slight variants of it, have been used before in research on the perception of prominence (e.g. Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997). There were various reasons why we limited the recordings to only one sentence. First, since the recordings would primarily be used as a basis for an observer-oriented rather than a speaker-oriented study (see experiments below), we wanted to optimize the experimental conditions for a perceptual study. Therefore, to make sure that participants in our experiments could focus as much as possible on the audiovisual correlates of

prominence, we avoided introducing lexico-syntactic variation which could distract our subjects from this primary goal. Second, the advantage of having exemplars of identical utterances facilitated the audiovisual manipulations to be discussed below, which consisted of creating artificial stimuli which consisted of various combinations of movies and sounds.

During the recording sessions, speakers were not given any instruction on how prosodic prominence should be realized in audiovisual speech, but they were told that they had to imagine that the target utterances were answers to various kinds of questions (''Who will go on Monday to Mali?''; ''When will Maarten go to Mali?''; ''Where will Maarten go to on Monday?''). The 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 to the aforementioned conditions, speakers were asked to utter sentences in a monotone, so without any auditory or visual markers of a prominent word, which were used for our second experiment discussed below. Fig. 1 presents two stills of one of our speakers (MS), taken from the middle part of an unaccented and accented syllable in a target word (producing the vowel /a/). As is already observable from this figure, the accented syllable appears to be produced with a greater articulatory movement, and is accompanied with some eyebrow movement.

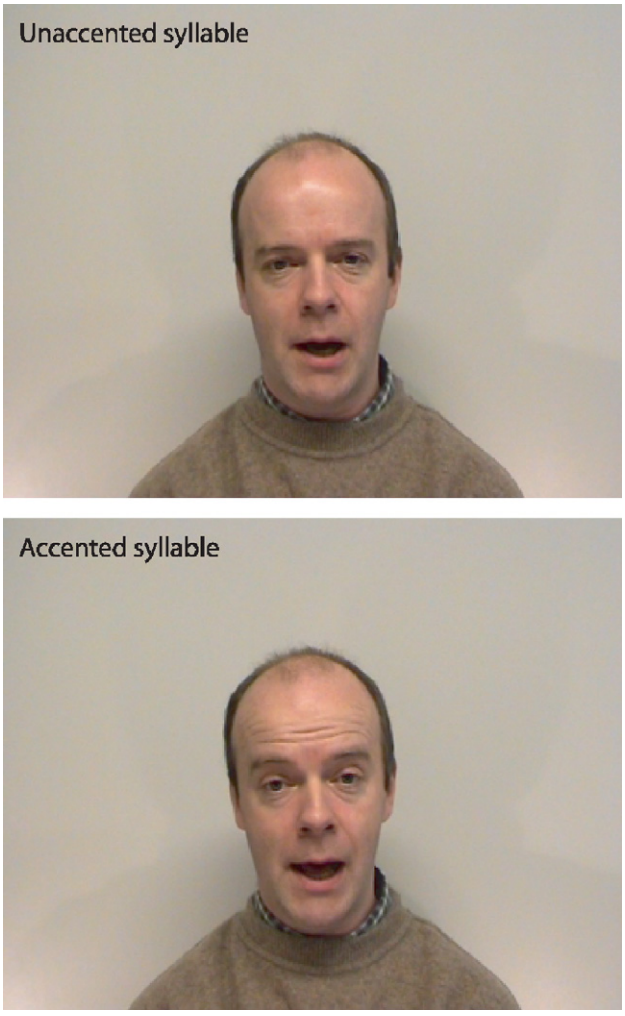

Fig. 1. Representative stills of a facial expression of one of our speakers while producing an unaccented (top) or accented (bottom) syllable in one of our target words.

The actual recordings were organized in different blocks of four sentence productions, in which a speaker was first asked to utter the sentence in a monotone, and then the three realizations with an prominence marking of the first, second or third target word. This whole procedure was repeated twice. The audiovisual recordings of all six speakers were made in a quiet 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 m 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 two spotlights next to the camera focused on the floor in order to minimize reflections. After the recording session, we presented all the utterances in audio-only format to two independent judges (unaware of the general purpose of our study) who were asked to indicate whether the sentences had an accent on W1, W2, or W3, or whether they did not contain any accent at all. These checks revealed that the accent distributions on the sentences were as intended, and that the monotonous sentences were indeed devoid from any audible markers of prominence. As a matter of fact, there was a 100% agreement between labelers on the presence or absence of clear auditory accents; this degree of consensus may be atypical in view of other prosodic labelling studies, but was due to the fact that the prosodic structures in our stimulus materials were very stylized.

These audiovisual recordings were used as a basis for the stimulus preparations of our two perception experiments. While the visual variation is identical in the two experiments, the auditory information is different, in that we use the versions with auditory markers of a prominent word for experiment 1 and without auditory markers (monotone renditions) for experiment 2, for reasons explained below.[1]

3. Experiment 1

3.1. Method

3.1.1. Stimulus preparations

The audiovisual recordings of the different utterances produced by our six speakers were manipulated with Adobe Premiere™ to obtain all the stimulus variants. First, the sound and video recordings were separated, after which these two modalities were combined again such that the video and audio channel always came from different recordings. In this way, we constructed two sets of stimuli. The first set contained so-called congruent utterances, i.e. utterances in which the auditory and visual markers of prominence occurred on the same word. The second set consisted of incongruent stimuli in which the auditory and visual markers were associated with different words, for instance, a visual marker on the third word and an auditory marker on the first or second one. Using a trial and error procedure, we chose the best matches of movie and speech as our stimuli for the following experiments, that is, the most synchronous combinations of movie and sound. Note that we decided to make use of artificial combinations for our experiment for both the incongruent and congruent conditions, to make the stimuli more comparable. That is, the congruent stimuli were also created from audio and video tracks from different exemplars of the same sentence. In this way, our participants in their perceptual judgments could not make use of the fact that some stimuli were artificial, and others were not. All the manipulations led to a total of 54 stimuli (three auditory markers, three visual markers, six speakers). Since only one sentence was used for all recordings, it turned out to be very easy to combine speech and movie, and the naturalness of the artificial stimuli was extremely good. This is in itself not surprising given findings that observers are perceptually very tolerant towards asynchronies between the auditory channel and visual information from the face. There is even evidence that sentence intelligibility scores in audiovisual condition of sentences with asynchronies between the two channels of up to 200 ms outperform those for audio-only conditions (Sakamoto, Tanaka, Tsumura, & Suzuki, 2007). An informal inspection of the data did not reveal cases of undesired lipsync problems, for instance leading to possible unwanted McGurk effects. To confirm these impressions, we asked a panel of two independent judges to check all the stimuli in terms of whether they felt the auditory or the visual signal was lagging behind, or whether the stimuli were completely synchronous. This additional check did not reveal any problematic cases of audiovisual mismatches.

