Acoustic Effects of Visual Beats

Marc Swerts, Emiel Krahmer

Communication and Cognition, Tilburg University, The Netherlands

{M.G.J.Swerts, E.J.Krahmer}@uvt.nl

Abstract

Speakers employ acoustic cues (pitch accents) to indicate that a word is important, but may also use visual cues (such as manual beat gestures, head nods, and eyebrow movements) for this purpose. Even though these acoustic and visual cues are related, the exact nature of this relationship is far from well understood. We investigate whether producing a visual beat leads to changes in how acoustic prominence is realized in speech. For this, we use an original experimental paradigm in which speakers are instructed to realize a target sentence with different distributions of acoustic and visual cues for prominence. Acoustic analyses reveal that the production of a visual beat indeed has an effect on the acoustic realization of the co-occuring speech, in particular on duration and the higher formants (F_2 and F_3), independent of the kind of visual beat and of the presence and position of pitch accents.

Index Terms: audiovisual speech production, gestures, manual beats, facial expressions,

1. Introduction

When a word in an utterance is important, for instance because it expresses new information, a speaker can signal this importance by making the word more prominent than other words in the utterance. Speakers can realize this prominence in a variety of ways, for instance by uttering the word while simultaneously making a manual beat gesture (a quick flick of the hand, Alibali et al. 2001, McNeill 1998) or a facial beat gesture (a rapid eyebrow movement, e.g., Ekman 1979, Krahmer & Swerts 2004; or a head nod, e.g., Hadar et al. 1983), but also by realizing the word with a pitch accent (e.g., Ladd 1996, Swerts et al. 2002), created by what, by analogy, might be called articulatory beat gestures.

It has been suggested that these visual and auditive cues to prominence are related. 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. Morgan (1953) noted that eyebrow movements have a tendency to follow pitch movements. This observation was fleshed out in Bolinger's (1985) "metaphor of up and down", which states that when the pitch rises or falls, the eyebrows go up or down as well. Bolinger stresses that the metaphor of up and down not only applies to eyebrow movements, but to all emphasizing gestures, including manual beat gestures.

Only a few studies have investigated the relation between pitch and (facial or manual) 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; an acoustic correlate for pitch) and the (left) eyebrow movement. They argue that their findings suggest that eyebrow and pitch movements do not coincide due to "muscular synergy", but for "communicative reasons". McClave (1998), in an explicit attempt to verify Bolinger's metaphor 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. On this basis, she concludes that "the correlation is not biologically mandated" (McClave 1998:87). These suggestive but inconclusive findings raise the question what the exact relation between visual beats and speech is, and whether this relation is the same for different kinds of beats.

In this paper, we concentrate on three kinds of visual beats, namely manual beat gestures, head nods and rapid eyebrow movements. Our hypothesis is that the production of visual beats is closely intertwined with speech production, so close, in fact, that the occurrence of a visual beat on a particular word is expected to have a noticeable impact on the speech itself. A research question is what the respective contributions of the three different visual beats are. Several possibilities exist: it might be that eyebrow movements have the biggest impact, since these were found to correlate with speech properties (pitch) in the study by Cavé et al. (1996), while no such correlation was found for manual beat gesture by McClave (1998); on the other hand, it has been claimed that manual gestures and articulatory gestures are controled by the same brain areas (e.g., Holden 2004), and thus the connection between manual beat gestures and prominence in speech might be closer than for facial beat gestures.

