PERCEPTUAL EVALUATION OF AUDIOVISUAL CUES FOR PROMINENCE

Emiel Krahmer,[1] *Zsófia Ruttkay,*[2] *Marc Swerts,*[3] *Wieger Wesselink*[4]

[1] BDM/Computational Linguistics, Tilburg University, The Netherlands
[2] CWI, Centre for Mathematics and Computer Science, The Netherlands
[3] CNTS, Antwerp University, Belgium and UCE, TU/e, The Netherlands
[4] Department of Mathematics & Computing Science, TU/e, The Netherlands

E.J.Krahmer@kub.nl, Zsofia.Ruttkay@cwi.nl, M.G.J.Swerts@tue.nl, J.W.Wesselink@tue.nl

ABSTRACT

This paper[1] reports on two experiments with a Talking Head that explore the ability of eyebrow movements to cue focus. The first experiment tests how listeners react to synthetic stimuli in which the eyebrow movements coincide with pitch accents versus those in which these two occur on different words. Results show that subjects prefer those utterances in which pitch and eyebrow movements are aligned on the same word. The second experiment investigates whether listeners are sensitive to eyebrow movements when they have to rate the prominence of particular words in audiovisual stimuli. This experiment shows that eyebrow movements both boost the perceived prominence of words that also receive a pitch accent, and downscale the prominence of unaccented words in the immediate context of the accented word.

1. INTRODUCTION

Speakers of languages such as Dutch and English have various linguistic strategies at their disposal to highlight specific information in their spoken utterances. In addition to morphosyntactic devices, they can exploit prosody to make words more prominent, e.g. using pitch accents as pointers to words that are new or contrastive. There have been claims in the literature that there exist additional visual cues to information status as well. In particular, different investigators argue that rapid eyebrow movements can have an accentuation role ([1,2,3,5,6,9].[2] However, only few of these studies are empirical in nature, and if so, they are generally purely speaker-oriented. As a result, it is yet unclear how such eyebrow movements are processed by listeners. One hypothesis is that they support the communication process, in that they represent an additional source of evidence for listeners to decide which words are important. An opposite hypothesis is that extra visual cues increase the cognitive load for a listener and are therefore counterproductive, since a distributed attention across different modalities may make linguistic interpretations of input utterances more difficult (see [4]).

The paper by [8] does tackle visual cues from a perceptual point of view. It reports on an experiment with a Talking Head, aimed at finding out the relative contributions of pitch accents and rapid eyebrow movements for the detection of focus. For this purpose, a "dialogue reconstruction" experiment was used: subjects had to perform a perceptual task in which they have to determine on the basis of the distributions of pitch accent and eyebrow movements on one utterance "blauw vierkant" (blue square) what the *preceding* utterance would have described: (1) a red square, (2) a blue triangle or (3) a red triangle. To perform this task subjects have to determine what the focus of the *current* utterance is: (1) the adjective ("blue"), (2) the noun ("square") or (3) both. Stimuli consisted of all possible combinations of pitch accents and eyebrow movements, so both cases where the two cues occurred on the same word (consistent stimuli), and cases where they did not (inconsistent stimuli). Results revealed that both pitch accents and eyebrow movements have a significant effect on the perception of focus, albeit that the effect of pitch is much larger than that of eyebrows. In particular, eyebrow movements primarily influenced the perception of focus when the pitch cues were ambiguous. However, the overall classification scores were less clear than those of an earlier experiment using speech-only stimuli ([11]), suggesting that the addition of another modality made the perceptual task more difficult.

That the eyebrow movements did not have a clear added value for focus detection in this study can potentially be ascribed to at least two factors, i.e., because they were *unnatural* and/or because they were *not functional*. First, various subjects indicated in a post-experimental interview that sometimes the eyebrows appeared to be poorly synchronized with the speech, in particular for inconsistent stimuli. They all reported that such mismatches made the animations less natural and caused considerable confusion. Second, it could be that eyebrow movements were not helpful because they are sometimes ignored by subjects when they have to determine the focus of an utterance; while the eyebrow movements were clearly above a perceptual threshold, they may simply not always be good indicators of focus, despite earlier claims found in the literature.

