Pitch, Eyebrows and the Perception of Focus

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

[1] Computational Linguistics, Tilburg University, The Netherlands
[2] CWI, Centre for Mathematics and Computer Science, The Netherlands
[3] CNTS, Center for Dutch Language and Speech, University of Antwerp, Belgium
[4] IPO, Center for User-System Interaction, Eindhoven University of Technology, The Netherlands

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

Abstract

We report on an experiment with a Talking Head, aimed at finding out the relative contributions of pitch accents and rapid eyebrow movements for the perception of focus. For this purpose, we use a "dialogue reconstruction" experiment: subjects have to perform a perceptual task in which they have to determine on the basis of the distributions of pitch accent and eyebrow movements what the focus is of the current utterance. Our results reveal that both pitch accents and eyebrow movements can have a significant effect on the perception of focus, albeit that the effect of pitch is much larger than that of eyebrows.

1. Introduction

In Germanic languages such as Dutch and English, speakers use intonation to encode the status of the information they convey to their listeners. In particular, the distribution of pitch accents marks how utterances should be integrated in the larger discourse context: accents tend to distinguish information that is in focus (new or contrastive) from information which is given from the prior context. Like pitch accents, rapid eyebrow movements can play an accentuation role (e.g., Birdwhistell 1970, Condon 1976).[1] It has even been argued that there is a one-to-one connection between the two; see, for instance, the so-called *Metaphor of Up and Down* (Morgan 1953, Bolinger 1985:202ff): when the pitch rises or falls, the eyebrows follow the same pattern. In fact, to see that there is indeed a close connection between pitch and eyebrows, one may try to utter a two word phrase, say "blue square", with a pitch accent (but no corresponding eyebrow movement) on the word "blue" and a rapid eyebrow movement (but no corresponding pitch accent) on the word "square". Most people find this a difficult exercise.

One of the few empirical studies devoted to the relation between pitch accents and eyebrow movements is Cavé et al. 1996, who report on a significant correlation between the two (in particular, and surprisingly, for the *left* eyebrow). It appears that rapid eyebrow movements often co-occur with pitch accents. The opposite is not the case: people do *more* with their pitch than with their eyebrows. Cavé and co-workers suggest that eyebrow movements and pitch do not link automatically (e.g., due to muscular synergy), but coincide for communicative reasons. Naturally, this raises the question what these communicative reasons might be. There is some evidence that pitch accents influence the listeners' processing of incoming utterances. Terken & Nooteboom (1987) found that people's reaction times are longer when given information is accented or when new information is deaccented. If eyebrow movements 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 the literature on Talking Heads (i.e., combinations of computer animations with speech), there is no consensus on the timing and placement of eyebrow movements. Pelachaud et al. (1996) note that the decision to raise the eyebrows is affect dependent, but in the examples they discuss, pitch accents and eyebrows coincide. Thus to the question *I know that Harry prefers* POTATO *chips, but what does* JULIA *prefer?* the Talking Head of Pelachaud et al. (1996:19) would respond with:[2]

$$(\overline{\text{JULIA}}\ \text{prefers})_{\text{theme}}\ (\overline{\text{POPCORN}})_{\text{rheme}}$$

Cassell et al. (2001) use eyebrow raising (or "flashes" as they call them) more sparingly. The eyebrows are raised when an *object* in the "rheme" is described. So in reply to the question above, the algorithm of Cassell et al. would not produce a 'flash' on "Julia". It should be noted that neither Pelachaud *et al.* (1996) nor Cassell et al. (2001) report on evaluation: it is not known whether the animations are effective in the way human listeners process the information. We get no insight in the contribution of the eyebrow movement: its function remains unclear. These issue are addressed in this paper. We present an experiment with a Dutch Talking Head, aimed at understanding the relative contributions of pitch accents and eyebrow movements for the perception of focus. It is studied how listeners' interpretation of phrases uttered by the Talking Head is influenced by the distribution of pitch accents and eyebrow movements. The experimental paradigm is methodologically new in that it is explicitly directed towards *functional* aspects of the animation. In the next section we describe the stimuli. Then we move to the design (section 3) and the results (section 4) of the experiment. We end with a discussion in section 5.

