Title: The Effects of Visual Beats on Prosodic Prominence

Authors: Marc Swerts and Emiel Krahmer

Affiliation: Communication and Cognition, Tilburg University

Running head: Beats and Prominence

Full address:
Marc Swerts
Communication and Cognition
Faculty of Arts
Tilburg University
P.O.Box 90153
NL-5000 LE Tilburg
The Netherlands
e-mail: M.G.J.Swerts@uvt.nl
phone: +31 13 4663070
fax: +31 13 4663110

Abstract

Speakers can employ a variety of means to indicate that a word is important, including acoustic cues such as pitch accents but also visual cues such as manual beat gestures, head nods and rapid eyebrow movements. Even though it has been noted that these acoustic and visual cues are related, the exact nature of this relationship is far from well-understood. In this paper, we look at the influence of the visual cues on the acoustic ones, based on data collected in an original experimental paradigm in which speakers were instructed to realize a particular target sentence with different distributions of acoustic and visual cues for prominence. In Experiment I ("hearing beats"), it is found that visual beats have a significant effect on the spoken realization of the target words. When a speaker produces a manual beat gesture, an eyebrow movement or a head nod, the word uttered while making this visual beat is produced with relatively more spoken emphasis, irrespective of the position of the acoustic accent. In Experiment II ("seeing beats"), 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, audiovisual speech, speech production, speech perception

Introduction

Speakers have a large repertoire of manual gestures and facial expressions at their disposal which they may use to support what they are saying. There is a growing awareness that spoken language and manual gestures are closely intertwined (e.g., Goldin-Meadow 2003, Mayberry & Nicoladis 2000, Wagner et al. 2004), as are spoken language and facial expressions (e.g., Barkhuysen et al. 2005, Krahmer & Swerts 2005, Munhall et al. 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*. In particular, 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.

That speech and gesture are related is an old observation (McNeill 1992 refers to Quintilian's *Institutio Oratoria* from 93 CE 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., Corbalis 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 et al. 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 et al. 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 1986, Schwartz et al. 2004, Tuomainen et al. 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 does not only work 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 any visual information from 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 et al. (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 modules 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, in working memory, while both de Ruiter and 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 et al. 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 different kinds of gestures communicate, if anything, (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 facial gestures such as rapid eyebrow movements (flashes) or head nods can perform a similar function (e.g., Birdwhistell 1970, Condon 1976, Eibl-Eibesfeldt 1972, Ekman 1979, Hadar et al. 1983, Pelachaud et al. 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 et al. 2002 among many others). Even though the exact meaning of pitch accents is 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 (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, to continue the musical motif of this introduction, 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 arm) gestures empirically. Cavé et al. (1996), for instance, report on a pilot production study with a limited number of speakers and they indeed found a significant correlation between fundamental frequency (F_0; 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 and 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 et al. 1996), or to support thinking (e.g., Alibali et al. 2001), but also for their hearers (e.g., Alibali et al. 2002, Özyürek 2002). However, only a few studies (e.g., Cassell et al. 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.

We will be studying three kinds of visual beat, 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), while no such correlation was found for manual beat gestures in the aforementioned studies of Cavé et al. (1996) and McClave (1998); on the other hand, as we have seen, 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.

Moreover, we hypothesize that seeing a gesture will increase the perceived prominence of that particular word, and that it will decrease the perceived prominence of the other words in the utterance. Again, a research question is whether this effect will differ 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), 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 will also lead 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. First, we collected audiovisual materials using an experimental approach, in which a number of speakers were instructed to produce a single target sentence in different conditions. The target sentence contained two proper names 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 variants the acoustic and visual prominence cues coincided, whereas in others there was a deliberate mismatch (or incongruency) between the two. 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 & 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 et al. (1996) and de Gelder and Vroomen (2000), among many others. The current approach is different from these studies in that we will not use experimental manipulations, but attempt to elicit incongruent utterances directly from speakers.

