Hearing and seeing beats

Krahmer, E.J.; Swerts, M.G.J.

Published in:
Proceedings of Speech Prosody 2006

Publication date:
2006

Hearing and Seeing Beats

The influence of visual beats on the production and perception of prominence

Emiel Krahmer & Marc Swerts

Communication and Cognition
Tilburg University, The Netherlands

Abstract

Speakers can employ a variety of means to indicate that a word is important, including auditory cues such as pitch accents and visual cues such as manual gestures, head nods and eyebrow movements (visual beats). In this paper, we look at the relation between visual and auditory cues for prominence, based on data collected with an original experimental paradigm in which speakers were instructed to realize a particular target sentence with different distributions of auditory and visual cues. The first experiment revealed that visual beats have a significant effect on the spoken realization of the target words. When a speaker produces a visual beat, the word uttered simultaneously is produced with relatively more spoken emphasis, irrespective of the position of the auditory accent. The second experiment showed that when participants *see* a speaker realize one of these beat gestures on a word, they perceive this word as more prominent than when they do not see the beat gesture.

1. Introduction

When speakers want to signal to their conversation partners that a word is important, for instance, because it represents new or contrastive information, they can do this with an auditory pitch accent but also with a variety of manual or facial gestures, including head nods, eyebrow movements and manual gestures.[1] Such facial and manual gestures are commonly referred to as visual beats, and it has been suggested on a number of occasions, that there appears to be a connection between pitch accents and visual beats. One of the earliest who made this connection is Dobogreav, as described in McClave (1998), who in 1931 noticed that when speakers were not allowed to make manual gestures, their speech displayed less variation in pitch. Similarly, Bolinger (1985) suggested that eyebrow movements and manual gestures have a tendency to follow pitch movements.

Only a handful of studies have investigated the relation between pitch and (facial or arm) gestures empirically. Cavé et al. (1996), for instance, report on a pilot production study with a limited number of speakers and they indeed found a significant correlation between fundamental frequency (F_0; a common acoustic representation for pitch) and the (left) eyebrow movement. McClave (1998), with the explicit aim to verify Bolinger's observation as applied to manual gestures, describes a microanalysis of three speakers, and found no significant correlations between pitch and manual gestures, although they do parallel each other on occasion. These inconclusive findings raise at least two questions: is there a different influence of different kinds of beats on speech, and how do addressees perceive these beats?

In an attempt to answer these questions, we studied the relation between pitch accents and different kinds of visual beats. First, we collected audiovisual materials using an experimental approach described in section 2, in which a number of speakers were instructed to produce a single target sentence in different conditions (a common procedure in experimental speech studies). In a number of variants the auditory and visual beats coincided, whereas in others there was a deliberate mismatch (or incongruency) between the two. The data thus collected were used in two experiments. In experiment I ("hearing beats") we investigate to what extent producing a visual beat influences the production of speech (section 3), while in experiment II ("seeing beats") we look at the influence of seeing a visual beat on prominence perception (section 4).

2. Data collection

Participants For the data collection, 11 speakers were recorded (age 20-45), 3 male and 8 female ones.

Task definition Participants were given the task to utter the four word sentence "Amanda gaat naar Malta" (*Amanda goes to Malta*), in a number of different variants. This target sentence is typical for studies of prominence and has been used before in studies of speech production and perception for Dutch (e.g., Gussenhoven et al. 1997). Throughout this paper, we refer to "Amanda" as the first target word (abbreviated as **W1**) and "Malta" as the second target word (abbreviated as **W2**).

Speakers were instructed to utter this sentence with a visual beat (either a manual beat gesture, a head nod or an eyebrow movement) on W1 or W2, and with an auditory accent on W1, W2 or on neither word. This gave rise to $3 \times 2 \times 3 = 18$ different realization tasks of the target sentence. Cases in which a gesture and a pitch accent should be realized on the *same* word are referred to as **congruent**, cases in which they are associated with *different* words are referred to as **incongruent**.