To gain further insight into the cue value of eyebrow movements for focus perception, we set up two experiments which address the two perceptual questions listed above. The first experiment investigates the naturalness of different combinations of pitch and eyebrow accents to see whether listeners have a preference for particular settings. The second experiment tests whether the perceived prominence of an accented word is affected by the presence or absence of an eyebrow movement. In the following, we will first describe how the stimuli of the two experiments were created, and then present the set-up and results of both experiments. The paper ends with a general discussion and conclusion.

2. MATERIALS

The stimuli for the two experiments described below are all an-

[1] Thanks to Han Noot, Matthew Stone and Mariet Theune for discussion.

[2] Non-rapid eyebrow movements can convey other meta-linguistic messages as well, such as surprise (raised) or doubt (frowned) (e.g., [5]).

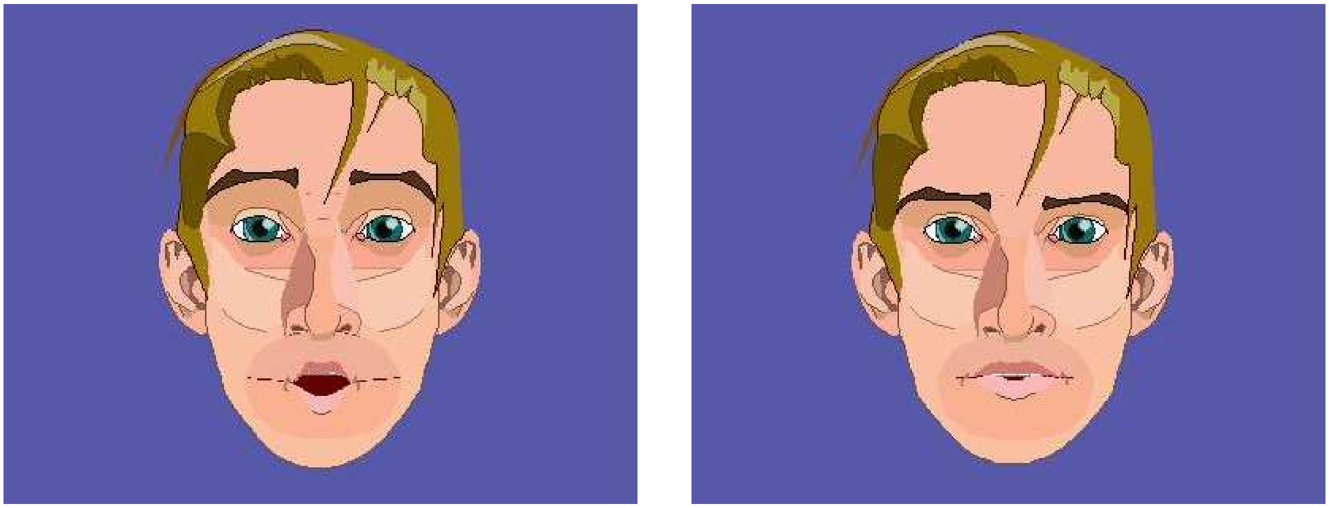

Fig. 1. Two stills from the Talking Head uttering "blauw vierkant" (blue square) with a raised eyebrow on the first word (left) and no eyebrow action on the second word (right).