[1]We invite the interested reader to visit the following website http://foap.uvt.nl/phonetics to view videos from speakers MS and EK that are representative for the stimuli used in the two experiments of this study.

3.1.2. Participants

Forty-two participants (18 male, 24 female) in total participated in this experiment on a voluntary basis, most of them students and colleagues at Tilburg university. The average age of the participants was 27.7 (youngest: 21, oldest: 50). They were all right handed, and had normal or corrected to normal vision and good hearing. All were naive to the experimental question, and none of them had been a speaker in the recording session.

3.1.3. Procedure

The stimulus materials were presented in one of four randomized orders to participants in an individually performed experiment. Participants saw clips of the speakers on a Philips True Color PC screen (107 T 17″) of 1024 by 768 pixels, and sound was played to them through loudspeakers located left and right of the computer screen. Stimuli were played using the Pamar software developed at the Psychology department of Tilburg University, which allows measurements of reaction times with audiovisual stimuli, and which has an error of maximally 25 ms, corresponding to the frequency with which the computer reads information from the keyboard. The participants were instructed to click on one of three buttons on their keyboard, marked with the numbers 1, 2 and 3, to indicate whether they had perceived the first, second or third word as being most prominent. Since the prominence ratings are relative judgments, they were told to click on the chosen button as soon as they decided what the most prominent word was, but in order to do so, they knew they had to listen to all three target words. The reaction times are measured with respect to the moment that a speaker finished uttering W3, which was entirely based on acoustic cues through auditory inspection. Thus, a reaction time of 0 means that a participant has clicked a button at exactly the same moment that a speaker finished uttering W3; a negative reaction time means that a participant has clicked before the end of the utterance, for instance because a participant has made a decision after hearing the /ma/ syllable in W3. The inter-stimulus interval was 500 ms, in which time frame participants had to respond. Note that we did not include a neutral option (''none of the words are prominent'') as a possible response category in our experiment to avoid that subjects would use this option too quickly in cases of minor or major doubt.

In addition, participants were told beforehand that after the test they would be asked to fill out a small questionnaire, in which they would have to answer a number of questions regarding the speakers who had been shown in the experiment. The participants were informed that the questions would refer to certain visual features of the speakers, such as gender or characteristics of their clothes. Participants were told that the person with most correct answers in the questionnaire would receive a book token. The reason to have this secondary task was to make sure that participants would always focus on the screen, and not for instance close their eyes to concentrate on the auditory signal alone. In theory, there is a slight danger that this procedure may have led participants to focus more on visual cues relevant to a speakers' gender or cloths, rather than on prominence cues, but we judged this problem to be of minor concern.

The actual experiment was preceded with a short practice test with six congruent stimuli, in order to make participants acquainted with the kinds of stimuli and the general experimental procedure. During the practice test, no feedback was given to the participants about the ''correctness'' of their responses. If there were no questions from the participants about the experimental set-up after the practice test, they could go on with the actual experiment in which it was no longer possible to communicate with the experimenter. The whole procedure, including practice test and questionnaire, took approximately 10 min per subject, of which about 8 min were used for the main experiment.

3.2. Results

The first experiment has a complete $3 \times 3 \times 6$ design with the following within-subject factors: auditory marker of prominence (three levels: prominence on W1, prominence on W2, prominence on W3), visual marker of prominence (three levels: prominence on W1, prominence on W2, prominence on W3), and speaker (six levels). (Order of stimulus presentation turned out not to be significant, and was not included in remaining analyses.) The data were first checked for the occurrence of possible outliers. Of a total of 2268 datapoints (54 sentences $\times$42 listeners), 38 cases were treated as outliers, i.e. those cases where the reaction times were at a distance of at least three standard deviations from the overall mean. The majority of these typically consisted

Table 1
Overview of perceived prominences for various combinations of auditory and visual markers to prominence on Word 1 (W1), Word 2 (W2) and Word 3 (W3)

Prominence		Chosen prominence		
Auditory	Visual	W1	W2	W3
W1	W1	247	4	1
	W2	226	26	0
	W3	235	3	14
W2	W1	17	233	2
	W2	1	248	3
	W3	8	233	11
W3	W1	44	3	205
	W2	13	58	181
	W3	3	2	247

Each row total is 252.

of cases in which a subject had produced very negative reaction times, basically meaning that they had responded a considerable time before the end of the utterance. Interestingly, 20 of those 38 outliers came from stimuli produced by speaker LL, who appeared to be the most visually expressive speaker of all. Outliers were then replaced with the overall average reaction time. No further manipulations of reaction times were performed.

First, Table 1 reveals which word (W1, W2, or W3) participants had chosen to be the most prominent one, as a function of various positions of an auditory and visual marker of prominence. Table 1 reveals that participants mostly designate that word in an utterance as being the most prominent one which also carries the auditory maker of prominence. Interestingly, that preference is stronger for cases where the chosen word also gets a visual marker: in other words, the congruent stimuli reveal a stronger preference for the auditory marker than the incongruent ones. Note that most confusion arises for cases where the auditory cue is positioned on W3, in line with earlier observations that later accents in an utterance are less salient (Krahmer & Swerts, 2001).

To get a first insight into the patterns of the reaction times, we conducted a *t*-test which compared averages, calculated per participant, for congruent and incongruent stimuli. Thus, for this test, we combined the two observations for incongruent stimuli and paired those to that for the congruent stimulus per participant. This *t*-test reveals that congruent stimuli differ significantly from incongruent ones in that the latter give consistently slower reaction times (congruent: 73 ms; incongruent: 150 ms) ($t_{(41)} = 4952$, $p < .001$). In addition, when we compared cases, again based on averages per participant, in which a participant's response matched with the position of the auditory marker of prominence with cases where there was a mismatch between these two, then it turns out that the matching conditions led to significantly faster reaction times (match: 84 ms; mismatch: 325 ms; $t_{(40)} = 3802$, $p < .001$).[2] However, since only a minority of 211 stimuli out of the total led to such mismatches, we collapsed these response times with those for the matching ones for subsequent analyses.