2. Materials

The stimuli used in the perception experiment consisted of animations of a male Dutch Talking Head uttering the phrase "blauw vierkant" (blue square). Six male voices are used in the experiment. Two voices are synthetic, four human. We use both synthetic and natural voices in order to see to what extent the naturalness of the voice influences the perception of focus. A

[1] And again like pitch accents, eyebrow movements can convey other meta-linguistic messages as well, such as surprise (raised) or doubt (frowned) (e.g., Ekman 1979). In those cases the eyebrow movements are typically *not* rapid.

[2] Here and elsewhere, SMALL CAPS indicate an accent, and overlined words are accompanied by a rapid eyebrow movement.

Figure 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).*

human voice has more natural and better sounding prosody, but a synthetic voice might be better suitable to accompany the visual counterpart of a synthetic character. The four human voices were collected in an earlier production experiment (Krahmer & Swerts 2001). This production experiment 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. Here we ignore initial dialogue contributions (it would be odd to reconstruct the preceding context for them, see below), so all properties are either given or contrastive. 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 (see Krahmer & Swerts 2001:395) 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%). The use of high boundary tones produces an intonation pattern often referred to as list intonation. Taken without context, it sounds basically like a question. Here the data from two high-ending and two low-ending human speakers were used. A Dutch diphone speech synthesizer was used for the generation of the two synthetic versions. In one version the synthesis system ended all its utterances with a high boundary tone, in the other only low boundary tones were used. The respective contours were copied ("prosody transplantation") from those of one high-ending and one low-ending speaker.

The animations were produced with the *CharToon* environment (Ruttkay et al. 1999). A 2D head of a male character formed the basis of the animations. CharToon animations are based on control points.[3] 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 100ms, while intermediate stages are computed using linear interpolation. Rapid eyebrow movements coincide with the stressed syllable of either the first ("blauw") or the second word ("vierkant"). Notice that these are the eyebrow counterparts of focus on the adjective and focus on the noun respectively.[4] This implies that in certain stimuli eyebrow movements are associated with non-focussed (and thus unaccented) information. Eyebrow movements were clearly perceivable and always had the following pattern: first, a 100ms dynamic raising part, then a static raised part of 100ms, and finally a 100ms 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 ($\pm$375ms, Cavé *et al.* 1996). We opted for slightly shorter movements due to the overall short duration of the stimuli.

3. Experimental setup

Since Dutch speakers encode the discourse context in the accent structure of the current utterance and may also use rapid eyebrow movements for this purpose, we want to investigate to what extent listeners are able to "reconstruct dialogue history" (Swerts, Krahmer & Avesani, to appear) when interpreting utterances produced by a Talking Head. We have the following research questions: (1) To what extent do pitch accents and rapid eyebrow movements contribute to the perception of focus? (2) What happens when eyebrows and pitch provide conflicting cues? (3) What is the influence of contour (high vs. low-ending)? (4) What is the influence of voice (synthetic vs. human)?

[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.

Table 1: *Summary of the results: classification of all 36 stimuli, for all 25 listeners ($N = 900$; the total for each row is 150 = 6 voices × 25 listeners). The left-hand side of the table characterizes the stimuli in terms of the distribution of pitch accents and eyebrow movements; the right hand side of the table records how often subjects perceived the focus on the first word, the second word or on both words.*

PITCH		EYEBROW		FOCUS PERCEIVED ON		
blue	square	blue	square	blue	square	both
yes	yes	yes	no	**45**	**41**	**64**
yes	yes	no	yes	**21**	**70**	**59**
no	yes	yes	no	**25**	**91**	**34**
no	yes	no	yes	**27**	**90**	**33**
yes	no	yes	no	**112**	**22**	**16**
yes	no	no	yes	**104**	**30**	**16**

Table 2: *The influence of placement of rapid eyebrow movement on the perception of focus for the three low-ending voices ($N = 450$: 18 stimuli × 25 subjects) and the three high-ending voices ($N = 450$: 18 stimuli × 25 subjects) respectively. The left-hand side of the table characterizes the stimuli in terms of the distribution of boundary tones and eyebrow movements; the right hand side of the table records how often subjects perceived the focus on the first word, the second word or on both words.*

BOUNDARY TONE	EYEBROW		FOCUS PERCEIVED ON		
	blue	square	blue	square	both
low (L%)	yes	no	**107**	**64**	**54**
	no	yes	**80**	**82**	**63**
high (H%)	yes	no	**75**	**90**	**60**
	no	yes	**72**	**108**	**45**

Subjects were 25 native speakers of Dutch, none with a background in speech research. They watched and listened to the Talking Head uttering the two-word phrase "blauw vierkant" (blue square), with a particular intonation contour (taken from its original context) and a rapid eyebrow movement on either the first or the second word. The task for the subjects was to determine by forced choice 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.