To address these issues, we proceeded as follows. We filmed a number of speakers using a novel experimental approach, in which the speakers were instructed to produce a single target sentence in different conditions. The target sentence contained two proper names ("Amanda" and "Malta") that might be marked for prominence, where speakers were instructed to signal this prominence with a pitch accent and/or with a visual beat. In a number of cases speakers were asked to realize the pitch accent and visual beat on the same word, whereas in others cases they were asked to realize them on different words, so that there was a deliberate mismatch (or incongruency) between the auditory and the visual beats. It has been argued that such mismatches are particularly useful when one wants to learn the relative impact of two related factors. According to Goldin-Meadow and Wagner (2005:236), "the best place to explore whether gestures can impart information to listeners is in gesture-speech mismatches." The mismatches Goldin-Meadow and Wagner (2005) refer to arise naturally in spoken communication (of children), but incongruencies between acoustic and visual information have also been used successfully with experimental manipulations as in, for instance, McGurk and MacDonald (1976) or de Gelder and Vroomen (2000), among many others. The current approach is different

from these studies in that we do not use experimental manipulations, but attempt to elicit incongruent utterances directly from speakers.

2. Method

2.1. Participants

For the data collection, 11 speakers were recorded (age 20-45), 3 males and 8 females. Due to missing data from one female participant, we could only analyse data from 10 speakers. They were all students and colleagues from Tilburg University (not involved with the study of audiovisual speech), and none objected to being recorded.

2.2. Materials

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 of 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 a rapid eyebrow movement) on W1 or W2 and with an acoustic pitch accent on W1, W2 or on neither of these.[1] 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*. The tasks were ordered in such a way that the congruent cases, which are assumed to be relatively easy precede the incongruent ones.

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 visual beat 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 and a rapid eyebrow movement, respectively.

2.3. Procedure

The audiovisual recordings of the 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. They were given 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; participants were told that words that were marked with such an icon should be uttered while making the corresponding gesture. Participants were also informed 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.

After the instruction, a training session started, during which speakers were asked to utter the sentence "Pietje gaat naar Polen" (*Little Pete goes to Poland*) in 4 variants of increasing complexity, illustrating all three visual beats, as well as the distinction between congruent and incongruent tasks. Since number of attempts is a potential factor of interest, we used a training sentence that is similar to the target sentence but not identical to it. When the final attempt of a speaker to realize a particular training sentence did not lead to a realization with the intended distribution of visual beats and acoustic accents (which happened rarely), this was pointed out by the experimenter. If the procedure was clear, the actual data collection phase started and there was no further interaction between speaker and experimenter (the latter was not in the visual field of the speaker during the collection phase).

For the collection phase, speakers were given a stack of 18 cards containing the tasks ordered from easy (congruent) to more difficult (incongruent). Speakers were instructed to go through this stack twice (referred to below as the first and second trial). 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.

Table 1: Average number of attempts per task sentence as a function of the trial (first or second), (in)congruency, and kind of visual beat (standard deviations between brackets).

Factor	Level	Number of Tries
Trial	First	1.24 (0.59)
	Second	1.20 (0.56)
Congruency	Congruent	1.11 (0.34)
	Incongruent	1.38 (0.78)
Visual beat	Head nod	1.22 (0.58)
	Eyebrow	1.27 (0.61)
	Manual	1.22 (0.60)

2.4. Data processing

The recordings were made with a digital video camera (MiniDV; 25 frames per second, a resolution of 720$\times$576 pixels, sampling of 4:2:0 (PAL), luma 8 bits chroma and 2 channel audio recording at 16 bits resolution and 48 kHz sampling rate). They were subsequently read and segmented per task.

Table 1 summarizes the number of attempts per task, as a function of trial (first or second one), of (in)congruency, and of kind of visual beat. Overall, the standard deviations are relatively high, which indicates that there is substantial variation among the speakers in the number of attempts they require. Some speakers never used multiple attempts, while others required 1.7 tries on average before they were satisfied with their final realisation. It can be seen that on average, speakers try as much in the first as in the second trial ($t = 0.66$, n.s.), but that they practice more on incongruent than on congruent ones ($t = 2.46, p < .05$). The presence and kind of gestures do not influence the number of attempts.

