The effects of visual beats on prosodic prominence

Krahmer, Emiel; Swerts, Marc

Published in:
Journal of Memory and Language

Publication date:
2007

Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 57 (2007) 396–414
Journal of Memory and Language
www.elsevier.com/locate/jml

The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception

Emiel Krahmer *, Marc Swerts

Communication and Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Received 9 January 2006; revision received 1 May 2007
Available online 24 July 2007

Abstract

Speakers employ acoustic cues (pitch accents) to indicate that a word is important, but may also use visual cues (beat gestures, head nods, 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, and whether it leads to changes in how prominence is perceived by observers. For Experiment I (''making beats'') 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-occurring 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. In Experiment II (''hearing beats''), it is found that visual beats have a significant effect on the perceived prominence of the target words. When a speaker produces a beat gesture, an eyebrow movement or a head nod, the accompanying word is produced with relatively more spoken emphasis. In Experiment III (''seeing beats''), finally, it is found that when participants *see* a speaker realize a visual beat on a word, they perceive it as more prominent than when they do not see the beat gesture.

Keywords: Gestures; Beats; Facial expressions; Audio-visual speech; Acoustics; Speech production; Speech perception

Introduction

Speakers have a large repertoire of potential cues at their disposal which they may use to support what they are saying, including gestures and facial expressions. There is a growing awareness that spoken language and manual gestures are closely intertwined (e.g., Goldin-Meadow, 2003; Mayberry & Nicoladis, 2000; Wagner, Nusbaum, & Goldin-Meadow, 2004), as are spoken language and facial expressions or head movements (e.g., Barkhuysen, Krahmer, & Swerts, 2005; Krahmer & Swerts, 2005; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Srinivasan & Massaro, 2003; Swerts & Krahmer, 2005). Still, the exact relation between auditory speech and visual gestures (of face, arm and body) is far from well understood. In this paper, we take a closer look at a particular kind of gesture that has received relatively little attention so far, namely *beats*. We are interested in the effects of these beat gestures on *prominence*, that is, the relative accentual strength with which words are realized in a spoken utterance. More specifically, we look at whether produc-

* Corresponding author. Fax: +31 13 4663110.
E-mail address: E.J.Krahmer@uvt.nl (E. Krahmer).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.06.005

ing a visual beat leads to a change in how prominence is realized in speech, and whether it leads to a change in how prominence is perceived by observers.

That speech and gesture are related is an old observation (McNeill, 1992 refers to Quintilian's *Institutio Oratoria* from 93 AD as an early source), and one that has been made in various disciplines. In work on the origin of speech, for instance, various researchers have suggested that language may originally have been encoded in gestures rather than in speech (e.g., Corballis, 1992; Fitch, 2000; Holden, 2004). This suggestion is based on the claim that the same brain areas control manual gestures and articulatory gestures, and it has indeed been proposed that a single mechanism may account for the underlying control of both manual gestures and oral gestures required for speech (e.g., Flanagan, Feldman, & Ostry, 1990). According to Holden (2004), evolutionary changes in the brain areas that control gestures might be responsible for the development of our language capacity.

In studies of speech perception, to give a second example, gestures have also played an important role. One of the central questions in speech perception is how listeners are able to map acoustic signals to linguistic elements such as phonemes. Three main theoretical perspectives on this issue have been developed in the past 50 years (Diehl, Lotto, & Holt, 2004). Two of these are based on the assumption that listeners recognize speakers' articulatory gestures, such as lip or tongue movements; intended gestures in *motor theory* (e.g., Liberman, 1957; Liberman & Mattingly, 1985) and real gestures in the *direct realist theory* (e.g., Fowler, 1991, 1996). Both these theories have claimed that the fact that human listeners use visual as well as acoustic information in speech perception (e.g., Dodd & Campbell, 1987; Schwartz, Berthommier, & Savariaux, 2004; Tuomainen, Andersen, Tiippana, & Sams, 2005) offers support for a gestural account of speech perception. A prime example of this is the *McGurk effect* (McGurk & MacDonald, 1976) in which an auditory /ba/ combined with a visual /ga/ is perceived as /da/ by most people. Interestingly, the McGurk effect not only works when articulatory gestures are *seen*; Fowler and Dekle (1991) had listeners identify synthetic /ba/ and /ga/ stimuli, while simultaneously touching the mouth of a talker producing either /ba/ or /ga/. Participants could not see the speaker, but still this haptic variant of the McGurk effect gave rise to reliable evidence of cross-modal effects on perception.

More recently, detailed analyses of speakers have confirmed that they produce speech and manual gestures in tandem, and among researchers in this field there appears to be a general consensus that speech and manual gesture should be seen as two aspects of a single process (e.g., Kendon, 1980, 1997; McNeill, 1992). But the jury is still out on *how* speakers co-produce speech and manual gestures. This can be illustrated by comparing various models for the combination of spoken language and manual gestures that were recently proposed, such as those of Kita and Özyürek (2003), Krauss, Chen, and Chawla (1996), and de Ruiter (2000), all based on the speech production model described by Levelt (1989). What these proposals have in common is the addition of a new gesture stream, which has a shared source with the speech production module but is otherwise essentially independent from it. The main difference between the proposed models lies in the location where the two streams (speech and gesture) part. According to Krauss and co-workers, for instance, this happens before conceptualization, while both de Ruiter as well as Kita and Özyürek argue that the separation takes place in the conceptualizer. McNeill and Duncan (2000) take a markedly different view and argue that speech and gesture should not be delegated to different streams, but rather are produced in close connection with each other, based on what they call "growth points". Thus, even though these researchers agree that speech and manual gestures are closely related, they disagree on how tight this relation is.

It is conceivable that different *kinds* of gestures should be integrated in different ways in speech models, although this aspect of speech–gesture interaction is still largely unexplored. In the literature on manual gestures, a distinction is usually made between representational gestures, "gestures that represent some aspect of the content of speech" (Alibali, Heath, & Myers, 2001) and beat gestures that do not represent speech content (see e.g., Alibali et al., 2001; Krauss et al., 1996; McNeill, 1992). Most of the proposed models focus on representational gestures, such as gestures indicating shape ("round") or motion ("upwards"). In fact, Alibali et al. (2001:84) stress "the need for further study of beat gestures and their role in speech production and communication."

A typical beat gesture is a short and quick flick of the hand in one dimension, for example up and down, or back and forth (McNeill, 1992). These gestures look somewhat like the gestures a conductor makes when beating music time (hence their name); they are sometimes also called "batons" (Efron, 1941; Ekman & Friesen, 1969), in reference to the slender rod used by conductors to direct an orchestra. There is an ongoing, general discussion about what, if anything, different kinds of gestures communicate (e.g., Goldin-Meadow & Wagner, 2005). According to Alibali et al. (2001) beat gestures have no semantic content. Still that does not mean that beats are without communicative value. According to McNeill (1992:15), "the semiotic value of a beat lies in the fact that it indexes the word or phrase it accompanies as being significant (...) for its discourse pragmatic content." A beat thus provides extra prominence for a word, for instance, because it expresses new information (McNeill 1992:169–170).