The stimuli were displayed on a high-resolution color PC screen, sound came over the loudspeakers to the left and the right of the screen. The experiment was individually performed and self-paced. Subjects could watch and listen to each stimulus as often as they desired, although not much use was made of this option. Before the actual experiment started, subjects entered a brief training session (consisting of three 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 36 stimuli (3 pitch accent distributions × 2 eyebrow versions × 6 voices). Naturally, subjects were not informed about the kinds of cues they could use for the context reconstruction. After the experiment subjects were briefly interviewed to test whether they understood the experimental set-up and to find out which cues they claimed to be focussing on. The entire experiment lasted approximately 10 minutes. The stimuli were presented in two different random orders, to compensate for possible learning effect.

4. Results

Except for one of the 25 subjects,[5] all subjects indicated in the post-experiment interview that they quickly understood the task. Table 1 summarizes the results. The total distribution is significantly different from chance: $\chi^2 = 292.2, df = 10, p < .001$. First consider the cases with a single pitch accents, i.e., the cases with a single prosodic focus on either the adjective or the noun. Notice that in these cases the majority of subjects indeed perceived the focus on the adjective or the noun respectively, no matter which of the words is accompanied by an eyebrow movement.

Certainly for these single prosodic focus cases, the distribution of pitch accents is more important for the perception of focus than the placement of eyebrow movements. This is also reflected by the fact that in the post-experiment interview, all subjects indicated that they paid most (if not all) attention to information in the auditory channel. Nevertheless, there is an overall effect of eyebrow movements: the distribution obtained with an eyebrow movement on the first word is significantly different from the distribution with a movement on the second word ($\chi^2 = 19$, $df = 8$, $p < .025$). Closer inspection of table 1 reveals that this is primarily due to cases with a double pitch accent. If we compare the cases in which the first word (the adjective "blauw") is associated with a rapid eyebrow movement with the cases in which the first word is not associated with such a movement, we see that in the former case 45 stimuli are perceived as having focus on the first word as opposed to 21 in the latter case. And, conversely, when we compare the cases in which the second word (the noun "square") is associated with a rapid eyebrow movement with the cases in which it is not, we see that in the former case 70 stimuli are classified as having focus on the noun as opposed to only 41 in the latter case. In other words, when the intonation contour provides less cues about the focus (since it contains two pitch accents), eyebrow movements have relatively more impact. Overall, the results for the four human voices are similar to the results for the two synthetic voices, albeit that the effect of eyebrow movements is a bit (but not significantly) more pronounced for the synthetic ones. One subject explicitly indicated that she "trusted" the human voices more than the synthetic ones, and thus paid special attention to pitch accents in the former situation.

Table 2 shows the results for the three low-ending (top) and the three high-ending voices (bottom) respectively. Interestingly, eyebrow movements appear to do more for low-ending speakers than for high-ending ones. In the former but not in the latter case, there is significant difference between a raised eyebrow on the first word and one on the second word ($\chi^2 = 6.82$, $df = 2$, $p < .05$ for low-ending speakers, $\chi^2 = 3.84$, *n.s.* for high-ending speakers). For the low-ending speakers we see that when the eyebrow movement occurs on the first word, it is most likely to be classified as having focus on the first word, whereas for a movement on the second word, the focus is more likely

[5] This one subject systematically scored the reverse of what we expected. He assumed that (visual or auditory) emphasis basically indicates "similarity" (i.e., givenness). His results are included, and account for part of the noise in the data.

to be perceived on the second word. The results for the high-ending speakers reveal a somewhat similar trend, but here there is a stronger overall bias in the direction of focus on the second word. This is probably due to the pronounced high boundary tones which makes the final word stand out perceptually (see Krahmer & Swerts 2001).