When a speaker produced multiple attempts for a given task, only the last attempt was selected for further analysis. For each speaker and task, the presence of the intended pitch accent

[1] To avoid a possible confusion: in a few tasks no words were marked for a pitch accent. It is usually assumed that each natural utterance should contain at least one pitch accent, and arguably these tasks are unnatural in this respect. But note that, as argued above, it might be that words that are marked for a visual beat but not for an acoustic one are still accented.

and visual beat was verified, which was indeed the case. This resulted in a corpus of 360 sentences (10 speakers $\times$ 18 tasks $\times$ 2 trials).

To see if and how the speech signal changed as a function of visual beats, we performed an automatic phonetic analysis of the recorded speech using the Praat software package (Boersma & Weenink 2006). For this we proceeded as follows. Since all 360 utterances contain the exact same phonemes, we applied an automatic alignment algorithm (based on a method of dynamic time warping) to mark phoneme boundaries in the wave form. Subsequently, an independent judge performed a manual check on this alignment. This person was unfamiliar with the research question, and performed his checks blind to condition. We then focussed on the /a/ segments in the stressed syllabi of W1 (amAnda) and W2 (mAlta). For each of these two segments, the duration (in seconds), the maximum fundamental frequency (F_0), the maximum values of three higher formants (F_1, F_2, F_3) and the intensity (energy) were measured automatically. It is well-known that syllables carrying a pitch accent are longer and louder than unaccented syllables, and that their F_0 is higher as well (e.g., Ladd 1996). It has been argued that accented syllables also have a noticeable effect on higher formants (e.g., van Bergem 1993). Whether visual beats have any influence on the speech signal is virtually unexplored. Since acoustic prominence is not an absolute property, but is established relative to the context, we use *difference scores* in the analyses, which are computed by subtracting the measured values for W2 from those of W1. Notice that a positive difference score, for duration for instance, indicates that the /a/ in W1 lasts longer than the /a/ in W2, and the other way around for a negative difference score.

2.5. Design and statistical analysis

The experiment has a complete 3 $\times$ 3 $\times$ 2 $\times$ 2 design with the following four factors: Pitch Accent (*no pitch accent, pitch accent on W1 (Amanda), pitch accent on W2 (Malta*), Type of Visual Beat (*head nod, eyebrow movement, manual beat gesture*), Position of the Visual Beat (*on W1, on W2*) and Trial (*First, Second*). For each of the acoustic difference scores (Duration, F_0, F_1, F_2, F_3 and Energy), a four-way Analysis of Variance (ANOVA) test for repeated measures was performed with the aforementioned within-subjects (i.e., speakers) factors. Mauchly's test for sphericity was used to test for homogeneity of variance, and when this test was significant we applied a Greenhouse-Geisser correction on the degrees of freedom (for the purpose of readability we report the normal degrees of freedom in these cases). The Bonferroni correction was applied for multiple pairwise comparisons.

3. Results

Table 2 gives the overall means for the different speakers. Table 3 lists the average difference scores for each of the main effects.

Below we first discuss the effects of auditory and visual beats on the various acoustic difference scores, beginning with an analysis of durational effects. As expected, a significant main effect of accent on duration was found ($F(2, 18) = 9.744, p < .001, \eta_p^2 = .52$). When W1 carried a pitch accent, the /a/ lasts relatively longer than its counterpart in W2, but when W2 carried a pitch accent the opposite holds, with the duration associated with no pitch accent lying in between (means and 95% confidence intervals: for accent on W1 $M = .0058$ $(-.0078, .0195)$, for accent on W2 $M = -.0162$ $(-.0281, -.0043)$, for no accent $M = -.0106$ $(-.0193, -.0019)$). All pairwise comparisons for the three levels no pitch accent, pitch accent on W1, and pitch accent on W2 are statistically significant at the $p < .05$ level, after a Bonferroni correction, with the exception of the comparison between no pitch accent and pitch accent on W2. Interestingly, we also found a significant main effect of position of the visual beat ($F(1, 9) = 16.444, p < .01, \eta_p^2 = .646$). When a visual beat occurred on W2, the /a/ lasts relatively long compared to the /a/ in W1 (for W1: $M = -.0009$ $(-.0117, .0098)$, for W2: $M = -.0131$ $(-.0222, -.0040)$). This effect did not differ for different types of visual beats ($F(2, 18) = 1.8, n.s.$). For duration, no other significant main or interaction effects were found.