Each individual task was displayed on a separate card, where words that should receive a pitch accent were marked in bold face and words that should receive a beat gesture were marked with a specific icon illustrating a hand, a head or an eye plus eyebrow as markers for a manual beat gesture, a head nod or a rapid eyebrow movement respectively.

Procedure The audiovisual recordings of the 11 speakers were made in a research laboratory at Tilburg University. Speakers were seated on a chair in front of a digital camera that recorded their upper body and face (25 fps). They were given

[1]The research was conducted as part of the VIDI-project "Functions of Audiovisual Prosody (FOAP)", sponsored by the Netherlands Organisation for Scientific Research (NWO), see `foap.uvt.nl`. Many thanks to Kelly de Jongh, Carel van Wijk, Edwin Commandeur and Lennard van de Laar.

a brief instruction, explaining the experimental setup and the task representations on the cards. They were told that only a word in bold face should be emphasized in speech. In addition, the three gesture icons (for head nod, eyebrow movement and manual gesture) were explained by the experimenter, and the intended gestures were illustrated; again participants were told that only words that were marked with such an icon should be uttered while making the corresponding gesture. Participants were told that they might find some of the tasks difficult to realize and that they were free to practice and repeat the sentence displayed on a card until they felt they could not further improve their realization in subsequent attempts.

For the collection phase, speakers were given a stack of 18 cards, each containing a single task, and speakers were instructed to go through this stack twice (referred to below as trial 1 and trial 2). They were asked to first read the task on the card, and then utter the sentence with the required distribution of beat gestures and pitch accents, using as many attempts as they felt necessary.

Data processing The video recordings were read into the computer and segmented per task. When a speaker produced multiple attempts for a given task, only the last attempt was selected. For each speaker and task, it was checked whether the intended words were accompanied by the intended manual or facial beat gestures; this was indeed the case. This resulted in a corpus of 396 sentences (11 speakers $\times$ 18 tasks $\times$ 2 trials).

3. Experiment I: Hearing beats

3.1. Method

All occurrences of W1 (*Amanda*) and W2 (*Malta*) were scored by three independent labellers in terms of prominence, where the following three-way distinction was made: a word was assigned a 0 if no pitch accent was noticed, a 1 if a minor pitch accent was heard and a 2 for a clear pitch accent. Labelling was performed individually on the basis of only the audio signal. Sentences were played in a random order, so that the labellers were always blind to condition, in order to avoid circularity.

A Pearson correlation analysis revealed that there was a high amount of agreement on the prominence-scores among the three labellers, with an average Pearson correlation of .65). The individual scores of the three labellers were summed to obtain one prominence score per word, which thus ranges from 0 (no pitch accent according to all three labellers) to 6 (a major pitch accent according to all three labellers). Finally, we computed an **auditory difference score** by subtracting the summed prominence scores for the second word from the summed prominence scores of the first word. This results in range from -6 to 6, where a positive difference score indicates that the first word is relatively more prominent than the second, while a negative score indicates that the second word is relatively more prominent.

3.2. Results

A full factorial Analysis of Variance (ANOVA) with auditory difference score as the dependent variable and with speaker as repeated measure was used to find out if and how the auditory difference scores depended on the within subjects factors auditory accent (with levels no pitch accent, pitch accent on W1, and pitch accent on W2), type of visual accent (head nod, eyebrow movement, manual beat gesture), position of the visual accent (W1, W2) and trial (first, second). The main effects are described in Table 1. Significant main effects were found of position of the visual accent ($F(1,9) = 15.486, p < .01, \eta_p^2 = .632$) and of auditory accent ($F(2,18) = 31.706, p < .001, \eta_p^2 = .779$). All pairwise comparisons for the three levels of the latter factor are statistically significant at the $p < .01$ level, after a Bonferroni correction. Neither type of visual accent nor trial had a significant effect ($F < 1$ in both cases), which means that for the auditory difference score it does not matter whether the target utterance was produced in the first round or in the second round, nor does it matter whether the visual accent was a head nod, an eyebrow movement or a manual beat gesture.