5. Discussion and future work

The results of the experiment can be summarized as follows: both auditory (accent distribution) and visual (eyebrow movement) cues can have a significant effect on the perception of focus. However, the effects clearly differ in magnitude; the impact of pitch accents is large, that of rapid eyebrow movements comparatively small. The visual cues contribute more when the auditory cues are inconclusive. In addition, the visual cues appear to be stronger for the low-ending contours than for the high-ending ones; the pronounced high boundary tone and the resulting bias for perceiving focus on the second word swamps the impact of eyebrows. That the auditory cues appear to be more important for focus perception may —with hindsight— be explained as follows: as noted in the introduction, human speakers do more with their pitch than with their eyebrows, so it is not unnatural that human listeners have learned to pay more attention to changes in pitch than to eyebrow movements.

That speech is dominant has two side effects. First of all, our results basically confirm the (Dutch) results of Swerts, Krahmer & Avesani (to appear), where the same experimental paradigm was applied to the speech-only data obtained with the four human voices. The difference is that there is overall somewhat more confusion in the current experiment. In part, the increase in confusion can be ascribed to the eyebrow movements. Certainly, they account for much of the "confusion" in the cases with a double pitch accent. Second, the dominance of speech also explains why 'inconsistent' cues (i.e., eyebrow movements on unaccented items) do not have a strong influence on the results. Interestingly, various subjects indicated in the post-experimental interview that sometimes the eyebrows appeared to be poorly synchronized with the speech. All these subjects reported that such mismatches made the animations less natural and they also claimed that the mismatches caused considerable confusion (compare the earlier cited findings of Terken & Nooteboom 1987).

Pilot work suggests that there is an interesting difference between native and non-native speakers of Dutch, in that the non-native speakers (in particular non-Germanic ones) appear to benefit more from eyebrow movements than native speakers (cf. the suggestion from Granström et al. 1999 that eyebrow motion is a more universal cue for prominence than pitch). This is probably also what one would expect; the prosodic cues are rather subtle, and it would not be surprising if non-native speakers have more difficulty in catching and correctly interpreting them. It would also be highly interesting to see what happens with Talking Heads for non-Germanic languages such as Italian. Italian has been claimed (e.g., Ladd 1996) to be intonationally different from Germanic languages such as Dutch in that it strongly resists deaccentuation within syntactic constituents. The acoustic data of Swerts, Krahmer & Avesani (to appear) indeed reveals this to be the case: their Italian speakers always put an accent on both the adjective and the noun ("TRIANGOLO NERO"), irrespective of prior context. They performed a dialogue reconstruction experiment for the Italian speech-only data. Interestingly, though not unexpectedly, the results reveal that Italian listeners systematically fail to correctly classify the Italian utterances in terms of dialogue history. We are currently planning to do the dialogue reconstruction experiment with an Italian Talking Head lifting its eyebrows on either the first ("triangolo") or the second word ("nero"). We would expect that rapid eyebrow movements have more impact for the Italian head than for the Dutch one, since the auditory cues are less informative for Italian than for Dutch.

Acknowledgements A preliminary version of this paper was presented at the Dagstuhl workshop on Coordination and Fusion in Multimodal Interaction (November 2001). Thanks are due to Han Noot, Erwin Marsi, Catherine Pelachaud, Koert van Stavenuiter, Matthew Stone, Mariët Theune and Raymond Veldhuis for discussion and comments.

6. References

[1] Bolinger, D., 1985, *Intonation and its parts*, London: Edward Arnold.

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

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

[4] 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.

[5] Condon, W., 1976, An analysis of behavioral organization, *Sign Language Studies*, 13:285-318.

[6] Ekman, P., 1979, About brows: Emotional and conversational signals, in: *Human ethology: Claims and limits of a new discipline*, M. von Cranach, K. Foppa, W. Lepenies, D. Ploog (eds.), Cambridge University Press, pp. 169-202.

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

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

[9] Ladd, D., 1996, *Intonational phonology*, Cambridge: Cambridge University Press.

[10] Morgan, B., 1953, Question melodies in American English, *American Speech* 2:181-191.

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

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

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

[14] Terken, J., Nooteboom, S., 1987, Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information, *Language and Cognitive Processes* 2 (3/4):145-163.