Title: Cognitive Processing of Audiovisual Cues to Prominence

Authors: Emiel Krahmer and Marc Swerts

Affiliation: Communication and Cognition, Tilburg University

Running head: Audiovisual cues to prominence

Full address: Emiel Krahmer
Communication and Cognition
Faculty of Arts
Tilburg University
P.O.Box 90153
NL-5000 LE Tilburg
The Netherlands
e-mail: e.j.krahmer@uvt.nl
phone: +31 13 4663070
fax: +31 13 4663110

Abstract

This article addresses two related questions regarding the cognitive processing of audiovisual markers of prominence in spoken utterances: (1) how important are visual cues to prominence from the face with respect to verbal 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 3 accents 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.

Keywords: audiovisual cues to prominence, reaction time experiments, classification tests, cognitive processing of accents

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 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" (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 the ears, but they also process information in the visual signal. A well-known example of how the perception of sound can depend on the perception of the visual modality is the so-called ventriloquism effect: when an observer is presented with synchronous but spatially discrepant visual and auditory information, the sound is perceived as being closer to the visual stimulus (Bertelson et al. 2000).

The interplay between auditory and visual signals on perception also appears when humans are exposed to a special case of sound, i.e., speech. This was spectacularly shown by McGurk and MacDonald (1976) who found that the display of an auditory /ba/ paired simultaneously with a silent movie of someone producing the syllable /ga/ often produces the percept /da/. An analogous effect at a different processing level has been reported for emotion perception: our perception of a visually displayed emotion (e.g. anger) can be biased by incongruent verbal cues, and vice versa (de Gelder & Vroomen 2000). Along the same lines, Pourtois et al. (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). These types of evidence on the synergy of speech and vision, both at a peripheral and more central processing level, show that auditory information is especially sensitive to visual information coming from a speaker's face.

The perceptual interdependencies of speech and facial cues have been of

practical use in a number of respects. Various studies show that the display of a speaking face can be used to augment an auditory signal which is degraded. It was found for instance that non-native language learners can be trained to make better use of phonetic information from visual cues to perceive a novel phonemic contrast. Hazan et al. (2005) report findings from native speakers of Japanese who, after audiovisual training, could improve their perception of the /v/-/b/-/p/ labial/labiodental distinction. Along the same lines, it was found that hearing-impaired people can more easily understand incoming speech when they are able to read the lips of their conversants, whose lips are synchronised with the incoming speech, even when their speaking partners are represented by means of a synthetic character (Beskow et al. 2004). Similarly, people without hearing deficits may also profit from visual cues coming from the face to better focus one's listening attention to the speech of one person in the context of a cacophony of surrounding noise, such as the voices of other people (the "cocktail-party" phenomenon) (Arons 1992). Finally, audiovisual interactions have been found to be useful to improve automatic speech recognition as well: the accuracy of an automatically decoded speech signal improves considerably when the system can also have access to the speaking face (Petajan 1985).

While previous work on how our perceptual system integrates auditory speech information and visual cues form 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, defined 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 sofar has focused on verbal cues to prominence, where it was found that accents are highlighted by means of variation in pitch, duration, loudness and voice quality (Ladd, 1996, Cruttenden, 1986). In more

recent years, it has regularly been reported that accents can also be marked by means of facial expressions, such as eyebrow movements or more exaggerated movements of the articulators (Cavé et al 1996, Granström et al. 1999, Keating et al. 2003, Dohen et al. 2004). Accordingly, such visual markers have been implemented in animated synthetic characters as markers of important bits of information (Cassell et al. 2001, Pelachaud et al 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 sofar have reported on how visual cues to prominence are processed by observers, and how they relate to auditory markers. Therefore, the current study will concentrate on 2 questions regarding their contribution for the perception of prominence: (1) how important are visual cues to prominence from the face with respect to verbal cues? and (2) are there differences between different facial areas in their cue value for prosodic prominence? Let us elaborate on these 2 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. 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 speech information (Mehrabian & Ferris 1967, Dijkstra et al. 2006). 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 sofar 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 et al. 2002), we found that people pay much more attention to auditory than to visual 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 visual features, yet at the same time provided evidence that the visual cues cannot be ignored either (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 2006), 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 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 verbal 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. While the manipulations 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 focus. 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 et al. 2002) which allows for more direct measurement of the import of accents on speech processing. Terken & 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. In line with earlier studies by Pourtois et al. (2002), one would predict that stimuli that are inconsistent regarding their use of visual and auditory cues to accent are more difficult to process than stimuli where the two types of cues do match.