Beat gestures of the form just described (''flick of the hand'') are not the only means speakers have to emphasize words. It has been argued that visual cues such as rapid eyebrow movements (flashes) or head nods can perform a similar function (e.g., Birdwhistell, 1970; Condon, 1976; Eibl-Eibesfelt, 1972; Ekman, 1979; Hadar, Steiner, Grant, & Rose, 1983; Pelachaud, Badler, & Steedman, 1996). In fact, such facial gestures are also referred to as beats (e.g., Ekman, 1979). Of course, emphasis can also be signalled prosodically, for instance via pitch accents (e.g., Cruttenden, 1997; Ladd, 1996; Swerts, Krahmer, & Avesani, 2002 among many others). Even though the exact meaning of pitch accents is a subject of discussion (e.g., Pierrehumbert & Hirschberg, 1990), it is generally assumed that pitch accents, like beat gestures, mark important (or 'significant') words. Indeed, there is some experimental evidence that correct placement of pitch accents (e.g., on new information) helps while incorrect placement (on old information) hinders processing of speech (e.g., Cutler, 1984; Terken & Nooteboom, 1987).

That there appears to be a connection between pitch accents and (manual and facial) gestures has been pointed out various times. One of the earliest who made this connection is Dobogreav, as described in Kendon (1980) and 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. (It is interesting to observe that professional singers often get the advice to raise the eyebrows when trying to reach a high note and to lower them for low notes, Wilson, 1991). Bolinger (1983, 1985) points out 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 at least two questions: is there a different influence of different kinds of visual beats on speech, and how do addressees perceive these beats?

Much work on gesture (including Cavé et al., 1996 & McClave, 1998) primarily addresses how and why speakers *produce* gestures. A number of studies have shown that speakers not only gesture for their own benefit, such as to enhance lexical access (e.g., Rauscher, Krauss, & Chen, 1996), or to support thinking (e.g., Alibali, Kita, & Young, 2000), but also for their hearers (e.g., Alibali et al., 2001; Özyürek, 2002). However, only a few studies (e.g., Cassell, McNeill, & McCullough, 1999) have looked at how addressees actually *perceive* these gestures. Still, both the speaker and addressee perspective are required to gain a full understanding of the interplay between speech and beat gesture during communication.

Here, we concentrate on three kinds of visual beats, namely manual beat gestures, head nods and rapid eyebrow movements. One underlying hypothesis is that speech and beats are indeed closely intertwined, so close, in fact, that the occurrence of a 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, as we have seen, it has been claimed that manual gestures and articulatory gestures are controlled 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.

Moreover, we hypothesize that seeing a gesture increases the perceived prominence of that particular word, and that it decreases the perceived prominence of the other words in the utterance. Again, a research question is whether this effect differs for different visual beats. Naturally, there might be differences in speech production which propagate into speech perception. But, in addition, it might be that different visual beats have different impacts on prominence perception. Manual beat gestures might have a bigger impact than facial gestures, because they might be easier to perceive than facial gestures (the amplitude of a manual gesture is substantially larger than that of an eyebrow movement). Alternatively, it might be that listeners pay special attention to the articulatory area (which is suggested by results on audiovisual speech perception, especially for speech perception in noisy conditions, see e.g., Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998), and since facial gestures are closer to the articulators than manual

gestures it might be that they have a bigger impact on prominence perception. For the same reason, it is expected that seeing a speaker produce an acoustic pitch accent also leads to an increased prominence perception, since there might be general visual correlates of acoustic accents. Keating et al. (2003), for instance, showed that acoustic accents are associated with a clearer visual articulation, while it has also been found that speech sounds louder when participants can look at the speaker, suggesting that audio cues are visually enhanced (Grant & Seitz, 2000; Schwartz et al., 2004).

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 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), Fowler and Dekle (1991), Massaro, Cohen, and Smeele (1996) and 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.

In this paper we describe three experiments. Experiment I (''making beats'') is focussed on the acoustic effects of making beats. We collect data from speakers in an experimental set-up, with pitch accents and three different kinds of visual beats on various positions. The recordings are analysed acoustically, to find out if and how the speech signal changes as a function of auditory and, especially, visual beats. Experiment II (''hearing beats'') concentrates on the auditory perception of words uttered while making a visual beat; three labellers rate the prominence of words with and without an accompanying beat. Experiment III (''seeing beats'') is a visual perception study, looking at the effects of seeing or not seeing visual beats on perceived prominence. In this experiment, participants are offered auditory stimuli with and without the corresponding visual information, and are asked to rate the prominence of the two target words in the utterance.

Experiment I: making beats

Method

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.

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 of studies of prominence and has been used before in studies of speech production and perception for Dutch (e.g., Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997; Rump & Collier, 1996). 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, listed in Appendix A. 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.

[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.

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 in the same order as listed in Appendix A. 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.

Data processing

The recordings were made with a digital video camera (MiniDV; 25 frames per second, a resolution of 720×576 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 realization. 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.

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)

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 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 syllables 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., Gussenhoven, 2004; 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.

Design and statistical analysis

The first 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 4-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.

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
Experiment I: average acoustic difference 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)

Speaker	Duration	F_0	F_1	F_2	F_3	Energy
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
Average	−.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 ($F(2,18) = 2.909,\ p = .08,\ \eta_p^2 = .244$). 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 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

Table 3
Experiment I: 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)

(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.

Summary

For Experiment I we analysed 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 Fig. 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 crucial question is whether these differences are perceptually relevant. The differences are sometimes rather small, so it is conceivable that listeners would not even notice them. Therefore, in Experiment II, we investigate 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. On the basis of the results of Experiment I one would expect that words