Table 2: Average scores for the 10 speakers in terms of Duration (in seconds), F_0, F_1, F_2, F_3 (all in Hz) and Energy (in dB).

Sp.	Dur.	F_0	F_1	F_2	F_3	En.
S1	.0048	-22.7	22.1	57.2	12.6	-.93
S2	.0068	16.1	15.6	-47.2	10.4	4.05
S3	-.0294	08.6	76.6	248.9	172.5	3.81
S4	.0043	-13.4	-13.1	235.6	152.0	3.66
S5	-.0191	4.6	-20.5	6.2	-283.6	-.25
S6	-.0176	-14.3	30.2	120.2	-14.5	1.49
S7	-.0168	33.2	-11.5	-190.6	-105.2	3.80
S8	-.0105	7.9	-.8	265.7	-98.6	1.32
S9	.0039	2.5	55.5	-286.7	-539.5	1.30
S10	.0021	29.7	20.1	66.1	-488.3	2.92
Av.	-.0072	5.2	17.4	47.6	-118.2	2.12

We also found the expected main effect of accent on F_0 ($F(2, 18) = 10.899, p < .001, \eta_p^2 = .548$, after a Greenhouse-Geisser correction on the degrees of freedom). When W1 carries a pitch accent, the peak F_0 in the /a/ of W1 is higher than the one in the /a/ of W2, and vise versa when W2 carries a pitch accent. The scores for the no accent condition lie in between these two (for no accent $M = 8.4$ $(-7.6, 24.2)$, for accent on W1 $M = 30.8$ $(13.6, 48.0)$, for accent on W2 $M = -23.5$ $(-49.3, 2.4)$). After a Bonferroni correction, all pairwise comparisons were significant at $p < .05$, except the one between no accent and accent on W2. The type and position of the visual beat did not have a significant effect on F_0 (in both cases $F < 1$). In fact, of all other main and interaction effects, only the complete 4-way interaction reached the significance threshold ($F(4, 36) = 3.077, p < .05, \eta_p^2 = .255$).

Accent also had a significant main effect on F_1, the first formant ($F(2, 18) = 5.277, p < .05, \eta_p^2 = .370$). This effect could not be attributed to any significant pairwise difference. Type and position of the visual beat did not significantly affect the F_1 difference scores (in both cases $F < 1$), nor was any other of the main or interaction effects statistically significant.

When looking at the F_2, the second formant, we found that accent showed a trend towards significance ($F(2, 18) = 2.909, p = 0.08, \eta_p^2 = .244$). But, interestingly, position of the visual beat revealed a significant main effect ($F(1, 9) = 16.6, p < .01, \eta_p^2 = .648$). When W1 is associated with a visual beat, the F_2 for this word is relatively low, and vise versa for when W2 is associated with a visual beat (for W1 $M = -60.5$ $(-211.3, 90.2)$, for W2 $M = 154.5$ $(15.0, 294.3)$). Notice, incidentally, that this pattern is virtually the same as

Table 3: Acoustic difference scores for Duration (in seconds), F_0, F_1, F_2, F_3 (all in Hz) and Energy (in dB) as a function of pitch accent, type of visual beat, position of visual beat and trial (std. errors between brackets).