imations produced with the *CharToon* environment ([10])[3] and had already been used in [8]. A 2D head of a male character formed the basis of the animations. CharToon animations are based on control points. By imposing a hierarchy on the control points, the number of parameters that control the movement of a face can be kept low. Visual speech is generated on the basis of a set of 48 visemes. Phonemes from the input are matched to corresponding visemes with a sampling rate of 100 ms, while intermediate stages are computed using linear interpolation. Eyebrow movements that were modeled using this environment always had the following pattern: first, a 100 ms dynamic raising part, then a static raised part of 100 ms, and finally a 100 ms dynamic lowering part (cf. Figure 1). The overall length of the movement is comparable to the average duration of rapid eyebrow movements of human speakers (±375 ms, [3]). We opted for slightly shorter movements due to the overall short duration of the stimuli. The sound produced by the Talking Head came from two human voices uttering different variants of the the phrase "blauw vierkant". These utterances were elicited in an earlier production experiment ([7]) that consisted of a set of dialogue games played by pairs of subjects, all native speakers of Dutch. During the game participants had to describe differently colored geometrical figures (including a blue square) on cards placed on a stack in front of them. The data obtained in this way allows for an unambiguous operationalization of focus: a property is defined to be *contrastive* if the previously described object had a different value for the relevant property, while it is *given* if the previously described had the same value for the relevant property. We say that a phrase is in *focus* if it is contrastive. By systematically varying the order of the cards in the stack, target descriptions ("blue square") were collected in three contexts: (*i*) focus on the adjective ("blue"), (*ii*) focus on the noun ("square") and (*iii*) all focus ("blue square"). A distributional analysis ([7]) reveals that for all the utterances used in the current experiment a word receives a pitch accent iff it is in focus. Interestingly, we had two kinds of speakers among our subjects: half of them happened to end their utterances with high boundary tones (H%), while the other speakers employed low boundary tones (L%). For the two perception experiments described below, the utterances from one high-ending and one low-ending human speaker were used. Given that the results for the different speakers were highly congruent in these different perceptual tasks, they will be collapsed below. The stimuli always had either a pitch accent on the first word, or on the second word, or on both. Eyebrow movements occurred on the first word, the second word, or the stimuli did not contain any eyebrow movements.[4]

3. EXPERIMENT 1

3.1. Goal

The goal of this experiment is to find out whether listeners' naturalness ratings of audiovisual stimuli are dependent on whether or not pitch accents and eyebrow movements occur on the same words.

3.2. Procedure

Audiovisual stimuli were presented to subjects in minimal pairs. The members of these pairs were always identical in terms of their sound properties, including the pitch accent distributions, i.e., with a pitch accent only on the first word ("blauw"), only on the second word ("vierkant"), or on both words. The members of the pairs differed in that one member had an eyebrow movement on the first word ("blauw"), whereas the other had such a movement on the second word ("vierkant"). As a result, the pitch accent and the eyebrow movement sometimes occurred both on the same word, sometimes they were located on different words, and sometimes there were two pitch accents, but only one eyebrow movement on either the first or the second word. Subjects were 25 native speakers of Dutch, none with a background in speech research. They watched

[3] See also `http://www.cwi.nl/projects/FASE/`.

[4] We did not include an eyebrow counterpart to "all focus," since this would involve either a raised eyebrow for a longer stretch of time or two rapid eyebrow movements in succession. For Dutch subjects both of these primarily have a non-focus signalling interpretation. Conversely, our elicited data did not show any cases of phrases without pitch accents, given the experimental set-up.

and listened to the Talking Head uttering the different pairs of the two-word phrase “blauw vierkant”. All pairs in both AB and BA order were randomly presented to subjects. The stimuli were displayed on a big screen in a group experiment, paced by one of the experimentors; sound came over a pair of loudspeakers. Subjects could watch and listen to each stimulus twice, and were encouraged to select - by forced-choice - the most natural animation from the stimulus pair after the first presentation and then give their definite response after the second one.[5] Before the actual experiment started, subjects entered a brief training session (consisting of three pairs of stimuli) to make them acquainted with the material and the setting of the experiment. No feedback was given on the ‘correctness’ of their answers and there was no communication with the conductor of the experiment. The experiment itself consisted of 12 stimuli (3 pitch accent distributions × 2 voices × 2 orders). Subjects were not informed about the kinds of cues they could use for their judgments. The entire experiment lasted approx. 5 minutes.

3.3. Results and discussion

Results of experiment 1 are given in Table 1. This shows that the pitch accent condition had a clear effect on subjects’ preferences for an eyebrow movement on either the first or the second word. The overall distribution is highly significant ($\chi^2 = 28.5$, df = 2, $p < 0.001$). Looking at the top row in this table, it can be seen that there is a moderate preference for the eyebrow movement on the first word when both words get a pitch accent. However, the next two lines with results on utterances with a single pitch accent clearly show that listeners strongly disprefer cases where the pitch accent and the eyebrow movement occur on different locations, and thus prefer cases where they are aligned on the same word. This effect is clearest when there is only a pitch accent on the first word. This result is in agreement with earlier observations on speech-only stimuli ([7]), which showed that nuclear accents that cue a single contrast in an utterance can be preceded but not followed by another smaller pitch accent. This preference for utterances in which pitch accents and eyebrow movements are aligned suggests that rapid eyebrow movements may indeed serve the same purpose as pitch accents, i.e., to render prominence to a word. This issue is investigated further in the next experiment.