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 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 (see also Dohen et al. 2004). 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, in press). Prosodic cues, however, tend to be located in the upper part of the face: 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). Similarly, de Gelder et al. (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, which prove 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 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. 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 surgery (Janzen 1977). 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 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 artefact like a vase. Thompson et al (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 spot. 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 cognitive processing of audiovisual markers of prominence in spoken utterances: (1) how important are visual cues to prominence from the face with respect to verbal cues? and (2) are there differences between different facial areas in their cue value for prosodic prominence? The following sections describe 2 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 3 accents 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 2 experiments themselves. We end this article with a general discussion about the implications of the various results for an audiovisual model of prosody perception.

Audiovisual recordings

As a basis for the two experiments described below, recordings were made of 6 native speakers of Dutch (4 male, 2 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. 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 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 receive an accent. Speakers were not given any instruction on how accents should be realized in audiovisual speech. 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 to the aforementioned conditions, speakers were asked to utter the sentence in a monotone, so without any auditory or visual markers of an accent. Figure 1 presents two stills of one of our speakers, 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.

insert figure 1 around here

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. This whole procedure was repeated twice. 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 minimize reflections.

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

Experiment 1

Method

Stimulus preparations

The audiovisual recordings of the different utterances produced by our 6 speakers were manipulated with Adobe Premiere™ to obtain all the stimulus variants. First, the sound and video recordings were separated, after which these 2 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 video 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; in this way, it was prohibited that our participants in their perceptual judgments could make use of the fact that some stimuli were artificial, and others were not. All the manipulations led to a total of 54 stimuli (3 auditory markers, 3 visual markers, 6 speakers). 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.

Participants

42 participants (18 male, 24 female) in total participated in this experiment on a voluntary basis, most of them recruited from the students population 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.

Procedure

The stimulus materials were presented in one of 4 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 to measure reaction times with audiovisual stimuli. The participants were instructed to click on one of three buttons on their keyboard, marked with the letters 1, 2 and 3, to indicate whether they had perceived the first, second or third word as being more prominent. Since the prominence ratings are relative judgments, they were told to click on the chosen button as soon as they thought what the most prominent word was, but in order to do so, they knew they were forced to listen to all three target words. The reaction times are measured with respect to the moment that a speaker finished uttering W3. Thus, a reaction time of 0 means that a participant has clicked a button on 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.

In addition, participants were told beforehand that after the test they would have to participate in 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 cloths. 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.

The actual experiment was preceded with a short exercise test with 6 congruent stimuli, in order to make participants acquainted with the kinds of stimuli and the general experimental procedure. If there were no questions from the participants about the experimental set-up after the pre-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 pretest and questionaire, took approximately 10 minutes per subject, of which about 8 minutes were

used for the central experiment.

Results

The first experiment has a complete $3 \times 3 \times 6$ design with the following factors: Auditory marker of an accent (3 levels: accent on W1, accent on W2, accent on W3), Visual marker of an accent (3 levels: accent on W1, accent on W2, accent on W3), and Speaker (6 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, 38 cases were treated as outliers, i.e. those cases where the reaction times were at a distance of at least 3 standard deviations from the overall mean. The majority of these typically consisted 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. Outliers were then replaced with the overall average reaction time. No further manipulations of reaction times were performed.

Before we embark on the results of the actual reaction times, let us first look at Table 1, which 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 an accent. Table 1 reveals that participants mostly designate that word in an utterance as being the more prominent one which also carries the auditory maker of an accent. 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 become less salient (Krahmer & Swerts 2001).

insert table 1 around here

Regarding the reaction times: a paired t-test which compares average times per speaker reveals that congruent stimuli differ significantly from incongru-

ent ones in that the latter give consistently slower reaction times (congruent: 73ms; incongruent: 150ms) ($t_{(41)}$= 4,952, $p < .001$). A three-way analysis for repeated measures was performed with the aforementioned within-subject variables as independent factors and with the reaction times (in milliseconds) as dependent variabele. Mauchy's test[1] was used to check the homogeneity of variance, and the Bonferroni correction was used for multiple pairwise comparisons. Main effects are displayed in Table 2. Main effects were found of Auditory marker of accent ($F_{(2,82)} = 20.523, p < .001, \eta_p^2 = .334$), Visual marker of accent ($F_{(2,82)} = 7.356, p < .01, \eta_p^2 = .152$) 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 accents later in the sentence. 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, which after closer inspection appears to be due to differences in the degree of speaker expressiveness with respect to visual or auditory cues. In addition, the anova gave a significant 2-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 do not significantly differ from the incongruent ones. The anova also gives significant 2-way and 3-way