Factor	Level	Duration	F_0	F_1	F_2	F_3	Energy
Accent	None	-.0106 (.004)	8.4 (7.0)	-9.9 (19.5)	50.3 (103.1)	-72.4 (88.3)	2.56 (.7)
	W1	.0058 (.006)	30.8 (7.6)	50.9 (12.4)	-66.1 (77.8)	-182.9 (111.1)	6.59 (.9)
	W2	-.0162 (.005)	-23.5 (11.4)	11.3 (10.9)	156.9 (46.6)	-99.2 (74.0)	-2.79 (.9)
Type	Head nod	-.0127 (.005)	4.3 (4.6)	17.2 (14.3)	6.8 (52.9)	-166.0 (64.9)	2.26 (.6)
	Eyebrow	-.0064 (.006)	6.3 (9.2)	22.7 (9.8)	75.3 (74.5)	-66.0 (103.0)	1.92 (.9)
	Hand	-.0019 (.003)	5.1 (6.5)	12.4 (13.1)	58.9 (76.8)	-122.6 (82.0)	2.18 (.5)
Position	W1	-.0009 (.005)	8.8 (6.6)	14.2 (12.5)	-60.5 (66.6)	-176.4 (95.6)	2.72 (.9)
	W2	-.0131 (.004)	1.7 (7.2)	20.6 (15.3)	154.5 (61.7)	-59.9 (69.0)	1.52 (.5)
Trial	First	-.0068 (.005)	6.8 (4.7)	17.5 (11.5)	77.7 (51.9)	-93.8 (80.9)	2.39 (.5)
	Second	-.0071 (.004)	3.7 (7.5)	17.3 (11.6)	16.3 (74.1)	-142.6 (84.0)	1.86 (.7)

the pattern for an accent on W1 and W2. Type of visual beat did not have a significant effect on the second formant ($F < 1$), but one 2-way and one 3-way interaction were significant (between trial and accent: $F(2, 18) = 3.970, p < .05, \eta_p^2 = .306$ and between trial, position and type: $F(2, 18) = 4.236, p < .05, \eta_p^2 = .320$).

For the third formant (F_3), accent did not have a significant effect ($F < 1$), but position of the visual beat approached significance ($F(1, 9) = 3.763, p = .08, \eta_p^2 = .295$). Type of visual beat again did not have a significant effect ($F(2, 18) = 1.437, n.s.$). Only one higher order interaction reached the significance threshold (between trial, position and accent, $F(2, 18) = 3.924, p < .05, \eta_p^2 = .304$).

Finally, for energy a main effect of accent was found ($F(2, 18) = 32.3, p < .001, \eta_p^2 = .782$). When W1 is accented, the energy of the /a/ segment was higher than that of W2, and vice versa when W2 was accented. When none of the words was accented, the energy difference score was in between these extremes (for no accent: $M = 2.56$ $(.96, 4.16)$, for accent on W1: $M = 6.59$ $(4.41, 8.76)$, for accent on W2: $M = -2.79$ $(-4.96, -.62)$). All pairwise comparisons were significant at the $p < .05$ level, after the Bonferroni correction. Position of the visual beat did not have a significant effect on energy ($F(1, 9) = 1.666, n.s.$), nor did type ($F < 1$) nor any other main or interaction effect.

4. Conclusion and Discussion

We analyzed data from 10 speakers realizing the sentence "Amanda gaat naar Malta" with a pitch accent on Amanda, Malta, or neither of these words, and with a visual beat (a manual beat gesture, a head nod, or an eyebrow movement) on Amanda or Malta. We performed acoustic analyses comparing the stressed /a/ in "amAnda" (W1) with the stressed /a/ in "mAlta" (W2). As expected, the presence of a pitch accent on a word resulted in a significant effect of duration (longer), energy (more intense) and F_0 (higher). Moreover, a significant effect for the first formant (F_1) and a trend towards significance for the second formant (F_2) were found. This indicates that our speech materials are 'normal', in the sense that an auditory accent has all the expected acoustic manifestations.