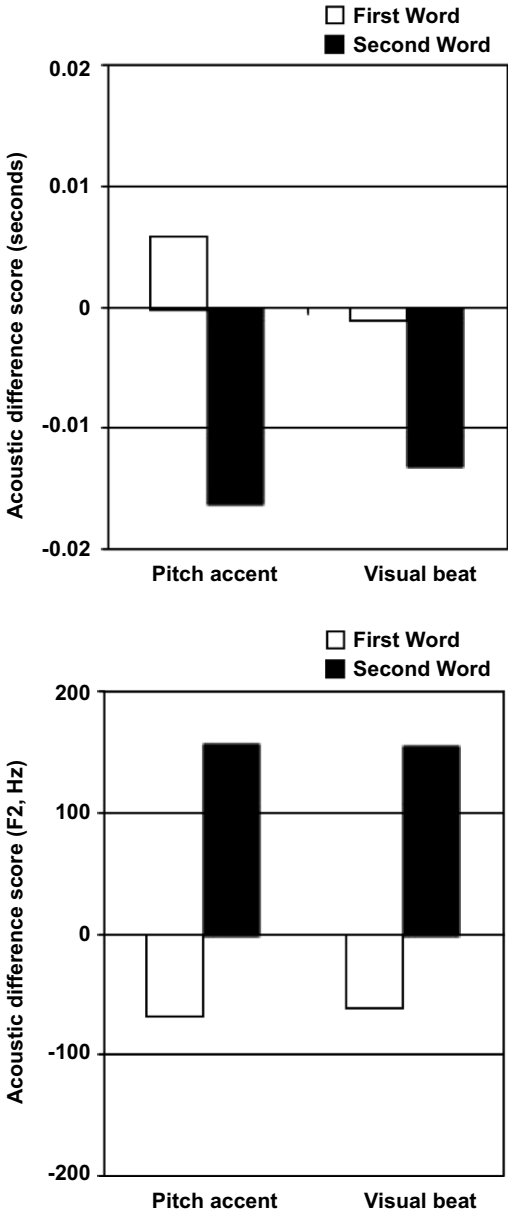

Fig. 1. Experiment I: acoustic effects of the presence of a pitch accent or a visual beat on either the first word (Amanda, W1) or the second one (Malta, W2), on duration (top) and the second formant (F_2) (below).

associated with a visual beat are indeed perceived to be more prominent, and that the effect of visual beats would perhaps be a little smaller than the effect of accent on perceived prominence. Type of visual beat and trial have no effect on the acoustic difference scores, and are therefore less likely to play a role for perceived prominence (if a difference is not there, it cannot be perceived either).

Experiment II: hearing beats

Method

Procedure

All occurrences of W1 (*Amanda*) and W2 (*Malta*) were scored by three independent labellers in terms of prominence, where, following the procedures outlined in Hirschberg, Litman, and Swerts (2004), a 3-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. Labellers could listen to a sentence as often as they desired.

Table 4 shows the Pearson correlations for the accent-scores among the three labellers. In general, the distinction between no accent or accent was easy to make, but the distinction between a minor and a major accent appeared to be more subjective. 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 a *perceived prominence difference score* by subtracting the summed prominence scores for the second word from the summed prominence scores of the first word. This results in a range from −6 to 6, where a positive score indicates that the first word is perceived to be more prominent than the second, while a negative score indicates that the second word is perceived as relatively more prominent.

Design and statistical analysis

The design and statistical analysis for the second experiment are the same as those of Experiment I, except that for Experiment II the dependent variable of the Analysis of Variance (ANOVA) is the (auditory) perceived prominence difference score.

Table 4
Experiment II: agreement (in terms of Pearson correlations) among labellers L_1, L_2 and L_3 for prominence scores [0, no pitch accent; 1, minor accent; 2, clear accent] for words W1 (*Amanda*) and W2 (*Malta*)

	W1			W2		
	L_1	L_2	L_3	L_1	L_2	L_3
L_1	—	0.62*	0.65*	—	0.67*	0.70*
L_2		—	0.58*		—	0.66*
L_3			—			—

* $p < .01$.

Results

In Table 5 the overall results are given, in terms of the auditory perceived difference scores (the raw scores for words W1 and W2 separately can be found in Appendix B). Two things can readily be observed from this table. First, when a pitch accent occurs on W1 this systematically leads to positive difference scores (i.e, W1 is perceived to be more prominent than W2), while a pitch accent on W2 always leads to negative difference scores (W2 is perceived as more prominent than W1). Moreover, when a visual beat occurs on W1, this systematically leads to more positive difference scores than when a visual beat occurs on W2. This indicates that both the position of a pitch accent and the position of a visual beat have an effect on the perceived prominence.

Table 6 lists the main effects from the statistical analysis. As expected, a main effect was found of pitch accent ($F(2,18) = 31.706,\ p < .001,\ \eta_p^2 = .779$, after a Greenhouse-Geisser correction): when a word carries a pitch accent, this word is indeed more prominent than the other. All pairwise comparisons for the three levels no pitch accent, pitch accent on W1, and pitch accent on W2 are statistically significant at the $p < .01$ level, after a Bonferroni correction. Interestingly, there was also a significant main effect of position of the visual beat ($F(1,9) = 15.486,\ p < .01,\ \eta_p^2 = .632$). Overall, when a speaker produces a visual beat, the word uttered while the speaker made this beat is produced with more spoken emphasis, irrespective of the position of the acoustic accent. Neither type of visual beat nor trial had a significant effect ($F < 1$ in both cases), which means that for the perceived prominence 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 beat was a head nod, an eyebrow movement or a manual beat gesture.

Table 6
Experiment II: average perceived prominence difference scores (*P*-diff) as a function of accent, type of visual beat, position of visual beat and trial (std. errors between brackets), with 95% confidence intervals in the last column

Factor	Level	*P*-diff (*SE*)	95%CI
Accent	None	−0.30 (0.17)	(−0.69, 0.08)
	W1	1.77 (0.25)	(1.20, 2.34)
	W2	−1.71 (0.40)	(−2.61, −0.81)
Type	Head nod	0.03 (0.24)	(−0.51, 0.57)
	Eyebrow	−0.12 (0.21)	(−0.59, 0.36)
	Hand	−0.16 (0.19)	(−0.59, 0.28)
Position	W1	0.60 (0.18)	(0.18, 1.02)
	W2	−0.76 (0.26)	(−1.34, −0.18)
Trial	First	0.01 (0.13)	(−0.27, 0.29)
	Second	−0.17 (0.21)	(−0.64, 0.29)

Fig. 2 illustrates the influence of pitch accents and visual beats on the perceived prominence difference score (the results for the different visual beats 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 a pitch accent on W2 results in a negative difference score (and recall that a positive perceived prominence 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 W2, the average difference score is negative. It is highly interesting to find that these two factors

Table 5
Experiment II: overall results

Position	Type	Trial	Pitch accent on		
			W1	None	W2
W1	Head nod	1	2.90 (0.41)	−0.20 (0.70)	−1.10 (0.78)
		2	2.40 (0.73)	1.00 (0.98)	−1.30 (0.21)
	Eyebrow	1	2.90 (0.59)	0.40 (0.82)	−1.20 (0.89)
		2	1.90 (1.15)	0.60 (0.79)	−0.70 (0.83)
	Hand	1	1.80 (0.46)	1.20 (0.65)	−1.70 (0.34)
		2	2.00 (0.58)	1.20 (0.66)	−1.30 (0.47)
W2	Head nod	1	1.00 (0.49)	−0.30 (0.80)	−2.00 (0.67)
		2	1.80 (0.44)	−1.40 (0.79)	−2.40 (0.78)
	Eyebrow	1	1.30 (0.76)	−1.60 (0.48)	−2.10 (0.78)
		2	0.80 (0.74)	−1.50 (0.61)	−2.20 (0.84)
	Hand	1	1.00 (0.52)	−1.00 (0.67)	−1.10 (1.04)
		2	1.40 (0.56)	−2.00 (0.58)	−3.40 (0.30)

Perceived prominence difference scores as a function of pitch accent, type of visual beat, position of visual beat and trial (std. errors between brackets).