[1] As a matter of fact, except for the 2-way interaction between auditory and visual markers, Mauchy'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 transparancy, we report on the normal degrees of freedom

interactions when Speaker is combined with the other factors, which again could be explained by the differences in overall expressiveness of speakers.

insert tables 2 and 3 around here

Discussion

The current experiment brought to light that visual cues have an impact on how accents are perceived, albeit that the visual markers appear to be not as strong as the auditory markers. While participants tend to focus on auditory cues, 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 or head nods) 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. This could 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. 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.

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.

Experiment 2

Method

Stimulus preparations

The stimuli used for this experiment are again based on the audiovisual recordings described in the earlier section. However, given that the current test is set up to learn more about the relative cue value of different facial areas, we did no longer include 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 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. Besides the original, we produced 4 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. Figure 2 gives some representative stills from one of our speakers (EK).

insert figure 2 around here

After having created these different versions, we made mirror images of all 5 versions of these stimuli. Figure 3 illustrates an original image together with its mirror.

insert figure 3 around here

All the manipulations led to a total of 180 stimuli: Visual marker of accent (3 levels: accent on W1, accent on W2, accent on W3), Speaker (6 levels), Facial area (5 levels: complete face, upper part visible, bottom part visible, left area visible, right area visible) and Display (2 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. An informal test confirmed that the monotonic utterances were indeed devoid of auditory accents.

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, and they had all had normal or corrected to normal vision and good hearing. None of the participants of experiment 2 had participated in experiment 1.

Procedure

The task was similar to that of our previous experiment, i.e., to indicate which word (W1, W2, or W3) was the more 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 highest amount of correctly detected 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 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

roughly 4.8 by 4.3 centimeters. In addition, we added the distance from the screen as a between-subjects factor (see also Jordan and 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 centimeters from the screen), in the middle condition participants 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 participants saw the original stimuli, and half of them saw their mirror versions. The actual experiment was again 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.

Results

The second experiment has a complete $3 \times 6 \times 5 \times 2 \times 3$ design with the following factors: Visual marker of accent (3 levels: accent on W1, accent on W2, accent on 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 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 more prominent one.

insert table 4 around here

The data were analysed with a multinomial 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. A customized model which only tests main effects revealed significant effects for Visual marker of 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 ($\chi^2 = 681.051, df = 2, p < .001$), and that the effect of Display was not significant. This model accounts for 24% of the variance. Table 5 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 than 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 due to differences in expressiveness between speakers, with relatively poor detection for stimuli from speaker MB and best results for speaker PB.

A model with only 2-way interactions between all factors presented above revealed significant interactions of Facial area with Speaker ($\chi^2 = 72.347, df = 20, p < .001$), with Distance ($\chi^2 = 31.620, df = 8, p < .001$), with Visual prominence (($\chi^2 = 16.562, df = 8, p < .05$), and with Display ($\chi^2 = 36.533, df = 4, p < .001$). In addition there were two more significant interactions between Visual prominence and Speaker ($\chi^2 = 230.116, df = 10, p < .001$), and between Visual prominence and Distance ($\chi^2 = 14.140, df = 4, p < .01$), with all the other interactions not being significant. This model with 2-way interactions could explain 32% of the variance. The interactions in which the factor Speaker

is involved can 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 accent 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 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 only reveals a significant interaction between Facial area and Display for horizontally blackened stimuli ($\chi^2 = 20.472, df = 8, p < .001$), 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 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 original left side is shown as the right side, while the reverse is true for the original right side becoming left side.

insert tables 5 and 6 around here

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's effect, as the observer, when making prominence judgments, tends to be biased to the side of a face which 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.

General discussion

This study has presented the results of 2 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 2 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 accented words can be detected in an utterance, albeit that they are less important than the auditory cues. So while this confirms earlier observations that verbal prosodic cues are more important than visual cues for the perception of prominence, it also makes clear that visual cues cannot be ignored, 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 et al. (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 accent are more difficult to process than stimuli where the two types of cues do match. Along the same lines, our

result agrees with what has previously been reported about the so-called ventriloquism effect, which is a classical example of how our perception of auditory cues is affected by visual information, which can both facilitate and hinder our perceptual capacities.