It is very interesting to see that the presence of a visual beat also had several significant effects on the acoustic difference scores. In particular: significant effects were found for duration and F_2, and a trend towards significance for F_3. What is particularly intriguing is that the effects of visual beats on duration and on the second formant are virtually the same as the effects of accents on these two acoustic measures. This can be seen in Figure 1: when a word is produced with either a visual beat or an accent, this word has a relatively longer duration (for a visual beat this holds especially for W2). Similarly, when a word is produced with either a visual beat or an accent, this word has a lower F_2 (recall that a positive difference score indicates that word W1 has a higher F_2, and a negative difference score thus indicates that W1 has a lower F_2, and conversely for W2). This suggests that visual beats have a very similar emphasizing function as accents.

These effects are the same for all three *types* of visual beats. In other words, it does not matter whether the visual beat is a manual beat gesture, a head nod or an eyebrow movement; the acoustic effects are the same. Trial did not have a significant effect, but a handful of significant interaction effects were found which always include this factor. This indicates that some of the acoustic differences are more pronounced in one of the trials.

The effect of visual beat gestures on speech realization has, to the best of our knowledge, not been studied before. However, we did come across one study that, independently and with a rather different experimental set-up, also looked at the influence of gestures on speech, namely Bernardis and Gentilucci (2006). They found, for Italian, that the production of representational gestures (such as waving bye-bye accompanying an utterance of "ciao") had a noticeable impact on the co-produced speech, in particular on the F_2. Interestingly, where they found that gestures lead to an increased F_2, we found a relative decrease. One possible explanation for this difference may involve differences between representational gestures (as studied by Bernardis & Gentilucci 2006) and non-representational gestures (this study). An alternative explanation may involve the fact that the measurements in this study were done on an /a/ phoneme. It has been argued that accentuating certain vowels (including /a/) in Dutch (and English) leads to a reduction of F_2 values (e.g., van Bergem 1993). Notice that this is exactly what we found concerning the effects of pitch accents on F_2. Since beat gestures have a similar accentuation function as pitch accents, it might be that this accounts for the reduction on F_2 which was found to accompany visual beats. It would be very interesting to further investigate this, for instance, by redoing the experiments with target sentences containing various vowels and with both representational and non-representational gestures.

The question naturally arises whether the revealed acoustic

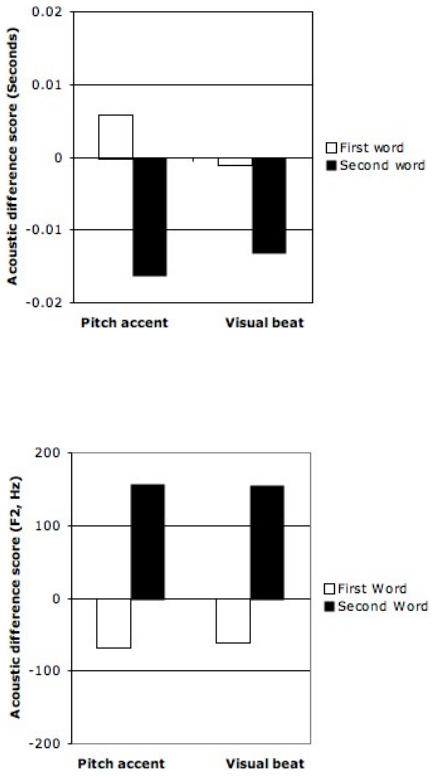

Figure 1: Acoustic effects of the presence of a pitch accent or a visual beat on either the first word (Amanda, W1) or the second one (Malta, W2), on duration (top) and the second formant (F_2) (below).

differences are perceptually relevant. The differences are sometimes rather small, so it is conceivable that listeners would not even notice them. Therefore Krahmer & Swerts (2007) also investigated whether producing a visual beat (either a manual beat gesture, a rapid eyebrow movement or a head nod) has a noticeable influence on the perceived prominence of the associated words, by letting three independent labellers score the perceived prominence of the two target words (blind to condition). It was indeed found that the production of a visual beat has a significant effect on the perceived prominence of the target words (W1, Amanda, or W2, Malta). When a speaker produces a visual beat while uttering one of these words, the perceived *spoken* prominence of that particular word increases, while the perceived prominence of the other word decreases (irrespective of which word carries a pitch accent). The effect is essentially similar for all three visual beats. This suggests that the different types of visual beats are indeed rather similar, and that they all stand in a similar relation to pitch accents. It is interesting to observe that these perception ratings are clearly in line with the acoustic findings from the results presented here, which suggests that these acoustic differences are perceptually relevant.