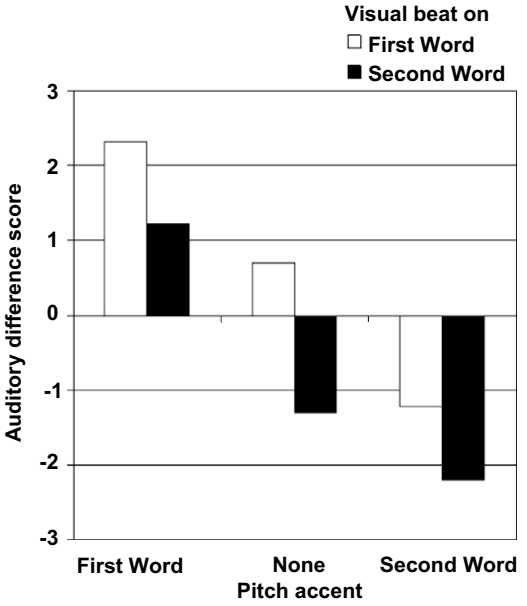

Fig. 2. Experiment II: average perceived prominence difference score as a function of the position of the pitch accent and the visual beat.

are independent; the interaction between accent and position of the visual beat is not significant ($F(2,18) = 1.888$, n.s.). As a result, congruent utterances lead to higher absolute perceived prominence difference scores than incongruent utterances.

In fact, of all possible interaction effects, only one reached significance, this is the 3-way interaction between accent, position of the visual beat and trial ($F(2,18) = 3.619,\ p < .05, \eta_p^2 = .287$). This interaction can be explained by looking at Table 5; it can be seen that the results for the interaction between accent and type of visual beat vary somewhat over the two trials, even though the effects are always in the same direction.

Summary

For Experiment II the auditory recordings from the 10 speakers were analysed, and the perceived prominence of the first (*Amanda*) and second (*Malta*) target words were scored by three independent labellers. On the basis of these scores, difference scores were computed by subtracting the prominence scores for the second word (W2) from those of the first one (W1).

The statistical analyses revealed that the production of visual beats has a clear impact on the perceived prominence of target words. When a speaker makes a visual beat while uttering one of the target words, the relative spoken prominence of that particular word increases, while the relative perceived prominence of the other word decreases. This is true irrespective of which word in the utterance is realized with a pitch accent, and irrespective of the kind of visual beat involved. Put differently: producing a visual beat on a word increases the prosodic prominence of that word. It is interesting to note that the perceived prominence effects are indeed in line with the acoustic results from Experiment I, suggesting that the reason why words with a visual beat are perceived to be more prominent has to do with the longer duration and changes in the higher formants of these words.

While Experiments I and II have focussed on the influence of visual beats on the production of speech, the last experiment focusses on the effects on prominence perception of *seeing* a visual beat.

Experiment III: seeing beats

Method

Participants

Twenty people participated in the third 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. They were not familiar with any of the speakers collected in Experiment I.

Stimuli

Data from three speakers (S3, S7 and S9), recorded during the production study, were used as stimuli for the perception study. These three speakers were selected randomly, but we made sure that their recordings were of a good quality and that both the production of speech and of visual cues was clear and consistent. The overall acoustic scores in Table 2 show that these three are representative for the variation among the speakers, and that their speech is comparable to that of the others. To keep the length of the third experiment manageable, we concentrated on eyebrow movements and manual gestures, as these seem to be the two most different from a visual perception perspective. This implies that 12 different utterances per speaker could be used, which were all selected from the second trial. All selected utterances 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 rated 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". A 10 point scale allows for fine-

grained judgments and, moreover, such a scale is typical of the Dutch school grading system so that all participants are familiar with it. The task was phrased in terms of "emphasis" without any reference to visual beats or pitch accents and their potential role in prominence perception. 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 during two separate experimental sessions in which they either focussed on the first or the second word.

To make sure that participants would look at the computer screen while rating prominence, they were given an additional memory task. Participants were told that following the experiment they would be asked a number of questions about the speakers, and that the person with most questions correct would receive a book token. Sample questions were "what was written on the grey sweater worn by one of the speakers?" and "how many speakers wore earrings?". The results of the memory test were not analysed (other than to find out who won the book token).

Procedure

The experiment was run on a laptop with a 15 in. screen and with separate loudspeakers 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) containing the different visual beats (eyebrow movements, manual gestures) and presentation formats (audiovisual and audio-only). Stimuli were preceded by an auditory beep and a number shown on the screen so that participants knew which stimulus was about to be shown, and followed by a 3 s 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.

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 conditions was always done in a different random order, so that possible learning effects could be compensated for.

In Experiment III 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 focus word, while a negative number indicates that seeing the speaker results in a decrease of perceived prominence for the focus word. Since the speech is the same in both conditions, any positive or negative differences must be attributed to the effect of seeing the speakers and thus their visual beats.

Design and statistical analysis

The third experiment has a complete $3 \times 2 \times 2 \times 3$ design with the following four factors: Pitch Accent (*no pitch accent, pitch accent on W1, pitch accent on W2*), Type of Visual Beat (*eyebrow movement, manual beat gesture*), Position of the Visual Beat (*W1, W2*) and Speaker (*S3, S7, S9*). Two 4-way Analysis of Variance (ANOVA) tests for repeated measures were performed with the aforementioned within-subjects (i.e., observers) factors and with the visual difference score as the dependent variable, one for the cases where participants focus on W1 and one for the cases where they focus on W2. Other than that, statistical analyses are performed as for Experiments I and II.

Results

Table 7 gives an overview of the visual difference score results (the raw audio-visual and audio-only scores from which the visual differences were computed can be found in Appendix B). Table 8 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 a 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 ($F(1,19) = 24.570,\ p < .001,\ \eta_p^2 = .564$ and $F(1,19) = 5.166,\ p < .05,\ \eta_p^2 = .214$, respectively). Inspection of Tables 7 and 8 reveals that seeing a manual beat gesture has a larger impact than seeing an eyebrow movement.

Position of the visual beat 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). Fig. 3 illustrates the main effect of position. Seeing a visual beat on W1 increases the perceived prominence of W1 and downscales the perceived prominence of W2. The reverse holds for seeing a visual beat on W2: when participants see a speaker produce a visual beat on W2, this increases the perceived prominence of W2, and reduces the perceived prominence of W1.

