Testing the effect of audiovisual cues to prominence via a reaction-time experiment

Emiel Krahmer and Marc Swerts

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

Abstract

This article discusses a perception experiment to investigate the relation between auditory and visual cues for marking prosodic prominence. The methodology makes use 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 markers of prominence were either congruent (occurring on the same word) or incongruent (in that the auditory and the visual cues were positioned on different words). Subjects were instructed to indicate as fast as possible which word they perceived as the most prominent one. Classification results show first of all that subjects' responses were much more dependent on auditory than on visual cues. In addition, however, we found that incongruent stimuli lead to slower reaction times than congruent stimuli, showing that visual cues do have an impact on the cognitive processing of prosodic prominence.
Index Terms: prominence, RT experiment, audiovisual speech

1. Introduction

The current article focuses on the interaction of auditory speech information and visual cues form a speaker's face. In particular, it will concentrate on 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). 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 (Ekman, 1979; Cavé et al 1996, Keating et al. 2003). Accordingly, such visual markers have been implemented in animated synthetic characters as markers of important bits of information (Cassell et al. 2001)

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. 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 subjects 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. We found that subjects pay much more attention to auditory than to visual information when they basically have to determine which word in an utterance represented new information (Krahmer et al. 2002). At the same time, a perception study revealed that observers tend to prefer visual and auditory cues to co-occur on the same word (congruent condition) rather than to be displaced on different words (incongruent), and that visual cues affect the perceived prominence of a word (Krahmer & Swerts, 2004)

Most of the tasks used in the experiments discussed above on prominence perception were offline, and consisted of elicited metalinguistic judgments of subjects on 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, a reaction time (RT) paradigm or eyetracking (Dahan et al. 2002) have been used successfully to more directly measure the import of accents on speech processing. For instance, Terken & Nooteboom (1987) found that people's reaction times are longer when given information is accented or when new information is deaccented. Sofar, 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. Therefore, the current study will make use of the RT paradigm to investigate the relative contribution of visual cues from the face for the perception of prominence. In the following, we describe the audiovisual recordings we used as a basis for our analyses, the procedure to run the RT experiment, and we end with a presentation of the results and a discussion.

2. Audiovisual recordings

As a basis for the experiment described below, recordings were made of 6 native speakers of Dutch (4 male, 2 female) between the ages of 20 and 40. In order to remove any visually distracting features, speakers did not wear any remarkable cloths, and were asked to take off their glasses during the data collection procedure. They were instructed to read out different variants of the sentence "Maarten gaat maandag naar Mali" (*Maarten goes Monday to Mali*) in such a way that the first (Maarten), second (maandag) or third content word (Mali) of the sentence would receive an accent. These three target words, which will be referred to as W1, W2 and W3 in the remainder of this paper, were

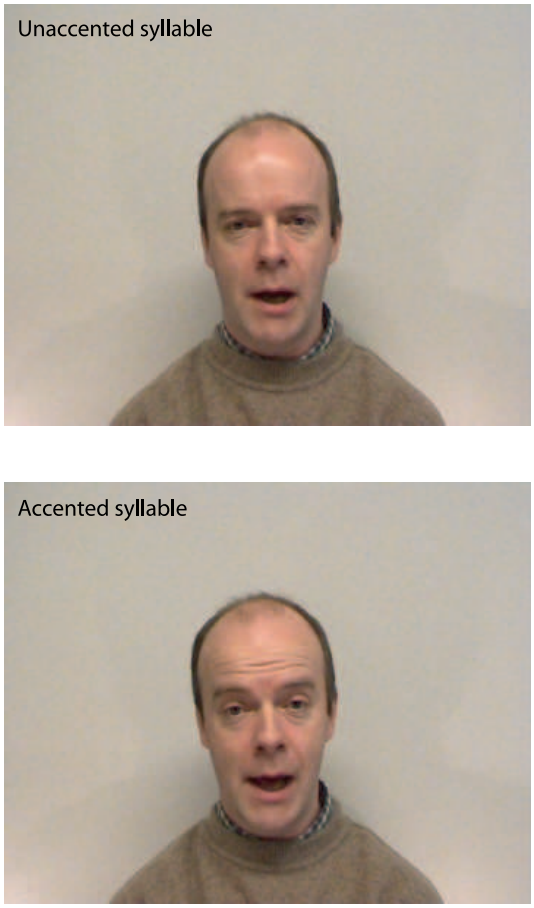

Figure 1: *Representative stills of a facial expression of one of our speakers while producing an unaccented (top) or accented (bottom) syllable in one of our target words.*

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. 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. 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 mimize reflections. These audiovisual recordings were used as a basis for the stimulus preparations of RT experiment.