So it appears that visual beats 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. This would be consistent with general theories of movement coordination (e.g., Turvey 1990, Flanders et al. 1992). Coordination can be seen as a means to make action coherent, and factors such as rhythm (Saltzman & Byrd 2000) and synchronization (Pikovsky, Rosenblum, & Kurths 2001) have been argued to play a role in this. Since the sophisticated motor control of arm movements and of the oral articulators would seem to be handled by the same underlying mechanism (e.g., Flanagan et al. 1990, Hammond 1990), it might well be that extra effort for one kind of gesture spills over into the other. To avoid a possible confusion, note that this is not in contradiction with the claims from McClave (1998) and Cavé et al. (1996) that the relation between pitch accents and visual beats (manual and eyebrows respectively) "is not biologically mandated" (McClave 1998) nor due to "muscular synergy" (Cavé et al. 1996). It is obvious that there is no 1-to-1 mapping between pitch accents and visual beats: speakers vary their pitch more than their manual gestures and their facial expressions, as both McClave (1998) and Cavé et al. (1996) show. Our findings in reveal that *if* a speaker produces a visual beat, this has a clear and noticeable effect on speech production.

5. Acknowledgements

The research described in this paper 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 for her help in collecting the data. Lennard van de Laar and Rob van Son have been tremendously helpful for the acoustic analyses, and Sander Canisius has been very helpful with the data processing. Many thanks also to Carlos Gussenhoven, Bob Ladd, Marie Nilsenova, and Vincent van Heuven for helpful discussions.

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] Bernardis, P. & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia 44*, 178–190.

[3] van Bergem, D. (1993). Acoustic vowel reduction. *Speech Communication*, *12*, 1–23.

[4] Boersma, P. & D. Weenink (2006). *Praat: doing phonetics by computer (Version 4.5.07)*, Retrieved from `http://www.praat.org/`.

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

[6] 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 the International Conference on Spoken Language Processing (ICSLP)* (pp. 2175–2179), Philadelphia.

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

[8] Flanders, M., Helms Tillery, S. & Soechting, J. (1992). Early stages in sensorimotor transformation. *Behavioral and Brain Sciences*, *15*, 309–362.

[9] de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289–311.

[10] 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.

[11] Hadar, U., Steiner, T., Grant, E., & Rose, F. (1983), Head movement correlates to juncture and stress at sentence level, *Language and Speech 26*, 117–129.

[12] Hammond, G. (1990). *Cerebral control of speech and limb movements*. Amsterdam: North-Holland.

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

[14] 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.

[15] Krahmer, E. & Swerts, M (2007), The Effects of Visual Beats on Prosodic Prominence: Acoustic analyses, auditory perception and visual perception, *Journal of Memory and Language*, to appear.

[16] Ladd, D. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

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

[18] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

[19] McNeill, D. (1992). *Hand and Mind: what gestures reveal about thought*. Chicago: University of Chicago Press.

[20] Morgan, B. (1953), Question melodies in American English, *American Speech, 2*, 181-191.

[21] Pikovsky, A., Rosenblum, M., & Kurths, J. (2001), *Synchronization. A Universal Concept in Nonlinear Sciences*. Cambridge: Cambridge University Press.

[22] Saltzman, E. & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, *19*, 499–526.

[23] Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: A comparative analysis. *Journal of Phonetics 30*, 629–654.

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