Table 7
Experiment III: overall results

Type	Position	Speaker	Pitch accent on					
			W1		None		W2	
			Scores for		Scores for		Scores for	
			W1	W2	W1	W2	W1	W2
Eyebrow	W1	S3	0.10 (0.45)	0.60 (0.46)	−0.10 (0.37)	0.15 (0.31)	0.35 (0.44)	0.70 (0.31)
		S7	1.60 (0.57)	−0.50 (0.31)	0.00 (0.62)	−0.60 (0.47)	−1.05 (0.42)	−0.75 (0.32)
		S9	−0.10 (0.38)	−0.20 (0.29)	0.45 (0.39)	0.65 (0.44)	−0.35 (0.39)	0.20 (0.31)
	W2	S3	0.25 (0.35)	0.25 (0.47)	−0.40 (0.27)	−0.25 (0.29)	0.80 (0.46)	−0.15 (0.22)
		S7	−0.20 (0.54)	0.30 (0.54)	−0.10 (0.30)	0.40 (0.48)	−2.55 (0.47)	−0.20 (0.34)
		S9	−0.40 (0.39)	0.50 (0.47)	−1.10 (0.43)	0.25 (0.35)	0.30 (0.30)	0.40 (0.37)
Hand	W1	S3	2.00 (0.47)	0.30 (0.36)	0.60 (0.32)	−0.30 (0.39)	1.05 (0.51)	0.25 (0.29)
		S7	0.85 (0.48)	−0.35 (0.27)	0.15 (0.63)	−1.25 (0.57)	0.05 (0.50)	0.10 (0.64)
		S9	1.35 (0.56)	0.25 (0.31)	1.25 (0.50)	−1.35 (0.72)	1.65 (0.41)	−0.65 (0.49)
	W2	S3	0.35 (0.30)	0.55 (0.34)	−1.20 (0.41)	1.45 (0.51)	−1.10 (0.51)	0.85 (0.36)
		S7	0.60 (0.49)	−0.20 (0.49)	0.45 (0.41)	0.00 (0.45)	−0.55 (0.33)	1.65 (0.43)
		S9	0.20 (0.33)	1.85 (0.32)	−0.20 (0.20)	1.70 (0.49)	0.45 (0.34)	0.45 (0.26)

Visual difference scores for W1 and W2 as a function of pitch accent, type of visual beat, position of visual beat and speaker (std. errors between brackets).

Table 8
Experiment III: average visual difference scores (*V*-diff) as a function of accent, type of visual beat, position of visual beat and speaker (std. errors between brackets), with 95% confidence intervals in separate columns

Factor	Level	Scores for W1		Scores for W2	
		V-diff. (*SE*)	95%CI	*V*-diff (*SE*)	95%CI
Accent	None	−0.01 (0.14)	(−0.30, 0.28)	0.07 (0.15)	(−0.24, 0.38)
	W1	0.55 (0.12)	(0.29, 0.81)	0.28 (0.11)	(0.04, 0.52)
	W2	−0.07 (0.18)	(−0.45, 0.29)	0.24 (0.10)	(0.03, 0.44)
Type	Eyebrow	−0.14 (0.10)	(−0.34, 0.07)	0.09 (0.07)	(−0.05, 0.24)
	Hand	0.44 (0.10)	(0.24, 0.65)	0.29 (0.09)	(0.10, 0.49)
Position	W1	0.55 (0.17)	(0.20, 0.90)	−0.15 (0.09)	(−0.33, 0.03)
	W2	−0.24 (0.08)	(−0.41, −0.07)	0.54 (0.12)	(0.29, 0.80)
Speaker	S1	0.22 (0.13)	(−0.04, 0.49)	0.37 (0.10)	(0.16, 0.57)
	S2	−0.06 (0.13)	(−0.33, 0.21)	−0.12 (0.11)	(−0.34, 0.11)
	S3	0.30 (0.11)	(0.06, 0.52)	0.34 (0.15)	(0.02, 0.65)

The effect of seeing the speaker is the same for both words: seeing speakers S3 and S9 has a small positive effect on the visual difference score, while seeing speaker S7 has a small negative effect. This effect is only significant for W2 (for W1: $F(2,38) = 2.778$, n.s., and for W2: $F(2,38) = 4.899, p < .05, \eta_p^2 = .205$).

Table 9 gives a complete listing of all potential interaction effects for W1 and W2. The most interesting one is the 2-way interaction between the type of visual beat and its position, which was significant for both words (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 Fig. 4. This figure reveals that when participants see a manual beat gesture on the focus word, this clearly increases the perceived prominence of that word (these are the positive white and black bars, for W1 and W2, respectively), while seeing such a gesture on the other word decreases the perceived prominence of the focus word (the negative bars). The effect of seeing an eyebrow movement is comparable, albeit less pronounced and only for the second word: seeing an eyebrow movement on the second word leads to a small decrease of perceived prominence for the first word and a small increase of perceived prominence for the second word.

As can be seen in Table 9, a few other interactions reached the significance threshold. It is interesting to observe that these always involve the factor speaker, which suggests that the way the three speakers produce their visual beats differs and that this has an impact on the perceived prominence. A closer inspection of Table 7

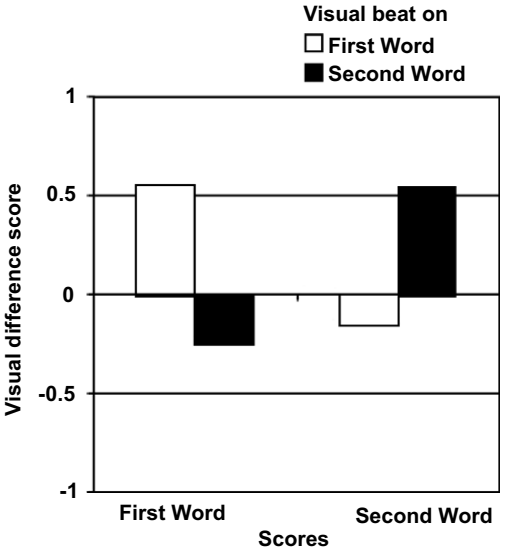

Fig. 3. Experiment III: average visual difference score as a function of position of the visual beat, which could occur either on the first target word (W1) or on the second one (W2). Left the scores for participants focussing on the first target word, right the scores for participants focussing on the second one.

indeed reveals that these interactions can be attributed to such speaker differences. For instance, focussing on the scores for W1 it can be seen that an eyebrow movement of speaker S7 generally has more impact on the visual difference scores than the eyebrow movements of speakers S3 and S9. Conversely, a manual beat gesture of speaker S7 seems to have a lesser impact than a similar movement from the other two speakers. This would explain the significant 2-way interaction between type of visual beat and speaker for W1. Comparable explanations can be given for the other significant interactions. For our current purposes, it is enough to know that speakers differ in how they realize their visual beats, which can have subtle effects on the perception of prominence.