A three-way analysis of variance for repeated measures was performed with the aforementioned within-subject variables as independent factors and with the reaction times (in milliseconds) as dependent variable. Mauchley's test[3] was used to check the homogeneity of variance, and the Bonferroni correction was used for multiple pairwise comparisons. Means are displayed in Table 2. Main effects were found of auditory marker of prominence ($F(2, 82) = 20.523$, $p < .001$, $\eta_p^2 = .334$), visual marker of prominence

[2]There was one participant whose responses always matched with the position of the auditory accent; his data were therefore not included in this *t*-test.

[3]As a matter of fact, except for the two-way interaction between auditory and visual markers, Mauchley's test for sphericity was significant for all main effects and other interactions. For these cases, we looked both at Greenhouse–Geisser and Huynh–Feldt corrections on the degrees of freedom, which gave similar results. For the sake of transparency, we report on the normal degrees of freedom.

Table 2
Average reaction times and standard deviations (in ms): main effects

Factor	Level	RT (S.d.)
Auditory prominence	W1	34 (62)
	W2	106 (55)
	W3	232 (44)
Visual prominence	W1	100 (55)
	W2	172 (54)
	W3	100 (55)
Speaker	EK	9 (56)
	LL	265 (65)
	MB	190 (47)
	ME	108 (48)
	MS	121 (50)
	PB	53 (57)

Table 3
Average reaction times (in ms) for various combinations of auditory and visual markers of prominence

Prominence		RT (in ms)						
Auditory	Visual	EK	LL	MB	ME	MS	PB	Average (S.d.)
W1	W1	−37	**240**	125	**−72**	**−12**	**−247**	**−19** (52)
	W2	−87	257	**35**	−10	49	66	52 (65)
	W3	**−105**	253	42	133	73	25	70 (74)
W2	W1	**−190**	**52**	149	311	74	263	**63** (63)
	W2	162	314	**63**	**89**	**34**	129	132 (53)
	W3	−92	141	231	233	126	**103**	124 (53)
W3	W1	172	**289**	465	212	288	115	257 (45)
	W2	294	496	386	399	335	91	333 (37)
	W3	**−38**	346	**210**	**−46**	**238**	**−69**	**107** (49)

Results broken down per speaker, and overall average (and standard deviation). The fastest response times for each level of auditory markers (W1, W2, W3) are in boldface.

($F(2,82) = 7.356, p < .01, \eta_p^2 = .152$) and speaker ($F(5,205) = 14.141, p < .001, \eta_p^2 = .256$). For auditory markers, all pairwise comparisons turned out to be significant: reaction times become increasingly slower for auditory markers later in the sentence. This is in itself a logical result as the marker for W3 occurs by its nature late in the sentence, so that participants have less time to process compared to the earlier markers for W1 and W2. Regarding visual markers, it appears that the reaction times on W2 words are significantly slower than the other two, whereas W1 and W3 do not differ from each other. It also turns out that speakers differ from each other in yielding slower or faster reaction times. In addition, the ANOVA gave a significant two-way interaction between auditory and visual markers ($F(4,164) = 10.362, p < .001, \eta_p^2 = .201$). This interaction can be explained by looking at Table 3, which displays average reaction times as a function of different combinations of auditory and visual markers: as can be seen, for W1 and W3 words (i.e. words at the edges of an utterance), it appears that congruent stimuli where visual and auditory markers co-occur on the same word, lead to faster reaction times than the incongruent stimuli, whereas in W2 words (the middle word in the utterance) the congruent stimuli are very similar to the incongruent ones. The ANOVA also gives significant two- and three-way interactions when speaker is combined with the other factors: Table 3 reveals that the congruent cases in W1 and W3 for stimuli from different speakers—with a few exceptions—lead

tocomparatively faster reaction times (the numbers printed in bold), but the average speed of these reactions is variable between speakers.

3.3. Discussion

The current experiment brought to light that visual cues have an impact on how prominence is perceived, albeit that the visual markers appear to be not as strong as the auditory markers. While participants tend to focus on auditory cues (for the first and final word in an utterance), they cannot ignore the visual markers: congruent stimuli lead to faster reaction times than incongruent ones. In this respect, it thus turns out that visual markers of prominence (such as eyebrow movements, head nods, or the velocity and amplitude of articulatory movements) can perform a similar function as pitch accents, confirming the expectation that a correct placement will enhance the listeners' processing of incoming speech, while incorrect placements may hinder it. Note, however, that this general effect interacted with a positional constraint: the impact of visual cues on processing time was only apparent if the auditory marker occurred on the first or last word of the sentence, while it disappeared for accents in medial positions. One could argue that this might be due to the fact that, in many languages, sentence edges represent important positions in an utterance, as they are often reserved for functionally important discourse information (e.g. Dik, 1978). Therefore, listeners may have a natural bias to focus on these positions when it comes to prominence detection, whereas they are less sensitive for middle positions. However, at least for French, several perception tests by Dohen (2005) on visual and audiovisual perception of contrastive focus in subject–verb–object sentences was never better for the object (end of utterance) than for the verb (middle of sentence). It is unclear at this stage whether the differences between the results from Dohen and those from the current study are due to linguistic differences between Dutch and French, or to the fact that different experimental paradigms were used to measure the processing of audiovisual cues to prominence. More research is needed to clarify this.

While experiment 1 thus showed that facial expressions matter in prominence detection, it remains to be seen which aspects of a face are more important for signalling prominence. The second experiment therefore focuses in more detail on the relative importance of different facial areas. In particular, we zoom in on differences both in the vertical and horizontal domain. The former distinguishes between a top and bottom part of the face, roughly coinciding with the areas around the eyes and the mouth, respectively. The latter dimension is concerned with a left–right distinction. These issues are addressed in experiment 2.