3. Reaction time experiment

3.1. *Method*

3.1.1. Stimulus preparations

The audiovisual recordings of the different utterances produced by our 6 speakers were manipulated with Adobe Premiere^{TM} 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 obtained two sets of stimuli. The first set contained so-called **congruent** utterances, i.e., utterances in which the auditory and visual cues to prominence occurred on the same word. The second set consisted of **incongruent** stimuli in which the auditory and visual cues 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 subjects in their perceptual judgments would 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). Note that the naturalness of the artificial stimuli was extremely good, and did not lead to unwanted perceptual effects.

3.1.2. Participants

42 subjects (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 subjects was 27.7 (youngest: 21, oldest: 50). They were all right-handed, and had normal or corrected to normal vision and good hearing.

3.1.3. 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 numbers 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 possible after the whole utterance was finished. The inter-stimulus interval was 500 ms, in which time frame subjects had to respond. Reaction times were measured with respect to the end of an audiovisually displayed utterance. 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

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

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. The actual experiment was preceded with a exercise test with 6 congruent stimuli, in order to make subjects 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 took approximately 10 minutes per subject, of which about 8 minutes were used for the central experiment.

4. Results

The first experiment has a complete $3 \times 3 \times 6$ design with the following factors: Auditory markers(3 levels: prominence on W1, W2, or W3), Visual markers (3 levels: prominence on W1, W2, or 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. We did not normalize RT's per subject.

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) subjects had chosen to be the most prominent one, as a function of various positions of an auditory and visual markers. Table 1 reveals that subjects mostly designate that word in an utterance as being the more prominent one which also carries the auditory cue. Interestingly, that preference is stronger for cases where the chosen word also gets a visual cue: in other words, the congruent stimuli reveal a stronger preference for the auditory accent 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.

Regarding the reaction times: a paired t-test which compares average times per speaker reveals that congruent stimuli differ significantly from incongruent ones in that the latter give consistently slower reaction times (congruent: 73ms; incongruent: 150ms) ($t_{(41)}$= 4,952, $p < .001$). 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 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 ($F_{(2,82)} = 20.523, p < .001, \eta_p^2 = .334$), Visual marker ($F_{(2,82)} = 7.356, p < .01, \eta_p^2 = .152$) 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. Also, it turned out that speakers differ from others 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 interactions when Speaker is combined with the

[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 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

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

other factors, which again could be explained by the differences in overall expressiveness of speakers.

5. 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 subjects 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. This general outcome is in line with earlier studies by Pourtois et al. (2002), who 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. 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 accent 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 the current experiment showed that facial expressions matter in prominence detection, it remains to be seen which aspects of a face are more important for signalling accents. There are reasons to believe that different facial areas are distinct in their cue value for signalling accents. In recent work, we zoomed in on facial 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. Our latest results bring to light that the top area of a face is functionally more important for prominence perception than the bottom part, and that the left side of a face is stronger than the right side.

We see different ways to further this research. First, the analyses presented in this article were based on data from 6 speakers. 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. 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.

6. Acknowledgments

This research is part of the FOAP project (http://foap.uvt.nl), funded by the Netherlands Organisation of Scientific Research (NWO). We thank Jean Vroomen (Tilburg University) for allowing us to make use of the Pamar software, and Marina Elegeert and Lennard van de Laar for experimental assistance.

7. References

Cassell, J., Vihjálmsson, H., Bickmore, T., (2001), BEAT: the Behavior Expression Animation Toolkit, Proc. 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, Proc. ICSLP, Philadelphia, pp. 2175-2179.

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

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

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Bernstein, L. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. in *Proc. ICPhS* (pp. 2071–2074), Barcelona.

Krahmer, E., Ruttkay, Zs., Swerts, M., & Wesselink, W., (2002) Pitch, Eyebrows, and the Perception of Focus. *Proc. Speech Prosody* 2002, pp. 443-446.

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

Ladd, D.R. (1996). *Intonational Phonology*. CUP.

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.

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