Summary

Experiment III addressed the effects of seeing a visual beat on prominence perception. In two separate sessions, 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 audio-visual AV condition), the spoken realization of this word is perceived as more prominent than when they do not see the beat gesture (the audio-only AO condition). In addition, seeing a beat gesture on one word also *decreased* the perceived prominence of the other word. These effects were stronger for the first word than for the second. Moreover, they are stronger for manual beat gestures than for rapid eyebrow movements. Finally, speakers differ in the way they realize visual beats, and this has subtle effects on the perceived prominence of words.

Since the speech is the same in the AV and the AO condition, the perceived prominence differences that were found must be due to the fact that participants see speakers produce visual beats. So, while Experiments I and II reveal that producing a visual beat has a noticeable impact on the spoken realization of words (longer duration, lower F_2, increased auditory prominence), Experiment III shows that, in addition, seeing such a visual beat increases the perceived prominence.

General discussion

When a word in an utterance is important, for instance because it expresses new or contrastive information, a speaker can signal this by making this

Table 9
Experiment III: overview of all interaction effects for W1 and W2 (effects marked with a † are reported after a Greenhouse-Geisser correction)

Interaction effect	W1	W2
Accent * Type	$F(2,38) < 1$, n.s.	$F(2,38) = 1.144$, n.s.
Accent * Position	$F(2,38) < 1$, n.s.	$F(2,38) = 1.134$, n.s.
Type * Position	$F(1,19) = 8.513, p < .01, \eta_p^2 = .309$	$F(1,19) = 15.483, p < .01, \eta_p^2 = .449$
Accent * Type * Position	$F(2,38) = 1.357$, n.s.	$F(2,38) = 1.720$, n.s.†
Accent * Speaker	$F(4,76) = 7.060, p < .01, \eta_p^2 = .271$†	$F(4,76) = 1.507$, n.s.
Type * Speaker	$F(2,38) = 4.937, p < .05, \eta_p^2 = .206$	$F(2,38) < 1$, n.s.†
Accent * Type * Speaker	$F(4,76) = 2.474$, n.s.†	$F(4,76) = 3.477, p < .05, \eta_p^2 = .155$
Position * Speaker	$F(2,38) < 1$, n.s.	$F(2,38) = 6.978, p < .01, \eta_p^2 = .269$
Accent * Position * Speaker	$F(4,76) = 3.191, p < .05, \eta_p^2 = .144$	$F(4,76) < 1$, n.s.
Type * Position * Speaker	$F(2,38) = 12.034, p < .001, \eta_p^2 = .388$	$F(2,38) = 2.975$, n.s.
Accent * Type * Position * Speaker	$F(4,76) < 1$, n.s.	$F(4,76) = 1.262$, n.s.

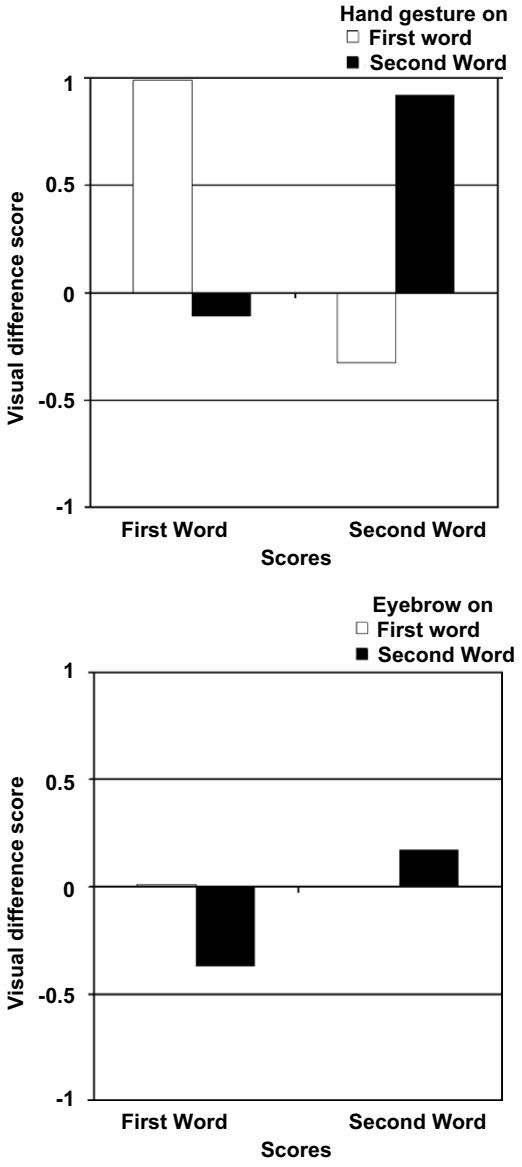

Fig. 4. Experiment III: average visual difference scores as a function of position of visual beat (which could occur either on W1 or on W2), for manual beat gestures (top) and eyebrow movements (bottom). Within each chart, the left hand side represents scores for participants focussing on W1, and the right hand side for participants focussing on W2.

word more prominent than the 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) or a facial beat gesture (a rapid eyebrow movement or a head nod), but also by realizing the word with a pitch accent (created by what, by analogy, might be called articulatory beat gestures).

In this paper, we have looked in detail at the influence of the visual cues on acoustic ones. For this purpose, utterances were gathered from speakers uttering the sentence *Amanda gaat naar Malta* (Amanda goes to Malta) with different distributions of acoustic and visual beats. These could be congruent, with a pitch accent and a visual beat (manual gesture, head nod, or eyebrow movement) on the same word, or incongruent, with a pitch accent on one word and a visual cue on the other. Speakers realized all these utterances in two trials.

In Experiment I (''making beats'') we analysed the speech of 10 speakers to find out whether accent, type of visual beat, position of the visual beat and trial had an effect on acoustic measures of duration, F_0, F_1, F_2, F_3 and energy. As expected, accent had a significant effect on most of these factors, such that words with pitch accent were generally more intense, higher and longer than words without. Interestingly, position of the visual beat also had a number of significant effects, in particular on duration (longer) and the second formant (lower), with a trend towards significance for the third formant.

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 in Experiment I 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.

In Experiment II (''hearing beats'') it was found that visual beats have 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 Experiment I, which suggests that these acoustic differences are perceptually relevant.

In Experiment III ("seeing beats") 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 & Swerts, 2001). Seeing a rapid eyebrow movement had somewhat similar effects, but much less pronounced. 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. Moreover, it was noteworthy that merely *seeing* the speaker realize an acoustic accent on a particular word resulted in an increased prominence perception for that word, confirming observations from Schwartz et al. (2004).