4. Experiment 2

4.1. Method

4.1.1. Stimulus preparations

The stimuli used for this experiment are again based on the audiovisual recordings described in Section 2. However, given that the current test is intended to learn more about the relative cue value of different facial areas, we no longer included auditory markers of prominence in our design. Therefore, as a basis for our stimulus preparations, we only made use of the monotone renditions of the utterances. Our procedure consists of three kinds of manipulations. The first one was similar to the one in our previous experiment, and consists of mixing the monotone realization of the utterances with the different visual realizations by our six speakers. In other words, in the current experiment, the auditory information was always identical for all the stimuli per speaker. Besides the original, we produced four additional versions from the video-recording of the full face, by blackening parts of the face, again using Adobe Premiere™ 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. Those black quadrants were of the whole image, not of the face per se. The size of the quadrants was slightly different for the different speakers, as its position was dependent on the size and the position of a speaker's head in a movie. Fig. 2 gives some representative stills from one of our speakers (EK).

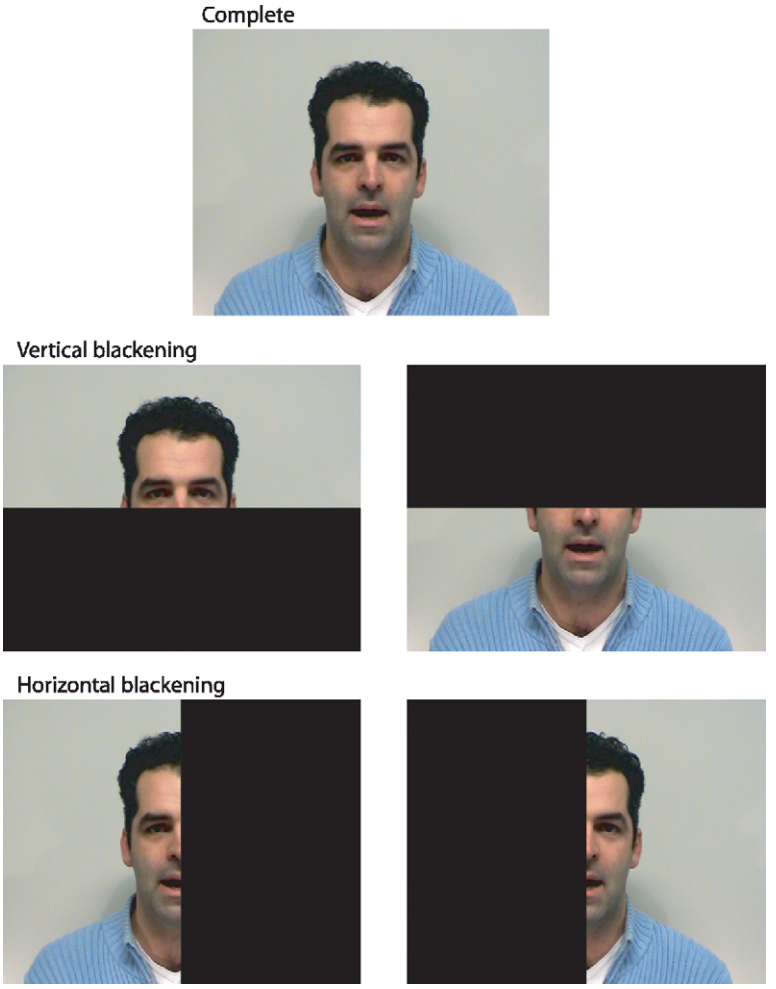

Fig. 2. Different stills which represent different versions of our stimuli as presented in experiment 2, in which the face of our speaker is either completely or partly visible.

Note that the blackened part of the screen did not move with the speaker, though this was not problematic since the speakers did not make extreme vertical or horizontal head movements.

After having created these different versions, we made mirror images of all five versions of these stimuli. Thus, in the mirror-condition, the speaker's left appears on the observer's left, as opposed to on the observer's right as would occur in the normal, non-mirrored presentation. Fig. 3 illustrates an original image together with its mirror.

All the manipulations led to a total of 180 stimuli: visual marker of prominence (three levels: prominence on W1, prominence on W2, prominence on W3), speaker (six levels), facial area (five levels: complete face, upper part visible, bottom part visible, left area visible, right area visible) and display (two levels: original, mirrored). Again, due to the uniformity of the words in the target sentence, audiovisual alignment was very good, and did not give rise to undesired side effects, as confirmed in checks with 2 independent judges (see procedure in experiment 1).

4.1.2. Participants

There were 66 participants (36 male, 30 female) who took part in this experiment on a voluntary basis, again students and colleagues from Tilburg University and other academic institutions nearby, and again all participants were naive to the experimental question. The average age of the participants was 25.5 years old,

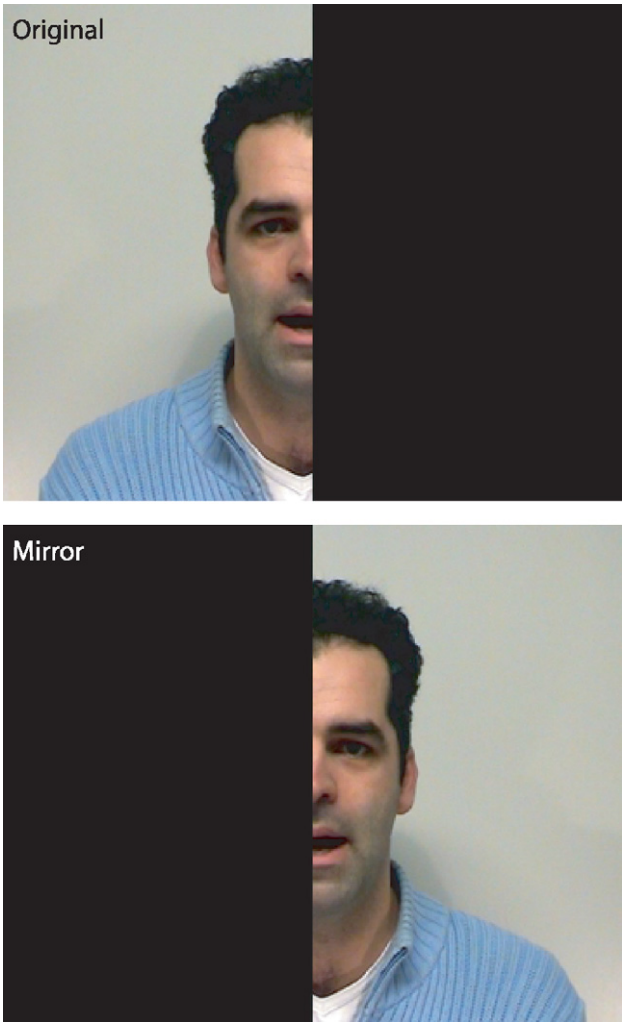

Fig. 3. Two representative stills of a facial expression presented in original or mirrored condition.

and they all had normal or corrected to normal vision and good hearing. None of the participants of experiment 2 had participated in experiment 1, and none had served as speaker in the recording session. We did not check whether or not a participant was right handed.