Following the data collection, two experiments were conducted. In the first one ("hearing beats"), the influence of the three kinds of visual beat gestures on the realization of spoken prominence was addressed. In the second ("seeing beats"), the effects of visual beats on perceived prominence were studied in a perception experiment during which participants were offered auditory stimuli with and without the corresponding visual information, and were asked to rate the prominence of the two target words in the utterance.

Data collection

Participants

For the data collection, 11 speakers were recorded (age 20-45), 3 males and 8 females. 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 et al. 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 beat gesture were marked with a specific icon illustrating a hand, a head or an eye plus eyebrow as markers for a manual beat gesture, a head nod and a rapid eyebrow movement respectively.

Procedure

The audiovisual recordings of the 11 speakers were made in a research laboratory at Tilburg University. Speakers were seated on a chair in front of a digital camera that recorded their upper body and face (25 fps). They were given a brief instruction, explaining the experimental setup and the task representations on the cards. They were told that only a word in bold face should be emphasized in speech. In addition, the three gesture icons (for head nod, eyebrow movement and manual gesture) were explained by the experimenter, and the intended gestures were illustrated; again participants were told that words that were marked with such an icon should be uttered while making the corresponding gesture. Participants were told that they might find some of the tasks difficult

[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.

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 tries is a factor of interest in the analyses, 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 trial 1 and trial 2). They were asked to first read the task on the card, and then utter the sentence with the required distribution of beat gestures and pitch accents, using as many attempts as they felt necessary.

Data processing and summary statistics

The video recordings were read into the computer and segmented per task.

Table 1 approximately here.

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

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 396 sentences (11 speakers × 18 tasks × 2 trials).

Experiment I: Hearing Beats

In the first experiment we look whether producing a visual beat (either a manual beat gesture, a rapid eyebrow movement or a head nod) has a noticeable influence on the production of speech.

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 et al. (2004), a three-way distinction was made: a word was assigned a 0 if no pitch accent was noticed, a 1 if a minor pitch accent was heard and a 2 for a clear pitch accent. Labelling was performed individually on the basis of only the audio signal. Sentences were played in a random order, so that the labellers were always blind to condition, in order to avoid circularity. Labellers could listen to a sentence as often as they desired.

Table 2 approximately here.

Table 2 shows the Pearson correlations for the accent-scores among the three labellers. Notice that the agreement is somewhat higher for second than for first word. 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 an *auditory difference score* by subtracting the summed prominence scores for the second word from the summed prominence scores of the first word. This results in range from -6 to 6, where a positive score indicates that the first word is relatively more prominent than the second, while a negative score indicates that the second word is relatively more prominent. The use of difference scores is motivated from the fact that prominence is not an absolute property, but is established relative to the context.

Design and statistical analysis

The first experiment has a complete 3 × 3 × 2 × 2 design with the following four factors: Pitch Accent (*no pitch accent, pitch accent on W1, pitch accent on W2*), Type of Visual Beat (*head nod, eyebrow movement, manual beat gesture*), Position of the Visual Beat (*W1, W2*) and Trial (*first, second*). A four-way Analysis of Variance (ANOVA) for repeated measures was performed with the aforementioned within-subjects (i.e., speakers) factors and with the auditory difference score as the dependent variable. Mauchy's test for sphericity was used to test for homogeneity of variance, and the Bonferroni correction was applied for multiple pairwise comparisons.

Results

TABLE 3 APPROXIMATELY HERE.

The main effects are described in Table 3. As expected, a main effect was found of pitch accent[2] ($F(2,18) = 31.706, p < .001, \eta_p^2 = .779$): when a word had to be emphasized (i.e., was printed in bold face on the task card), 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 making this beat is produced with relatively more spoken emphasis, irrespective of the position of the acoustic accent. Neither type of visual accent nor trial had a significant effect ($F < 1$ in both cases), which means that for the auditory difference score it does not matter whether the target utterance was produced in the first round or in the second round, nor does it matter whether the visual beat was a head nod, an eyebrow movement or a manual beat gesture.

Table 4 approximately here.