Table 1: *Average* **auditory difference scores** *as a function of accent, type of gesture, position of gesture and trail (std. errors between brackets).*

Factor	Level	A-diff (s.e.)
Accent	None	-.30 (.17)
	W1	1.77 (.25)
	W2	-1.71 (.40)
Type	Head nod	.03 (.24)
	Eyebrow	-.12 (.21)
	Manual beat	.16 (.19)
Position	W1	.60 (.18)
	W2	-.76 (.26)
Trial	First	.01 (.13)
	Second	-.17 (.21)

Table 2: *Average auditory difference score as a function of the position of the auditory and the visual acecent respectively (std. errors between brackets).*

		Pitch accent on		
		W1	None	W2
Visual	W1	2.32 (.40)	0.70 (.25)	-1.22 (.42)
Beat on	W2	1.22 (.30)	-1.30 (.39)	-2.20 (.46)

The significant main effects can be illustrated using Table 2 which illustrates the influence of pitch accents and visual beats on the auditory difference score (the results for the different beat gesture and trials are collapsed as these did not have a significant influence on the results). First, it can be observed that on average a pitch accent on W1 results in a positive difference score and an auditory accent on W2 results in a negative difference score (and recall that a positive auditory difference score indicates that the first word is relatively more prominent, while a negative score indicates that the second word is more prominent). The same can be observed for the visual beats: if one of these occurs on W1, the difference score is positive on average and if one occurs on, the average difference score is negative. It is highly interesting to find that these two effects are independent (there is no significant interaction between the two factors). As a result, congruent utterances lead to higher absolute difference scores than incongruent utterances.

3.3. Conclusion

The first experiment revealed that visual beat gestures have a clear impact on the *spoken realization* of target words. When a speaker produces a visual beat while uttering the first or second

word of interest (i.e., Amanda or Malta), the relative spoken prominence of that particular word increases, while the relative spoken prominence of the other word decreases. This is true irrespective of which word in the utterance is realized with a pitch accent. Interestingly, the kind of visual beat (eyebrow, head nod or manual beat) does not have a significant effect. In the second experiment, the effects of *seeing* a visual beat gestures on prominence perception will be addressed.

4. Experiment II: Seeing beats

4.1. Method

Participants Twenty people participated in the second experiment, 9 men and 11 women, with an average age of 35. None were involved with the production study, and none had experience with audiovisual research.

Stimuli Data from three speakers, recorded during the data collection phase, were used as stimuli for the perception study. These three speakers were selected because their recordings were of a good quality (no background noise) and because they spoke most clearly throughout the production phase. The perception study concentrated on eyebrow movements and manual gestures (head nods were left out to keep the length of the perception experiment reasonable). This implies that 12 different stimuli per speaker could be used, which were all selected from the second trial. All fragments were offered in two variants to the participants: an audiovisual variant (i.e., as original recordings) and an audio-only variant (with a black screen). In total, we used 72 stimuli (3 speakers $\times$ 12 utterances $\times$ 2 conditions [audiovisual, audio-only]). Audiovisual and audio-only stimuli were interleaved, and offered in one of two random orders.

Task Participants had to rate the perceived prominence of the first (W1) and the second word (W2) on a 10 point scale, where 1 indicated "no prominence" and 10 indicated "strong prominence". Such a 10 point scale allows for fine-grained judgments and, moreover, such scales are typical of the Dutch school grading system so that all participants are familiar with it. The participants were confronted with the 72 stimuli in two blocks, and were instructed to concentrate on one of the two target words per block. All participants rated the prominence of both W1 and W2 in all stimuli during two separate experimental sessions in which they either focus on the first or the second word.

Procedure The experiment was run on a laptop with a 15 inch screen and with separate loud speakers positioned to the left and right of the computer. The experiment was individually performed. After participants were instructed about the goal of the experiment (prominence perception), a brief training session started, consisting of 4 stimuli (from a fourth speaker not used in the actual experiment) illustrating the different visual beats (eyebrow movements, manual gestures) and presentation formats (audiovisual and audio-only). Stimuli were preceded by a visual stimulus ID and an auditory beep, and followed by a 3 second interval in which a white screen was displayed and during which participants could rate the prominence of the target word on an answer form. If participants had no questions about the procedure, the actual experiment started and there was no further interaction between participant and experimenter.