4.1.3. Procedure

The task was similar to that of our previous experiment, i.e. to indicate which word (W1, W2, or W3) was the most prominent one in a stimulus utterance, except that this time the experiment was a paper-and-pencil test and participants were not requested to react as fast as possible. Participants were also told that the person with the greatest accuracy in detecting prominent words 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 among 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 185×165 pixels, corresponding to roughly 4.8×4.3 cm. In addition, we added the distance from the screen as a between-subjects factor (see also Jordan & Sergeant, 2000 for a similar procedure), in the sense that one third of the participants had to do the experiment at a "normal" distance from the screen (approximately 50 cm from the screen), in the middle condition participants were positioned at 250 cm from the screen, and in the far

Table 4
Distribution of participants' chosen prominences for different visual prominences

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

Differences in row totals are due to the fact that outliers were left out of the calculations.

condition at 380 cm 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 × 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 s, in which time frame participants had to indicate in a multiple-choice on an answer sheet whether they thought the first, second or third target word was the most prominent one (forced choice). All stimuli were only presented once. Half of the participants saw the original stimuli, and half of them saw their mirror versions. The mirror/original condition was a between-subject factor in order to reduce experimental time, and to avoid that participants would notice the manipulations. The actual experiment was again preceded by a short test phase (with no feedback from the experimenters) to make participants acquainted with the general set-up. The experiment, including instructions and test phase, lasted about 20 min per subject.

4.2. Results

The second experiment has a complete 3 × 6 × 5 × 2 × 3 design with the following factors: visual marker of prominence (three levels: prominence on W1, prominence on W2, prominence on W3), speaker (six levels), facial area (five levels: complete face, upper part visible, bottom part visible, left area visible, right area visible), display (two levels: original, mirrored) and distance (three levels: close, middle, far). Table 4 gives a first overall impression of how the responses are distributed for various positions of the visual markers. As can be seen from the numbers on the diagonal in the confusion matrix, participants tend to perceive the word which receives the visual marker as being the most prominent one.

The data were analysed with a logistic regression with the aforementioned variables as independent factors, and the participants' perceived prominence scores as dependent variable. Scores were represented as a binary variable, either as correct (the response is identical to the position of the visual marker) or incorrect. To gain insight in the added value of main effects and interactions on the explained variance, we ran separate models with only main effects, and models which include estimates for interactions as well. A customized model which only tests main effects revealed significant effects for visual marker of prominence ($\chi^2 = 9.537, \mathrm{df} = 1, p < .01$), facial area ($\chi^2 = 319.441, \mathrm{df} = 4, p < .001$), speaker ($\chi^2 = 176.433, \mathrm{df} = 5, p < .001$) and distance ($\chi^2 = 681.051$, $\mathrm{df} = 2, p < .001$), while the effect of display was not significant. This model accounts for 24% of the variance. Table 5 reveals that initial markers of prominence are most often detected correctly, whereas detection is poorer for markers in middle and last sentence position, which is statistically confirmed from a pairwise comparison of the parameter estimates. With respect to the effect of facial area, we see that a whole face presentation leads to the best prominence detection, whereas displays of the upper and left part of the face lead to significantly better results than displays of 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 5 also shows that stimuli from different speakers lead to markedly different results, with relatively poor detection for stimuli from speaker MB and best results for speaker PB.

A model which also includes two-way interactions between all factors presented above revealed significant interactions of facial area with speaker ($\chi^2 = 72.347, \mathrm{df} = 20, p < .001$), with distance ($\chi^2 = 31.620, \mathrm{df} = 8$, $p < .001$), with visual prominence ($\chi^2 = 16.562, \mathrm{df} = 8, p < .05$), and with display ($\chi^2 = 36.533, \mathrm{df} = 4, p < .001$).

Table 5
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

In addition there were two more significant interactions between visual prominence and speaker ($\chi^2 = 230.116, \mathrm{df} = 10, p < .001$), and between visual prominence and distance ($\chi^2 = 14.140, \mathrm{df} = 4, p < .01$), with all the other interactions not being significant. This model with the two-way interactions included could explain 32% of the variance. The interactions in which the factor speaker is involved can possibly be related to speaker-specific variation in expressiveness: first, while all speakers exhibit the same pattern of the main effect of visual display, for some speakers the differences between conditions are larger than for others; second, there are differences between speakers as to which markers of prominence in an utterance (W1, W2, W3) gets the highest proportion of correct scores. The interaction between visual prominence and distance is due to the fact that the differences in scores for W1, W2 and W3 become bigger when distance increases: whereas the prominence scores for the three words are about the same in the close and the middle conditions, the scores for W1 are markedly higher (56.6%) than for W2 (47.1%) and W3 (48.5%) in the far condition. Similarly, the differences between facial conditions become bigger at a larger distance, which explains the interaction between facial area and distance (the prominence scores for different conditions are most dissimilar in the far condition). The most intriguing interaction is that between facial area and display, as it turns out that display (original view or mirrored view) does not have an effect when faces are shown in full or with the vertical manipulations, whereas display does matter for faces that are horizontally manipulated: the original (i.e. speaker's) left side always gets higher correct scores than the original right side, but when the left side is shown in mirror image the scores get lower, while the reverse is true for the case in which the original right side is displayed as the left side.

To get more insight into the latter result, we 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 reveal a significant interaction only between facial area and display for horizontally blackened stimuli ($\chi^2 = 20.472, \mathrm{df} = 8, p < .001$), but not for whole face stimuli, or for stimuli manipulated in the vertical domain (both $p > .1$). This can be explained using the data given in Table 6 which reveals that the scores for prominence 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 from those 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

Table 6
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

original display, we observe that prominence detection is 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 original left side is shown as the right side, while the reverse is true for the original right side becoming left side.

4.3. 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 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. It is a speaker effect since 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. However, that left side is perceived as being less prominent when it is shown as a speaker's right side, which appears to be related to an observer effect, as the observer, when making prominence judgments, tends to be biased to the side of a face that occurs in his or her left field of vision. The reverse effects are true for the speakers' right side of a face, whether shown in original or mirrored display.