4. EXPERIMENT 2

4.1. Goal

The goal of this experiment is to find out whether listeners are sensitive to eyebrow movements when they have to rate the prominence of audiovisual stimuli.

4.2. Procedure

Audiovisual stimuli again consisted of pairs of utterances that have identical sound properties (including pitch accent distributions), but are different in terms of visual characteristics. Unlike the previous test, however, the pairs now consisted of one utterance without any eyebrow movement, whereas the other had a movement on either the first (“blauw”) or the second word (“vierkant”). Given the findings of the previous experiment that listeners prefer cases where pitch accents and eyebrow movements occur on the same word, the eyebrow movement was always accompanied by a pitch accent, thus excluding cases where these two occurred on different words in the same phrase. Subjects were the same 25 native speakers of Dutch that participated in the previous experiment. The current experiment consisted of four different sessions, two with speech from the low-ending speaker, two with speech from the high-ending one. The procedure and experimental set-up were exactly the same as in the previous experiment, except that subjects were now instructed to rate the prominence of words. In two sessions, they were asked to pay attention to the first word (“blauw”) of the two utterances in a pair, and by forced choice pick the one which was perceived as most prominent; in two other sessions, they had to rate the prominence of the second word (“vierkant”) in the utterance pair. The stimuli were presented in two different random orders, to compensate for possible learning effect, and were presented to subjects in a list of 8 pairs, 4 of which were distractors, that consisted of utterance pairs that were not only different in terms of eyebrow movement distribution, but also in their sound properties, in an attempt to deliberately confuse people about the purpose of the experiment. Before a session started, subjects again entered a brief training session (consisting of one stimulus pair) to make them acquainted with the material and the setting of the experiment. Again no feedback was given on the ‘correctness’ of their answers and there was no communication with the experimentor. The second experiment lasted approximately 15 minutes.

Table 1. *Preference judgments for different utterance pairs with pitch accents on both words, only on first word or only on second word, and eyebrow movements on either the first or the second word ($N = 300$: 12 stimuli × 25 subjects).*

Pitch accent on	Eyebrow-movement on		Total
	First word	Second word	
Both words	60	40	100
Second word alone	38	62	100
First word alone	75	25	100

4.3. Results and discussion

Results of experiment 2 are given in Tables 2 (results for “blauw”) and 3 (results for “vierkant”), respectively. All pairwise comparisons of interest show a significant preference for one of the stimuli ($28.8 < \chi^2 < 42.3$, df = 1, $p < 0.001$). Both for judgments on the first and the second word it is true that the presence of an eyebrow movement has a clear effect on the perceived prominence of a word in two ways. First, they boost the accent strength on those words on which there was already a pitch movement for accentuation. The 2^{nd} and the 3^{rd} row of Table 2 illustrate this. E.g., in the former case, when both words are accented, people consider the first word (‘blauw’) more prominent when it is accompanied by an eyebrow movement than when it is not. In this way, pitch and eyebrows have an additive effect for ratings of prominence. Second, eyebrow movements downscale the perceived prominence of the words that appear in the immediate preceding or following context of the accented word. This is illustrated by the 1^{st} and the 4^{th} row of Table 2. For instance, when both words are accented, people consider the first word (‘blauw’) less prominent when the second word (‘vierkant’) is accompanied by an eyebrow movement than when no word is accompanied by an eyebrow movement. This is compatible

[5] Subjects were asked to choose the animation in which sound and image were best synchronized.