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 very much 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 transmitted information. Subsequent empirical research has often provided support for the predominance of visual signals for cuing emotion (e.g. Hess et al. 1988, Walker & Grolnick 1983) (See also Massaro & Egan (1996) and 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 (be it that 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).

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, had to produce 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 mouth positions, and head movements. One of the reasons why information from the upper part of the face, such as variation in eyebrow movement, turned out not to be so dominant, may be due to the fact that their production measures did not accurately measure all visually available information: whereas the mouth area was modelled using 17 dots, the top part was represented by only 2 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. 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 which 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 which 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 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.

In addition to such perceptual data, there is also speaker-related evidence from studies on emotional expression that faces are asymmetric, though results are not always entirely consistent. Borod et al (1998) report that most studies on emotional expression reveal that the left hemiface (which has greater connectivity to the right cerebral hemisphere) is more intense or moves more extensively than the right hemiface 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 et al. 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 this study were still in a preliminary stage of exploration. Given that pitch has been claimed to be one of the primary indicators of accent in spoken utterances, it would seem natural to expect that the left eyebrow is comparatively more relevant for the expression of focus than the right eyebrow.

We see different ways to further this research. First, the analyses presented in this article were based on data from 6 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 2 speakers were the authors while the other 4 were completely naive to the experimental question. Rather, the effects seemed more due to the fact that speakers differ in their degree of expressiveness. In the future, we could try and replicate our data with more natural utterances, with different speakers, to see to what extent our first results have general validity. Second, we have seen that our first experiment gave clear processing differences 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 sentences positions which have been shown to be functionally marked. 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 2006). 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.

Acknowledgments

This research was conducted in the context of the FOAP project, which is funded by the Netherlands Organisation of Scientific Research (NWO). We thank Jean Vroomen (Tilburg University) for allowing us to make use of the Pamar software, Marleen Roffel, Marina Elegeert and Gwendolyn Tabak for help with the data collection and the perceptual evaluations, and Lennard van de Laar for technical assistance.

References

Arons, B. (1992) A review of the cocktail party effect. *Journal of the American voice I/O society* **12**: 35–50.

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., Karlsson I., Kewley J. & Salvi G. (2004). SYNFACE - A Talking Head Telephone for the Hearing-impaired. In K. Miesenberger, J. Klaus, W. Zagler, D. Burger (eds.) *Computers helping people with special needs*, pp. 1178–1186, Berlin/Heidelberg: Springer.

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*

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*, Phildelphia, pp. 2175–2179.

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. (in press). Audio-visual speech perception off the top of the head. *Cognition* in press

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

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. von 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 Grim Cabral, L. & Morais, J. (Ed.), *Investigando a linguagem.* (pp. 262-266). Florianópolis: Mulheres.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, **14**(3):289-311.

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

Hazan, V., Sennema, A., Iba, M. & Faulkner, A. (2005) Effect of audiovisual perceptual training on the perception and production of consonants in Japanese learners of English. *Speech Communication*, **47**, 360-378.

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.

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

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., & Bernstein, L. (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 *Proc. SPIE*, **3644**: 34–44.

Kowatari, Y., Yamamoto, M., Takahashi, T., Kansaku, K., Kitazawa, S., Ueno, S., & Yamane, S. (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. *Proc. Speech Prosody 2002, Aix-en-Provence*, pp. 443–446.

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

Krahmer, E. & Swerts, M. (2006). Hearing and Seeing Beats: The influence of visual beats on the production and perception of prominence. *Proc. Speech Prosody 2006*, Dresden, Germany, May 2006.

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. & 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. & 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.

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

Petajan, E. D. (1985). Automatic lipreading to enhance speech recognition. *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, Atlanta, Georgia.

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**, 99105.

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.

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.

Captions to figures:

Figure 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

Figure 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

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

Table 1: Overview of perceived prominences for various combinations of auditory and visual markers to prominence.