5. General discussion

This study has presented the results of two experiments on the perceptual processing of visual markers of prominence, i.e. words that 'stand out' with respect to other words in a spoken utterance. Experiment 1 (a reaction-time experiment) was concerned with the general question how important visual markers in a speaker's face (such as eyebrow movements or more pronounced movements of the articulators) are with respect to auditory markers (prosodic cues such as pitch, duration and loudness), which traditionally have received much more scholarly attention than the former. Experiment 2 (a classification experiment) investigated whether different facial areas (both in the vertical and the horizontal domain) differ in their cue

value for signalling prominence. Let us discuss the main findings of these two experiments in view of the existing literature on the processing of faces in general, and of prosodic prominence in particular.

Experiment 1 presented evidence that visual cues to prominent information do have an effect on the speed with which prominent words can be detected in an utterance, albeit that they are less important than the auditory cues. So while this confirms earlier observations that auditory prosodic cues are more important than visual cues for the perception of prominence, it also makes clear that visual cues have some import, as was also already clear from previous metalinguistic judgments tasks on the naturalness and perceived prominence of audiovisually produced prominences in utterances generated by a synthetic head (Krahmer & Swerts, 2004). The general effect is also consistent with results of others who presented observers with audiovisual stimuli which were either congruent or incongruent regarding the use of auditory or visual cues to communicatively important information. For instance, Pourtois, Debatisse, Despland, and de Gelder (2002) showed that listeners find it more difficult to process words spoken with a certain emotional tone (e.g. happy), when they are simultaneously looking at a face that expresses an incongruent emotion (e.g. sad). Similarly, stimuli that are inconsistent regarding their use of visual and auditory cues to prominence are more difficult to process than stimuli where the two types of cues do match.

Experiment 1 also brings to light that the relationship between auditory and visual cues, and especially the relative cue strength of these two modalities for signalling certain aspects of communication, is a nuanced one. Previous studies have stressed the predominance of visual information for highlighting paralinguistic information, such as attitudinal and emotional correlates of particular utterances. This has led people like Mehrabian and others to maintain that visually observable variation from the face can account for more than 90% of the emotional content of a message (Mehrabian & Ferris, 1967). Subsequent empirical research has often provided support for the predominance of visual signals for cuing emotion (e.g. Hess, Kappas, & Scherer, 1988; Walker & Grolnick, 1983). (See also Massaro & Egan, 1996; Srinivasan & Massaro, 2003 for discussion about the relative cue value of auditory and visual features.) However, this finding does not necessarily generalize to all types of functionally relevant elements of spoken interaction, as is clear from the current study on prominence perception. In retrospect, this may explain why a vast majority of prior studies on emotion, beginning with the early seminal work by Darwin, have very much concentrated on facial displays of emotion (although most of that work was restricted to analyses of still images), whereas people dealing with correlates of prominence often have exclusively restricted their analyses to prosodic cues in speech-only stimuli (loudness, pitch, duration, spectral features).

Also, note that there is an important difference between the results of experiment 1 and those reported earlier on the McGurk effect. The latter relates to the observation by McGurk and MacDonald (1976) that the display of an auditory /ba/ paired simultaneously with a silent movie of someone producing the syllable /ga/ often produces the percept /da/. This outcome provides evidence that information from an auditory and visual source are integrated at one point during the perceptual process to form one coherent percept. In contrast, the results of the prominence experiment do not point out that multisensory information about prominence is integrated in the same way as in typical McGurk studies. In the latter, the perceived sound may be a compromise between conflicting visual and auditory cues, as observers perceive a sound (e.g. /da/) which is different from both the auditory (e.g. /ba/) and visual (e.g. /ga/) signal. In the prominence decisions reported here, however, the observers choose for either the visual or the auditory cue, where the auditory cues are clearly predominant. Note, however, that in the current experiment, participants were forced to choose between three options, whereas the McGurk data consists of free responses. The fact, however, that the prominence decisions slow down in the case of incongruent stimuli, does show that, similarly to McGurk effects, both visual and auditory cues are weighted during prominence decisions, and integration is more difficult when the two cues do not match (cf. Massaro, 1998, 2002; Massaro, Cohen, & Smeele, 1996).

Experiment 2 revealed that facial areas are not equivalent in their cue value for signalling prominence. When we look at the face from a vertical axis, our data reveal that the top part of the face has more cue value than the bottom part. This finding is in line with earlier claims by Lansing and McConkie (1999) that people 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). It is also in agreement with work by de Gelder et al. (1999) who report that judgments of paralinguistic information are easier when observers are exposed to the upper part of the face rather than the lower part. However, at first sight, our results seem inconsistent with findings

by Keating et al. (2003) who studied three male American speakers who, in addition to speaking words with different lexical stresses, produced sentences that differed in phrasal stress. Using small reflective dots that were attached to the speakers' faces, a number of articulatory measures was obtained for various facial areas, such as displacement of left eyebrow, head, lip and chin. They found that all their measures distinguished stressed from unstressed words, but that there was also some speaker variation; a perceptual study revealed that visual perceivers could most easily recover information about phrasal stress from larger and faster mouth opening movements, more open mouth positions, and head movements. More research is needed to find out why their and our study are at variance regarding the relative cue value of information from the upper part of the face, such as variation in eyebrow movement. In general, it appears that their production measures were especially focused on the mouth area, which was modelled using 17 dots, whereas the top part was represented by only two dots.

With respect to the horizontal variation, which a priori might seem less relevant for prominence perception, we found that the left side of a speaker's face has stronger cue value for prominence marking than his/her right side. It appears that this effect was due to a combined speaker and observer effect. Inspection of the literature reveals a left dominance of the face, both from a speaker- and observer-related perspective. There is some speaker-related evidence from studies on emotional expression that different facial areas differ in expressiveness, though results are not always entirely consistent. Moreover, one has to be cautious to interpret results from studies on emotion, as our own study was dealing with prominence, which may be processed quite differently in production and perception than paralinguistic information (see below). Borod, Koff, Yecker, Santschi, and Schmidt (1998) report that most studies on emotional expression reveal that the left half of the speaker's face (which has greater connectivity to the right cerebral hemisphere) is more intense or moves more extensively than the right half during facial expression of emotion. However, while this left dominance has repeatedly been reported for negative emotions, there is some evidence that positive affect tends to be associated with greater activity in the right region of the face (Richardson, Bowers, Bauer, Heilman, & Leonard, 2000). More directly related to our current study, Cavé et al. (1996) report that the left eyebrow more strongly correlates with intonation patterns than does the right eyebrow, although the analyses in that study were still in a preliminary stage of exploration. Given that pitch has been claimed to be one of the primary indicators of prosodic prominence in spoken utterances, it would seem natural to expect that the left eyebrow is comparatively more relevant for the expression of prominence than the right eyebrow.