Table 2. *Prominence judgments for the first word in the utterance pairs ("blauw"). 8 stimuli, for all 25 listeners ($N = 200$; the total for each row is $50 = 2$ voices $\times$ 25 listeners). The left-hand side of the table characterizes the stimuli in terms of the distribution of pitch accents; the right hand side records how often subjects chose "blauw" to be most prominent in conditions where an eyebrow movement was absent, or the movement occurred on the first word or on the second word.*

Pitch accent on	Eyebrow-movement on		Total
Both words	No word	Second word alone	
	48	2	50
Both words	No word	First word alone	
	2	48	50
First word alone	No word	First word alone	
	3	47	50
Second word alone	No word	Second word alone	
	42	8	50

Table 3. *Prominence judgments for the second word in the utterance pairs ("vierkant"). 8 stimuli, for all 25 listeners ($N = 200$; the total for each row is $50 = 2$ voices $\times$ 25 listeners). The left-hand side of the table characterizes the stimuli in terms of the distribution of pitch accents; the right hand side records how often subjects chose "vierkant" to be most prominent in conditions where an eyebrow movement was absent, or the movement occurred on the first word, on the second word.*

Pitch accent on	Eyebrow-movement on		Total
Both words	No word	Second word alone	
	6	44	50
Both words	No word	First word alone	
	42	8	50
First word alone	No word	First word alone	
	44	6	50
Second word alone	No word	Second word alone	
	4	46	50

with earlier observations on speech-only stimuli discussed in [7] which showed that prominence judgments are very much dependent on the prosodic context, for instance, in that an isolated pitch peak is perceived as more prominent than the same peak presented in the context of an intonationally comparable pitch peak ("prosodic masking"). Along the same lines, the current test has shown that the prominence judgement of an unaccented syllable varies as a function of what happens in terms of visual cues in the immediate context. Note that results for both speakers used in the experiment and for both words are virtually identical.

5. DISCUSSION AND CONCLUSION

This paper has reported on two experiments with a Talking Head, to gain more insight into the cue value of eyebrow movements for the perception of prominence. The first experiment tested how listeners react to audiovisual stimuli in which the eyebrow movements coincide with pitch accents versus those in which these two occur on different words. Results showed that our subjects prefer those utterances in which pitch and eyebrow movements are aligned on the same word. The second experiment investigated whether listeners are sensitive to eyebrow movements when they have to rate the prominence of particular words. This experiment showed that eyebrow movements both boost the perceived prominence of words that also receive a pitch accent, and downscale the prominence of unaccented words in the immediately preceding or following context of the accented word. While both experiments thus confirm earlier claims that eyebrow movements are relevant for prominence perception, it remains puzzling why they were only minimally used in a more functional task described in [8]. It might be that eyebrow movements are exploited more consistently as a cue to different kinds of discourse information, or that listeners are more biased to using auditive information rather than visual information for focus perception. To gain more insight into this, it is useful to investigate real speaker behaviour in natural interactions. For the experiments described here, use was made 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 be nice to supplement the current results with findings of observations on real speakers to see whether they indeed use eyebrow movements for the determination of focus as suggested here, or whether these mainly signal other types of information, if any.

6. REFERENCES

[1] Birdwhistell, R., 1970, *Kinesics and context*, University of Pennsylvania Press.

[2] Cassell, J., Vihjálmsson, H., Bickmore, T., 2001, BEAT: the Behavior Expression Animation Toolkit, *Proceedings of SIGGRAPH'01*, pp. 477-486.

[3] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R., 1996, About the relationship between eyebrow movements and F_0 variations, *Proceedings ICSLP*, Phildelphia, pp. 2175-2179.

[4] Doherty-Sneddon, G., Bonner, L., Bruce, V., 2001, Cognitive demands of face monitoring: Evidence for visuospatial overload, *Memory & Cognition* 29 (7): 909-919.

[5] Ekman, P., 1979, About brows: Emotional and conversational signals, in: *Human ethology* M. von Cranach, et al. (eds.), Cambridge University Press, pp. 169-202.

[6] Granström, B., House, D., Lundeberg, M., 1999, Prosodic cues to multimodal speech perception, *Proceedings 14th ICPhS*, San Francisco.

[7] Krahmer, E., Swerts, M., 2001, On the alleged existence of contrastive accents, *Speech Communication* 34:391-405.

[8] 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.

[9] Pelachaud, C., Badler, N., Steedman, M., 1996, Generating facial expressions for speech, *Cognitive Science* 20:1-46.

[10] Ruttkay, Zs., ten Hagen, P., Noot, H., 1999, CharToon; A system to animate 2D cartoon faces, *Proceedings Eurographics*.

[11] Swerts, M., Krahmer, E., Avesani, C., to appear, Prosodic marking of information status in Dutch and Italian: A comparative analysis, *Journal of Phonetics*.