Prominence		Chosen prominence			
Auditory	Visual	W1	W2	W3	Total
W1	W1	247	4	1	252
	W2	226	26	0	252
	W3	235	3	14	252
W2	W1	17	233	2	252
	W2	1	248	3	252
	W3	8	233	11	252
W3	W1	44	3	205	252
	W2	13	58	181	252
	W3	3	2	247	252

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

Factor	Level	RT (in ms)
Auditory prominence	W1	34
	W2	106
	W3	232
Visual prominence	W1	100
	W2	172
	W3	100
Speaker	EK	9
	LL	265
	MB	190
	ME	108
	MS	121
	PB	53

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

Prominence		
Auditory	Visual	RT (in ms)
W1	W1	-19
	W2	52
	W3	70
W2	W1	63
	W2	132
	W3	124
W3	W1	257
	W2	333
	W3	107

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

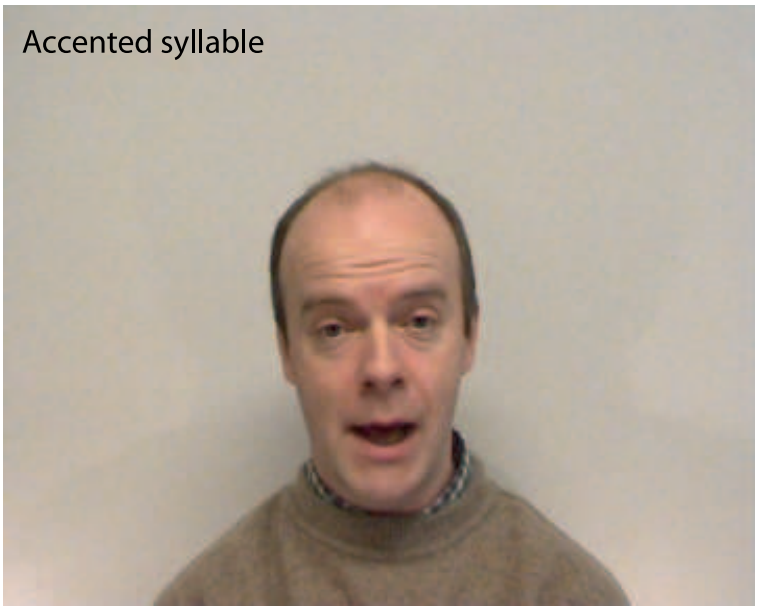
	Chosen prominence			
Visual prominence	W1	W2	W3	Total
W1	1416	307	255	1978
W2	425	1361	191	1977
W3	433	210	1337	1980

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

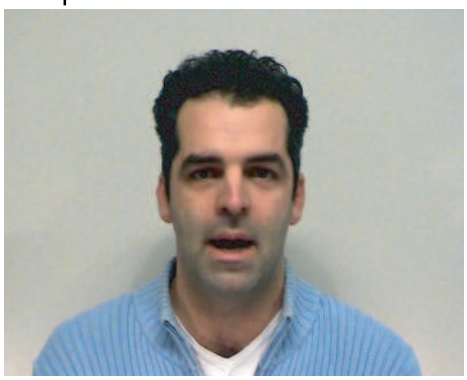
Table 6: Percentage correct prominence detection as a function of combined settings of display, distance, and facial area

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

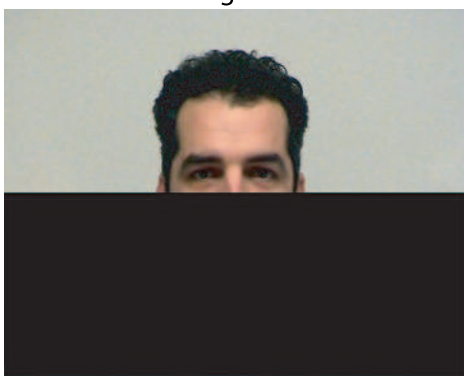


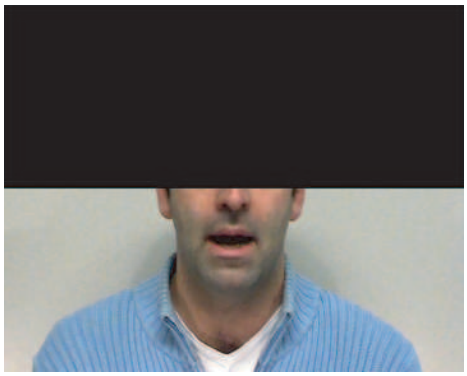
(figure 1)