Table 4 illustrates the influence of pitch accents and visual beats on the auditory 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 auditory difference score indicates that the first word is relatively more prominent, while a negative score indicates that the second word is more prominent). The same can be observed for the visual beats: if one of these occurs on W1, the difference score is positive on average and if one occurs on W2, the average difference score is negative. It is highly interesting to find that

[2] Mauchy's test for sphericity was significant for this factor, hence we looked at the Greenhouse-Geisser and Huynh-Feldt corrections on the degrees of freedom, which resulted in p-values of less than .001 in both cases. For the sake of transparency, we report on the normal degrees of freedom.

these two effects are independent (there is no significant interaction between the two factors). As a result, congruent utterances lead to higher absolute auditory difference scores than incongruent utterances.

Discussion

The first experiment revealed that the production of visual beats has a clear impact on the *spoken realization* of target words. When a speaker makes a visual beat while uttering the first or second word of interest (i.e., Amanda or Malta), the relative spoken prominence of that particular word increases, while the relative spoken prominence of the other word decreases. This is true irrespective of which word in the utterance is realized with a pitch accent, and irrespective of the kind of visual beat involved.

In the next experiment, the effects of seeing a visual beat on prominence perception will be addressed.

Experiment II: Seeing Beats

Method

Participants

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

Stimuli

Data from three speakers, recorded during the production study, were used as stimuli for the perception study. These three speakers were selected because their recordings were of a good quality and because they spoke most clearly throughout the production phase. To keep the length of the second 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 stimuli per speaker could be used, which were all selected from the second trial. All fragments were offered in two variants to the participants: an audiovisual variant (i.e., as original recordings) and an audio-only variant (with a black screen). In total, we used 72 stimuli (3 speakers × 12 utterances × 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 inch screen and with separate loud speakers positioned to the left and right of the computer. The experiment was individually performed. After participants were instructed about the goal of the experiment (prominence perception), a brief training session started, consisting of 4 stimuli (from a fourth speaker not used in the actual experiment) illustrating the different gestures (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 second interval in which a white screen was displayed and during which participants could rate the prominence of the target word on an answer form. If participants had no questions about the procedure, the actual experiment started and there was no further interaction between participant and experimenter.

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

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

Design and statistical analysis

The second experiment has a complete 3 × 2 × 2 × 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 (*S1, S2, S3*). A four-way Analysis of Variance (ANOVA) for repeated measures was performed with the aforementioned within-subjects (i.e., observers) factors and with the visual difference score as the dependent variable. Mauchy's test for sphericity was used to test for homogeneity of variance, and the Bonferroni correction was applied for multiple pairwise comparisons.

Results

Table 5 approximately here.

Table 5 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 Table 5 reveals that seeing a manual beat gesture has a larger impact than seeing an eyebrow movement. Position is the most interesting main effect, and is also the most consistently strong of the four main effects ($F(1, 19) = 14.234, p < .001, \eta_p^2 = .428$ for W1 and $F(1, 19) = 18.513, p < .001, \eta_p^2 = .494$ for W2). Seeing a visual beat on W1 increases the perceived prominence of W1 and downscales the perceived prominence of W2, while the reverse holds for seeing a visual beat on W2. The effect of seeing the speaker is the same for both words: seeing speakers S1 and S3 has a small positive effect on the visual difference score, while seeing speaker S2 has a small negative effect. This effect is only significant for W2 (for W1: $F(2, 38) = 2.778$, n.s., and for W2: $F(2, 38) = 4.899, p < .05, \eta_p^2 = .205$).

Table 6 approximately here.

For both words, a significant two-way interaction between the type of gesture and the position of the gesture was found (for W1: $F(1, 19) = 8.513, p < .01, \eta_p^2 = .309$; for W2: $F(1, 19) = 15.483, p < .001, \eta_p^2 = .449$). This interaction can be explained by looking at the average visual difference scores depicted in Table 6. This table reveals that when participants see a manual beat gesture on the focus word, this clearly increases the perceived prominence of that word, while seeing such a gesture on the other word decreases the perceived prominence of the focus word. The effect of seeing an eyebrow is comparable, albeit less pronounced. Only a few other interactions reached the significance threshold, and these always involve the factor speaker. A closer inspection of the data revealed that these interactions could be attributed to the fact that while the effect of hand gestures were the same for all three speakers, the effects of eyebrow movements seemed to differ per speaker (for one speaker the eyebrows did not seem to have an effect, while for the others it did).