Arguably, one disadvantage of Experiment II is that it is based on the perceived prominence scores of three labellers (albeit ones experienced in intonation labelling). The audio-only scores of Experiment III offer a small scale replication of the labelling for Experiment II (on a one-third subset of the stimuli from three speakers). Hence, as a check, we computed new perceived prominence difference scores for these 36 stimuli by subtracting—per utterance—the average score of the 20 participants for W2 from the average score for W1. Subsequently we computed a correlation between the difference scores thus obtained for Experiment III with the perceived prominence difference scores for the same subset of stimuli from Experiment II. This revealed that there indeed was a strong correlation between the two ($r = .69, p < .001$), which confirms that the labelling was done correctly.

The results from Experiments I and II indicate 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., Bernstein, 1967; Flanders, Helms Tillery, & Soechting, 1992; Turvey, 1990). 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 McGurk and MacDonald (1976) 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 Experiments I and II reveal that *if* a speaker produces a visual beat, this has a clear and noticeable effect on speech production.

Another relevant question is what the consequences of Experiments I and II are for models of speaking. The experiments were not designed to test specific hypotheses about the nature of gesture in speech production, but the findings indicate that, at least for manual beat gestures, there is indeed a very close connection between speech and gesture. This close connection between speech and gesture might be somewhat easier to explain in the context of McNeill and Duncan's (2000) integrated approach than for models in which gesture and speech form essentially separate streams with a shared origin. The current results are also consistent with the conjecture that gestures with different functions may have different sources in a general model for speaking (e.g., manual beat gestures arising relatively late, say in the formulator, while representational gestures may arise in some earlier stage). Suggestive evidence for such a model might be found if it were the case that beat gestures have a stronger influence on speech production than representational gestures, and we hope to address this question in future research.

Several other lines for future research suggest themselves. Our experiments were based on data from speakers who were instructed to realize a particular sentence in a number of different ways. This approach has clear advantages as it allows for far more experimental control than unsolicited data would do, and the use of incongruent stimuli enables us to separate the influences of acoustic and visual

cues to prominence. But a potential downside of this approach is that some of the tasks speakers had to utter (in particular the incongruent ones) are less natural than the others (but note that our findings apply to the congruent and the incongruent tasks alike). The question naturally arises whether the interdependencies between visual beats and prosodic prominence in the two experiments are likely to be a feature of normal communication as well. We conjecture that this is indeed the case. One group of suggestive evidence in this direction comes from the work of Kelso and colleagues who show that speakers who simultaneously perform a finger tapping task while speaking increase stress with longer finger movements (Kelso & Holt, 1980; Kelso, Tuller, & Harris, 1983), suggesting a further link between the subsystems of speaking and manual performance. Such a link is also revealed by Hiscock and Chipuer (1986), who show that finger tapping slows down speech, both when the tapping rhythm is compatible (congruent) and when it is incompatible with the rhythmic structure of the sentence (incongruent). In a similar vein, various studies have suggested that visual cues have an impact on prominence perception. Krahmer and Swerts (2004) report on a series of experiments using a virtual computer character with Dutch and Italian participants, showing that eyebrow movements boost the perceived prominence of the word they co-occur with, while they downscale the prominence of the words in the immediate context, and Glave and Rietveld (1979), in a different setting, showed that perceived speech loudness increases when participants see the speaker. Nevertheless, it would be interesting to supplement the current findings with data about gestures (both manual and facial) in spontaneous speech, although it is difficult to see how incongruent utterances could be triggered naturally. It is also worth pointing out that the connection between pitch accents and visual beats is likely to be language and/or culture dependent. In particular, it can be hypothesized that a similar connection is less likely to be found in languages such as Japanese or certain dialects of Basque, where accent is lexically determined and not associated with communicative prominence. It would be worthwhile to test this.

Finally, it would be interesting to gain further insights in how addressees *process* visual beats. In Experiment III it was shown that seeing visual beat gestures leads to an increase in perceived prominence. From the work of Cutler (1984), Terken and Nooteboom (1987) and others it is known that the correct placement of pitch accents (on 'important words') helps processing while incorrect placement (on 'non-important' words) does not. Given the similarities between acoustic and visual beats, an interesting question is whether correct placement of visual beats facilitates (speeds up) processing in a similar way, and whether incorrect placement similarly hinders processing. This question is addressed in Swerts and Krahmer (2007), where we also experiment with covering parts of the visual stimuli to find out what the relative contributions of different face parts are for prominence perception.

Acknowledgments

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 our statistical consultants Carel van Wijk and Edwin Commandeur for some highly useful assistance and discussion, and to Bob Ladd and Vincent van Heuven for helpful discussions on the acoustic analyses. Finally, we thank Carlos Gussenhoven, Marie Nilsenova and the anonymous reviewers for detailed and constructive comments on previous versions of this paper.

Appendix A. List of stimuli

This appendix contains the list of stimuli used for the data collection, in order of actual usage. Stimuli may contain an acoustic and/or a visual cue (a manual beat gesture, a head nod or a rapid eyebrow movement), positioned on either "Amanda" (W1) or "Malta" (W2)

Task	Cue position		Type of visual beat
	Acoustic	Visual	
1.	—	Amanda	Manual
2.	—	Amanda	Head nod
3.	—	Amanda	Eyebrow
4.	—	Malta	Manual
5.	—	Malta	Head nod
6.	—	Malta	Eyebrow
7.	Amanda	Amanda	Manual
8.	Amanda	Amanda	Head nod
9.	Amanda	Amanda	Eyebrow
10.	Malta	Malta	Manual
11.	Malta	Malta	Head nod
12.	Malta	Malta	Eyebrow
13.	Amanda	Malta	Manual
14.	Amanda	Malta	Head nod
15.	Amanda	Malta	Eyebrow
16.	Malta	Amanda	Manual
17.	Malta	Amanda	Head nod
18.	Malta	Amanda	Eyebrow

Appendix B. Raw scores for Experiments II and III

The table below lists the raw scores for Experiment II: the average perceived prominence scores for both W1 and W2, as a function of pitch accent, type and position of visual beat and trial

Position	Type	Trial	Pitch accent on					
			W1		None		W2	
			Scores for		Scores for		Scores for	
			W1	W2	W1	W2	W1	W2
W1	Head nod	1	5.70	2.80	4.10	4.30	4.20	5.30
		2	5.60	3.20	4.40	3.40	4.50	5.80
	Eyebrow	1	5.70	2.80	4.30	3.90	4.20	5.40
		2	5.30	3.40	4.30	3.70	4.50	5.20
	Hand	1	5.40	3.60	4.80	3.60	4.10	5.80
		2	5.50	3.50	4.70	3.50	4.00	5.30
W2	Head nod	1	5.90	4.90	4.10	4.40	3.70	5.70
		2	5.90	4.10	3.40	4.80	3.10	5.50
	Eyebrow	1	5.70	4.40	3.30	4.90	3.40	5.50
		2	5.10	4.30	3.40	4.90	3.30	5.50
	Hand	1	5.80	4.80	3.70	4.70	3.70	4.80
		2	5.80	4.40	3.00	5.00	2.60	6.00

These scores range from 0 (“no pitch accent according to all 3 labellers”) to 6 (“a major pitch accent according to all 3 labellers”).