Table 3: *Average* **visual difference scores** *as a function of accent, type of gesture, position of gesture and trail (std. errors between brackets).*

Factor	Level	Scores for W1 V-diff. (s.e.)	scores for W2 V-diff (s.e.)
Accent	None	-.01 (.14)	.07 (.15)
	W1	.55 (.12)	.28 (.11)
	W2	-.07 (.18)	.24 (.10)
Type	Eyebrow	-.14 (.10)	.10 (.07)
	Hand	.44 (.10)	.29 (.09)
Position	W1	.55 (.17)	-.15 (.09)
	W2	-.24 (.08)	.54 (.12)
Speaker	S1	.22 (.13)	.37 (.10)
	S2	-.06 (.13)	-.12 (.11)
	S3	.30 (.11)	.34 (.15)

Half of the participants started rating the prominence of the first word W1, "Amanda", (in all 72 stimuli), the other half started rating the second W2, "Malta" (in all stimuli). After rating the prominence for one word, participants could take a short break before starting to rate the other word. Scoring for different words was always done in a different random order, so that possible learning effects could be compensated for.

In Study II the primary interest is in the effect of *seeing* (congruent and incongruent) beat gestures on prominence perception. We therefore define a **visual difference score**, by subtracting the prominence score in the audio-only condition from the prominence score in the audiovisual condition: if the result is a positive number, this indicates that seeing the speaker increases the perceived prominence of the target word, while a negative number indicates that seeing the speaker results in a decrease of perceived prominence for the target word.

4.2. Results

A full factorial Analysis of Variance (ANOVA) with observer as repeated measure was used to find out how the visual difference score depended on the within subjects factors auditory accent (with levels no pitch accent, pitch accent on W1, pitch accent on W2), type of visual accent (eyebrow movement and manual beat gesture), position of the visual accent (W1 or W2) and speaker (S1, S2 or S3). Separate analyses were performed for W1 and W2, corresponding to the two experimental sessions.

Table 3 lists the main effects for W1 and W2. Accent had a significant effect on W1 ($F(2,38) = 4.986, p < .05, \eta_p^2 = .208$): seeing the speaker utter W1 with a pitch accent increases the perceived prominence of W1, while seeing the speaker utter W2 with a pitch accent leads to small decrease in perceived prominence of W1. Accent did not have a significant influence when the participants focus on W2 ($F(2,38) < 1$, n.s.). Type of visual beat has a significant influence on the visual difference score for both W1 and W2 ($F(1,19) = 25.570, p < .001, \eta_p^2 = .564$ and $F(1,19) = 5.166, p < .05, \eta_p^2 = .214$, respectively). Inspection of Table 3 reveals that seeing a manual beat gesture has a larger impact than seeing an eyebrow movement. Position is the most interesting main effect, and is also the most consistently strong of the four main effects ($F(1,19) = 14.234, p < .001, \eta_p^2 = .428$ for W1 and $F(1,19) = 18.513, p < .001, \eta_p^2 = .494$ for W2). Seeing

Table 4: *Average visual difference scores as a function of type and position of visual beat, for both W1 and W2 (std. errors between brackets).*

	Scores for W1		Scores for W2	
	W1	W2	W1	W2
Eyebrow	.01 (.20)	-.37 (.11)	.00 (.12)	.17 (.14)
Hand	.99 (.17)	-.11 (.11)	-.33 (.13)	.92 (.16)

a visual beat on W1 increases the perceived prominence of W1 and downscales the perceived prominence of W2, while the reverse holds for seeing a visual beat on W2. The effect of seeing the speaker is the same for both words: seeing speakers S1 and S3 has a small positive effect on the visual difference score, while seeing speaker S2 has a small negative effect. This effect is only significant for W2 ($F(2, 38) = 2.778$, n.s. for W1 and $F(2, 38) = 4.899, p < .01, \eta_p^2 = .494$ for W2).