Discussion

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

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 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, data were collected from 11 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.

In Experiment I ("hearing beats") It was found that visual beats have a significant effect on the spoken realization of the target words (W1, Amanda, or W2, Malta). When a speaker produces a visual beat while uttering one of these words, the relative *spoken* prominence of that particular word increases, while the relative spoken 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 kinds of of visual beats are indeed rather similar, and that they all stand in a similar relation to pitch accents. It might be that visual beats are governed by the same brain area which also controls articulatory gestures (which would be consistent with Holden 2004) and it would be interesting to investigate this.

In Experiment II ("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 and 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), but with data from real speakers.

The results from Experiment I 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, 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 et al. 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 Experiment I reveal that *if* a speaker produces a visual beat gesture, this may have a clear and noticeable effect on speech production.

Another relevant question is what the consequences of Experiment I are for models of speaking. The experiment was 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 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 different kinds of gestures have different functions (Alibali et al. 2002), and might 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. Experiments I and II 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, Tuller & Harris 1983, Kelso & Holt 1980), 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 experiment 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 (1975), 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 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 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 II 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 Krahmer and Swerts (2006), 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 conducting the experiments and to Lennard van de Laar for technical assistance. Many thanks also to our statistical consultants Carel van Wijk and Edwin Commandeur for some highly useful assistance. Finally, we would like to thank Carlos Gussenhoven and the anonymous reviewers for detailed and constructive comments on a previous version of this paper.

Appendix

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).

	Cue Position		
Task	Acoustic	Visual	Type of Visual Beat
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

References

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.

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.

Barkhuysen, P., Krahmer, E. & Swerts, M. (2005). Problem Detection in Human-Machine Interactions based on Facial Expressions of Users. *Speech Communication*, *45*, 343–359.

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

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

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

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

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review*, *96*, 168–174

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 eyebrow movements and F_0 variations. *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.

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

Cutler, A. (1984). Stress and accent in language production and understanding. In: D. Gibbon and H. Richter (eds.), *Intonation, accent and rhythm. Studies in Discourse Phonology* (pp. 77–90). Berlin: de Gruyter.

Cruttenden, A. (1997). *Intonation* (2nd edition). 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 Lipreading.* 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.

Frick-Horbury, D. (2002). The use of hand gestures as self-generated cues for recall of verbally associated targets. *American Journal of Psychology 115*, 1–20.

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

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., 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

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

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Berstein, L. (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: *The production of speech*, P. MacNeilage (ed.), Springer Verlag: New York, pp. 137–173.

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. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, *27*, 175–200.

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 and C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents* (pp. 191–216). Dordrecht: Kluwer Academic Press.

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

Krahmer, E. & M. Swerts (2006), Cognitive Processing of Audiovisual Cues to Prominence, in preparation.

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 (1985), The motor theory of speech perception revised, *Cognition 21*, 1–36.

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.

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.

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: Cohen, P., Morgan, J., & Pollack, M. (Eds), *Intentions in Communication* . Cambridge, MA: MIT 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.

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.

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

Saltzman, E. & Kelso, J. (1987). Skilled action: A task-dynamic approach. *Psychological Review 94*, 84–106.

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

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

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.

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.

Table 1: Average number of tries per task sentence as a function of the trial (first or second), (in)congruency, and kind of gesture (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)
Gesture	Head nod	1.22 (0.58)
	Eyebrow	1.27 (0.61)
	Manual	1.22 (0.60)

Table 2: 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			—			—

* : $p < .01$

Table 3: Average auditory difference scores (A-diff) as a function of accent, type of visual beat, position of visual beat and trial (std. errors between brackets).

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

Table 4: Average auditory difference score (A-diff) as a function of the position of the pitch accent and the visual beat (std. errors between brackets).

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

Table 5: 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).

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

Table 6: Average visual difference scores (V-diff) as a function of type and position of visual beat, for both W1 and W2 (std. errors between brackets).

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