Complete

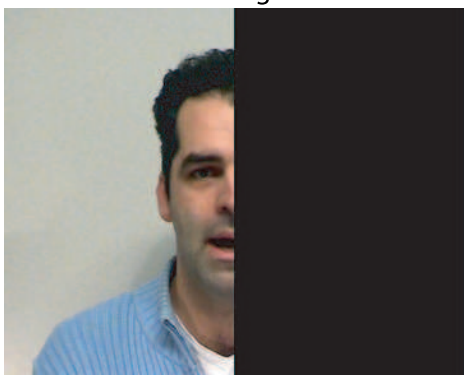


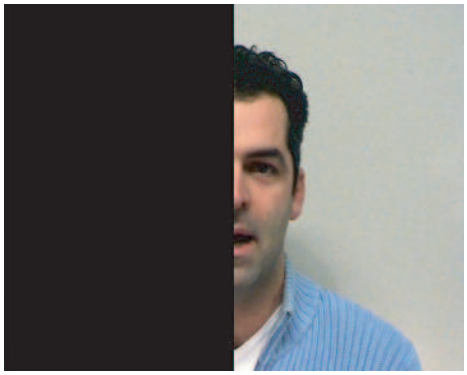
Vertical blackening



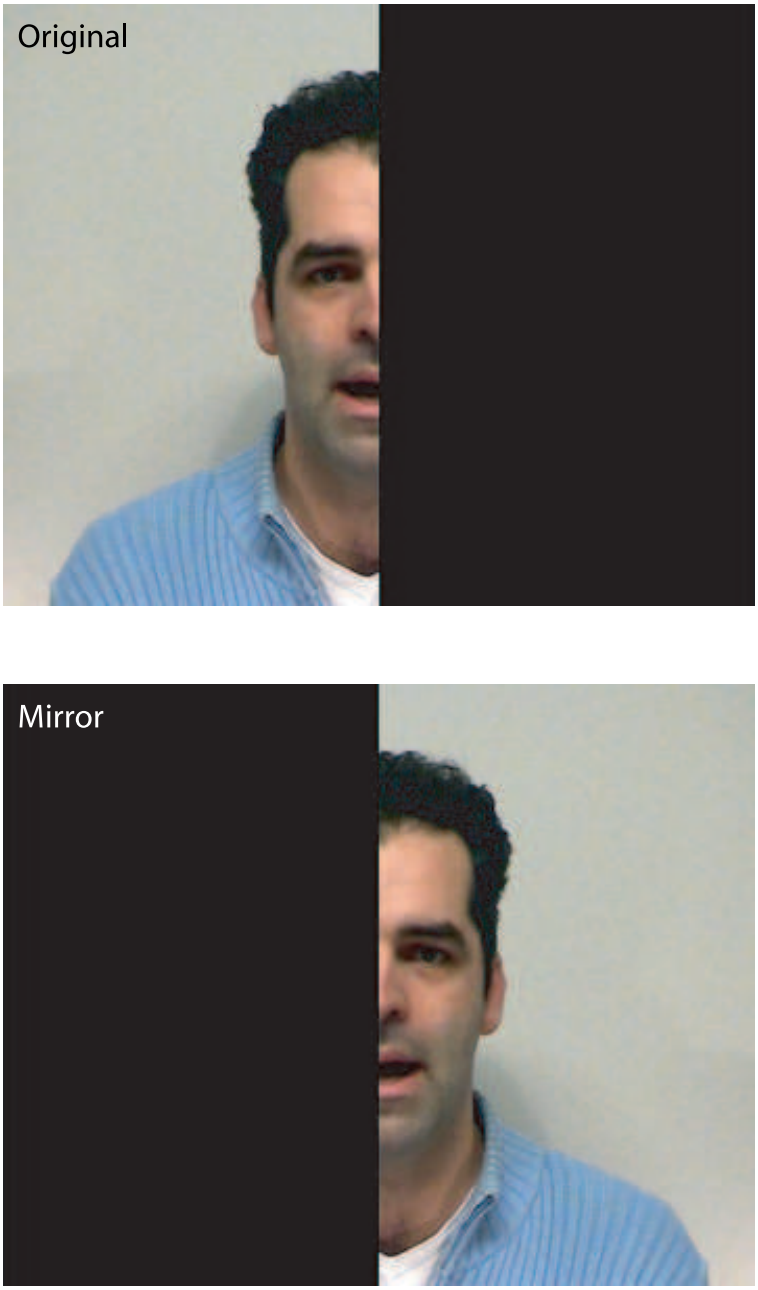


Horizontal blackening





(figure 2)



(figure 3)