For both words, a significant two-way interaction between the type of gesture and the position of the gesture was found (for W1: $F(1, 19) = 8.513, p < .01, \eta_p^2 = .309$; for W2: $F(1, 19) = 15.483, p < .001, \eta_p^2 = .449$). This interaction can be explained by looking at the average visual difference scores depicted in Table 4. This table reveals that when participants see a manual beat gesture on the focus word, this clearly increases the perceived prominence of that word, while seeing such a gesture on the other word decreases the perceived prominence of the focus word. The effect of seeing an eyebrow is comparable, albeit less pronounced. Only a few other interactions reached the significance threshold, and these always involve the factor speaker. A closer inspection of the data revealed that these interactions could be attributed to the fact that while the effect of hand gestures were the same for all three speakers, the effects of eyebrow movements seemed to differ per speaker (for one speaker the eyebrows did not seem to have an effect, while for the others it did).

4.3. Conclusion

The second experiment addressed the effects of *seeing* visual beat gestures on prominence perception. Participants had to rate the prominence of both target words (W1, Amanda, and W2, Malta) with and without seeing the speaker. It was found that when participants see a speaker perform a manual beat gesture on a word, the spoken realization of this word is perceived as more prominent than when they do not see the beat gesture. In addition, seeing a manual beat gesture on one word also *decreased* the perceived prominence of the other word. The effect of seeing eyebrow movements was less consistent.

5. Final remarks

In this paper, we have looked at the connections between visual and auditory beats for the production and the perception of prominence. In the first experiment, it was found that visual beats have a significant effect on the spoken realization of the target words (W1, Amanda, or W2, Malta). When a speaker produces a beat gesture while uttering one of these words, the relative spoken prominence of that particular word increases, while the relative prominence of the other word decreases (irrespective of which word carries a pitch accent). This effect holds for all three visual gestures under consideration, which suggests a close connection between auditory and visual cues.

In the second experiment, it was found that when participants *see* a manual beat gesture on a word, they perceive the spoken realization of this word as more prominent than when they do not see the beat gesture. This effect was stronger for the first word than for the second. This might be due to the fact that in Dutch the nuclear ('most important') accent usually comes late in the sentence, an 'early' nuclear accent (i.e., one that occurs in a non-default position) therefore stands out perceptually (see e.g., Krahmer and Swerts 2001). Seeing an eyebrow movement had somewhat similar effects, but much less pronounced, presumably because they are less noticeable. It was interesting to find that visual cues not only increase the perceived prominence of the word they co-occur with, but also reduce the perceived prominence of the other word of interest.

The results from experiment I indicate that visual beat gestures have a noticeable effect on the spoken realization of the associated word. An obvious question is why this is the case. Apparently, the muscular activity required for visual beats leads to increased muscular activity for articulation. It might be that visual beats are governed by the same brain area which also controls articulatory gestures (e.g., Holden 2004). This would be consistent with general theories of movement coordination (e.g., Turvey 1990). Another interesting follow-up question is what the consequences of these results are for the recently proposed models of speaking which treat gesture and speech as closely related systems (see e.g., Kita and Özyürek (2003). The current results are consistent with the conjecture that different kinds of gestures have different functions (Alibali et al. 2001), and might have different sources in a general model for speaking. We hope to address these issues in future research.

6. References

[1] Alibali, M., Heath, D. & Myers, H. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language 44*, 169–188.

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

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

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

[5] Holden, C. (2004). The origin of speech. *Science 303*, 1316–1319.

[6] Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?, *Journal of Memory and Language*, *48*, 16–32.

[7] Krahmer, E. & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*, 391–405.

[8] Krahmer, E. & Swerts, M. (2004). More about brows, In: Zs. Ruttkay and C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents* (pp. 191–216). Dordrecht: Kluwer Academic Press.

[9] McClave, E. (1998). Pitch and Manual Gestures. *Journal of Psycholinguistic Research*, *27*, 69–89.

[10] Turvey, M. (1990). Coordination. *American Psychologist 45*, 938–953.