The perceptual dominance of the left side of the face has been demonstrated repeatedly, both for static and dynamic images. Kowatari et al. (2004) present an overview of studies that show that a person's left side is depicted more often than the right side in portraits, and that reaction times in face recognition are shorter when the left side of a face is presented than the right side. In their own study using functional magnetic resonance imaging (fMRI), they found that photographs of left 3/4 view of a face elicit stronger neural responses (in comparison with right 3/4 views) in areas of the brain that are known to be involved in face recognition, where there is a right hemisphere bias. These results are consistent with the outcome of an investigation by Butler et al. (2004) who explored eye-movement patterns in a study of gender decisions for which they used chimeric images (stimuli in which male and female stimuli are blended into a complete face). They found that, when viewers have to determine the speaker's sex of such pictures, they more consistently used information from the left side of the face (see also Mertens et al., 1993). While the previous studies were based on processing of static images (photographs), Thompson et al. (2004) investigated spatial attention across a talker's face during auditory-visual speech discourse processing (movie clips). The participants' task was to detect dots that were superimposed onto a talker's face for 17 ms. Results reveal that dot detection performance was greater for the talker's left compared to their right side.

It seems relevant to compare previous and current findings of lateralized processing of facial cues to prominence with the literature on neural correlates of prosody. Previous studies have revealed hemispheric differences between the processing of different levels of prosody. Melodic and emotional aspects of prosody have often been argued to be more lateralized to the right, whereas the linguistic aspects of prosody are processed more to the left or in both hemispheres (Baum, Pell, Leonard, & Gordon, 1997; Borod, Andelma, Oble, Tweed, & Welkowit, 1992; Walker, Daigle, & Buzzard, 2002). Similarly, several experiments reported by Joanette, Goulet, and Hannequin (1990) have confirmed that rapid and local processing of acoustic information seems to be carried out by the left hemisphere, whereas processing over longer stretches of

syllables appears to be controlled by the right hemisphere. The fact that we find that cues to prominence in the left part of a speaker's face (which is in the observer's right visual field, and therefore processed by the observer's left hemisphere) are more important from a perceptual perspective, may therefore be used as indirect evidence that facial markers of prominence are processed as other types of linguistic (rather than paralinguistic or emotional) information. Obviously, more research is needed to confirm or disconfirm this claim.

We see different ways to pursue this research. First, the analyses presented in this article were based on data from six speakers. While our primary interest was to gain insight into the perceptual processing of audiovisual features, it is interesting to see that the participants' judgments varied as a function of the speaker presented. This did not seem to be related to the fact that two speakers were the authors while the other four were completely naive to the experimental question. Rather, the effects seemed more due to the fact that speakers differ in their degree of expressiveness. Our judgments so far on speaker expressiveness were based on our own impressions, and not supported by quantitative analyses of facial markers, as in the work by Dohen (2005) or Keating et al. (2003). It would be very useful to conduct such measurements on the kinds of data we have gathered for our two experiments. One of the reasons we did not include visible markers on our speakers' faces was because these would make the recordings less suitable for a perception study, which was our primary goal. In the future, we plan to replicate our data with more natural utterances, with different speakers, to see to what extent our first results have general validity. While our prominence judgment tasks may have been a bit metalinguistic in nature, there is evidence from speaker studies that eyebrow movements indeed occur on stressed vowels (Keating et al., 2003), so that such variation is likely to be used in speech perception as well. Second, we have seen that our first experiment gave clear processing differences in terms of reaction times for words that occurred in sentence-initial or final position (resp. W1 and W3), whereas words in the middle of the sentence (W2) did not show any effect of visual cues. We hypothesized that this could be due to an observer's bias for sentence positions that have been shown to be functionally marked, even though there is conflicting evidence (albeit for French) by Dohen (2005). However, it is possible that the effect could also be due to syntactic or semantic factors. This could be investigated with other stimulus materials with different lexico-syntactic structures. Third, we have limited the research to a study of facial cues. It could be useful to extend the research to include other potentially useful bodily markers such as hand and arm movements, which have also been shown to serve as beat gestures (Krahmer & Swerts, in press). Finally, in order to determine more exactly to what extent the prominence ratings are due to a speaker or observer effect, we intend to perform a follow-up study in which we measure participants' eye gaze behaviour to learn more about which facial areas are dominant for this task.

Acknowledgements

This research was conducted in the context of the FOAP project, which is funded by the Netherlands Organization of Scientific Research (NWO) (see http://foap.uvt.nl). We thank Jean Vroomen (Tilburg University) for allowing us to make use of the Pamar software, Marleen Roffel, Marina Elegeert, Gwendolyn Tabak, Femke Wiene and Lauraine Sinay for help with the data collection and the perceptual evaluations, and Lennard van de Laar for technical assistance. Parts of this research have been presented at the 2005 AVSP workshop in Vancouver, BC (Canada) (Summer 2005) and at the Interspeech 2006 conference in Pittsburgh, PA (USA). We have also benefitted tremendously from very helpful comments from Marion Dohen, Dani Byrd, and two anonymous reviewers of the Journal of Phonetics.

References

Baum, S., Pell, M., Leonard, C., & Gordon, J. (1997). The ability of right- and left-hemisphere-damaged individuals to produce and interpret prosodic cues marking phrasal boundaries. *Language and Speech*, *40*, 313–330.

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, *62*, 321–332.

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. *Proceedings of Interspeech* (pp. 1272–1275), Pittsburgh, PA.

Borod, J. C., Andelma, F., Oble, L. K., Tweed, J. R., & Welkowit, J. (1992). Right hemisphere specialization for the appreciation of emotional words and sentences: Evidence from stroke patients. *Neuropsychologia*, *30*, 827–844.