The following table lists the raw scores for Experiment III: these are the average audiovisual (AV) and audio-only (AO) scores for W1 and W2 as a function of pitch accent, type of visual beat and position of visual beat

Type	Position	Pitch accent on											
		W1				None				W2			
		Scores for				Scores for				Scores for			
		W1		W2		W1		W2		W1		W2	
		AV	AO	AV	AO	AV	AO	AV	AO	AV	AO	AV	AO
Eyebrow	W1	6.10	5.57	5.12	5.15	4.63	4.52	5.00	4.93	4.90	5.25	7.18	7.13
	W2	6.03	6.15	5.73	5.38	4.15	4.67	5.58	5.45	4.52	5.00	6.95	6.93
Hand	W1	7.03	5.63	4.48	4.42	6.05	5.38	4.36	5.33	5.67	4.75	6.63	6.73
	W2	7.40	7.02	6.02	5.28	4.28	4.60	6.30	5.25	3.98	4.40	7.82	6.83

These scores range from 1 (“no prominence”) to 10 (“strong prominence”).

References

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.

Alibali, M., Kita, S., & Young, A. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*, 593–613.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). Problem detection in human–machine interactions based on facial expressions of users. *Speech Communication, 45*, 343–359.

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

Bernstein, N. (1967). *The coordination and regulation of movements*. London: Pergamon.

Birdwhistell, R. (1970). *Kinesics and context*. University of Pennsylvania Press.

Boersma, P. & Weenink, D. (2006). *Praat: Doing phonetics by computer (Version 4.5.07)*, <http://www.praat.org/>.

Bolinger, D. (1983). Intonation and gesture. *American Speech, 58*, 156–174.

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

Cassell, J., McNeill, D., & McCullough, K. (1999). Speech–gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognition, 7*, 1–33.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eye-

brow movements and F_0 variations. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (pp. 2175–2179). Philadelphia.

Condon, W. (1976). An analysis of behavioral organization. *Sign Language Studies, 13*, 285–318.

Corballis, M. (1992). On the evolution of language and generativity. *Cognition, 44*, 197–226.

Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge: Cambridge University Press.

Cutler, A. (1984). Stress and accent in language production and understanding. In D. Gibbon & H. Richter (Eds.). *Intonation, accent and rhythm studies in discourse phonology* (pp. 77–90). de Gruyter Berlin.

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

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge: Cambridge University Press.

Diehl, R., Lotto, A., & Holt, L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.

Dodd, B., & Campbell, R. (1987). *Hearing by eye: The psychology of lip-reading*. New Jersey: Lawrence Erlbaum Associates.

Efron, D. (1941). *Gesture and environment*. New York: King's Crown Press.

Eibl-Eibesfelt, I. (1972). Similarities and differences between cultures in expressive movements. In R. Hinde (Ed.), *Non-verbal communication*. Cambridge: Cambridge University Press.

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.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavioral categories: Origins, usage, and coding. *Semiotica, 1*, 49–98.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Science, 4*, 258–267.

Flanagan, R., Feldman, A., & Ostry, D. (1990). Control of human jaw and multi-joint arm movements. In G. Hammond (Ed.), *Cerebral control of speech and limb movements* (pp. 29–58). Amsterdam: North-Holland.

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

Fowler, C. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America, 88*, 1236–1249.

Fowler, C. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*, 1730–1741.

Fowler, C., & Dekle, D. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 877–888.

Glave, R., & Rietveld, A. (1979). Bimodal cues for speech loudness. *Journal of the Acoustical Society of America, 66*, 1018–1022.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S., & Wagner, S. (2005). How our hands help us learn. *Trends in Cognitive Sciences, 9*, 234–240.

Grant, K., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*, 1197–1208.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

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.

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.

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

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication, 43*, 155–175.

Hiscock, M., & Chipuer, H. (1986). Concurrent performance of rhythmically compatible or incompatible vocal and manual tasks: Evidence for two sources of interference in verbal-manual timesharing. *Neuropsychologia, 24*, 691–698.

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

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., & Auer, E. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. In Proceedings 16th International Conference of the Phonetic Sciences (ICPhS) (pp. 2071–2074). Barcelona, Spain.

Kelso, J., & Holt, K. (1980). Exploring a vibratory systems account of human movement production. *Journal of Neurophysiology, 43*, 1183–1196.

Kelso, J., Tuller, B., & Harris, K. (1983). A "Dynamic Pattern" perspective on the control and coordination of movement. In P. MacNeilage (Ed.), *The production of speech* (pp. 137–173). New York: Springer Verlag.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). The Hague: Mouton.

Kendon, A. (1997). Gesture. *Annual Review of Anthropology, 26*, 109–128.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and gesture. *Journal of Memory and Language, 48*, 16–32.

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

Krahmer, E., & Swerts, M. (2004). More about brows. In Zs. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating embodied conversational agents* (pp. 191–216). Dordrecht: Kluwer Academic Press.

Krahmer, E., & Swerts, M. (2005). How children and adults signal and detect uncertainty in audiovisual speech. *Language and Speech, 48*, 29–54.

Krauss, R., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in*

experimental social psychology (pp. 389–450). San Diego: Academic Press.

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

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge: MIT Press.

Liberman, A. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29*, 117–123.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Massaro, D., Cohen, M., & Smeele, P. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 100*, 1777–1786.

Mayberry, R., & Nicoladis, E. (2000). Gesture reflects language development: Evidence from bilingual children. *Current Directions in Psychological Science, 9*, 192–196.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27*, 69–89.

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

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D., & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture*. Cambridge: Cambridge University Press.

Morgan, B. (1953). Question melodies in American English. *American Speech, 2*, 181–191.

Munhall, K., Jones, J., Callan, D., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science, 15*, 133–137.

Özyürek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46*, 688–704.

Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science, 20*, 1–46.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication*. Cambridge, MA: MIT Press.

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization. A universal concept in nonlinear sciences*. Cambridge: Cambridge University Press.

Rauscher, F., Krauss, R., & Chen, U. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226–231.

Rump, H., & Collier, R. (1996). Focus conditions and the prominence of pitch accented syllables. *Language and Speech, 39*, 1–17.

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

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*, B69–B78.

Srinivasan, R., & Massaro, D. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech, 46*, 1–22.

Swerts, M. & Krahmer, E. (2007). Facial expressions and prosodic prominence: Comparing modalities and facial areas. *Journal of Phonetics*, in press.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53*, 81–94.

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.

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

Tuomainen, J., Andersen, T., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*, B13–B22.

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

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

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*, 926–940.

Wagner, S., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language, 50*, 395–407.

Wilson, G. B. (1991). Three Rs for vocal skill development in the choral rehearsal. *Music Educators Journal, 77*, 42–46.