Borod, J. C., Koff, E., Yecker, S., Santschi, C., & Schmidt, J. M. (1998). Facial asymmetry during emotional expression: Gender, valence, and measurement technique. *Neuropsychologica*, *11*, 1209–1215.

Butler, S., Gilchrist, I. D., Burt, D. M., Perrett, D. I., Jones, E., & Harvey, M. (2004). Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns? *Neuropshychologia*, *44*, 52–59.

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 of the ICSLP* (pp. 2175–2179), Philadelphia.

Cho, T., & McQueen, J. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries phrasal accent and lexical stress. *Journal of Phonetics*, *33*, 121–157.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*, 292–314.

Davis, C. & Kim, J. (2007). Audio-visual speech perception off the topic of the head. *Cognition, 100*, B21–B31.

Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating uncertainty—the contribution of different audiovisual prosodic cues to the perception of confidence. *Proceedings of the Speech Prosody 2006*, Dresden, Germany, May 2006.

Dik, S. C. (1978). *Functional grammar*. Amsterdam: North-Holland.

Dohen, M. (2005). *Deixis prosodique multisensorielle*: *Production et perception audiovisuelle de la focalisation contrastive on français*. PhD thesis, Institut National Polytechnique de Grenoble (INPG), Grenoble.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, *44*, 155–172.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. van Cranach, et al. (Eds.), *Human ethology* (pp. 169–202). Cambridge: Cambridge University Press.

Erber, N. P. (1974). Effects of angle, distance, and illumination on the visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research*, *17*, 99–112.

Erickson, D., Fujimura, O., & Pardo, B. (1998). Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech*, *3–4*, 399–417.

de Gelder, B., Vroomen, J. H. M., & Bertelson, P. (1999). The role of face parts: The perception of emotions in the voice and face. In L. Grim Cabral, & J. Morais (Eds.), *Investigando a linguagem* (pp. 262–266). Florianópolis: Mulheres.

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the IEEE international conference on automatic face and gesture recognition* (pp. 381–386), Washington, DC.

Granström, B., House, D., & Lundeberg, M. (1999). Prosodic cues to multimodal speech perception. *Proceedings of the 14th ICPhS*, San Francisco.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America*, *102*, 3009–3022.

Hess, U., Kappas, A., & Scherer, K. (1988). Multichannel communication of emotion: Synthetic signal production. In K. Scherer (Ed.), *Facets of emotion: Recent research* (pp. 161–182). Hillsdale, NJ: Erlbaum.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proceedings of Eurospeech 2001* (pp. 1957–1960), Denver, CO.

Janzen, E. K. (1977). A balanced smile—a most important treatment objective. *American Journal of Orthodontics*, *72*, 359–372.

Joanette, Y., Goulet, P., & Hannequin, D. (1990). *Right hemisphere and verbal communication*. New York: Springer.

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audio-visual speech recognition. *Language and Speech*, *43*(1), 107–124.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., et al. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. In *Proceedings of the international conference of phonetic sciences* (*ICPhS*) (pp. 2071–2074), Barcelona, Spain.

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

Kowatari, Y., Yamamoto, M., Takahashi, T., Kansaku, K., Kitazawa, S., Ueno, S., et al. (2004). Dominance of the left oblique view in activating the cortical network for face recognition. *Neuroscience Research*, *50*, 475–480.

Krahmer, E., Ruttkay, Zs., Swerts, M., & Wesselink, W. (2002). Pitch, eyebrows, and the perception of focus. *Proceedings of speech prosody 2002* (pp. 443–446), Aix-en-Provence.

Krahmer, E. & Swerts, M. (2001). On the alleged existence of contrastive accent. *Speech communication*, *34*, 391–405

Krahmer, E., & Swerts, M. (2004). More about brows. In Zs. Ruttkay, & C. Pelachaud (Eds.), *Evaluating ECAs*. Dordrecht: Kluwer Academic Press.

Krahmer, E. & Swerts, M. (in press). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, accepted for publication.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

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.

Massaro, D. (1998). *Perceiving talking faces: From speech perception to a behavioural principle*. Cambridge: MIT.

Massaro, D. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granström, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Dordrecht: Kluwer.

Massaro, D., Cohen, M., & Smeele, P. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, *100*, 1777–1786.

Massaro, D., & Egan, P. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review*, *3*, 215–221.

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

Mehrabian, A., & Ferris, S. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, *31*, 248–252.

Mertens, I., Siegmund, H., & Grüsser, O.-J. (1993). Gaze motor asymmetries in the perception of faces during a memory task. *Neuropsychologia*, *31*, 989–998.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. Head movement improves auditory speech perception. *Psychological Science*, *15*, 133–137.

Munhall, K. G., & Vatikiotis-Bateson, E. (1996). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II. Advances in the psychology of speechreading and auditory–visual speech*. London: Psychology Press.

Nicholls, M. E. R., Searle, D., & Bradshaw, J. L. (2004). Read my lips: Asymmetries in the visual expression and perception of speech revealed through the McGurk effect. *Psychological Science*, *15*, 138–141.

Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, *20*, 1–46.

Pourtois, G., Debatisse, D., Despland, P. A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research*, *14*, 99–105.

Richardson, C. K., Bowers, D., Bauer, R. M., Heilman, K. M., & Leonard, C. M. (2000). Digitizing the moving face during dynamic displays of emotion. *Neuropsychologia*, *38*, 1028–1039.

Sakamoto, S., Tanaka, A., Tsumura, K., & Suzuki, Y. (2007). Effect of speed difference between time-expanded speech and talker's moving image on word or sentence intelligibility. *Proceedings AVSP conference*, Hilvarenbeek, The Netherlands, September 2007.

Srinivasan, R., & Massaro, D. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, *46*, 1–22.

Swerts, M., & Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. *Proceedings of speech prosody 2004*, Nara, Japan.

Terken, J., & Nooteboom, S. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, *2*, 145–163.

Thompson, L. A., Malmberg, J., Goodell, N. K., & Boring, R. L. (2004). The distribution of attention across a talker's face. *Discourse Processes*, *38*(1), 145–168.

Walker, A., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behavior and Development*, *6*, 491–498.

Walker, J. P., Daigle, T., & Buzzard, M. (2002). Hemispheric specialization in processing prosodic structures: Re-visited. *Aphasiology*, *16*, 1155–1172.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*, 555–568.