The Temporal Relation between Beat Gestures and Speech

Thomas Leonard and Fred Cummins*
University College Dublin
thomas.leonard@ucdconnect.ie,fred.cummins@ucd.ie

September 6, 2010

Running title: Beat Gesture Timing

Tel: +353 1 716 2902
Fax: +353 1 269 7262

Corresponding author:
Fred Cummins,
UCD School of Computer Science and Informatics, University College Dublin, Dublin 4 , Ireland

Preprint of article to appear in Language and Cognitive Processes, 2010.

Abstract

The temporal relation between beat gestures and accompanying speech are examined in two experiments. In the first, we find that subjects are very quick to spot altered timing between gesture and speech if the gesture is later than normal, but are considerably less sensitive to alterations that result in an earlier gesture. This suggests an asymmetry in the expectation on the part of listeners/watchers and raises immediate questions about *which* elements within the speech are being perceived as linked to *which* elements in the gestural series. We therefore examine the variability between several kinematic landmarks in a beat gesture, and three potential anchor points in the accompanying speech. We find the least variable relationship obtains between the point of maximum extension of the gesture and the accompanying pitch accent. Together, these findings contribute to our understanding of both the production and perception of beat gestures along with speech, and support an account of speech communication as a strongly embodied activity.

Acknowledgements

This work was funded by Principal Investigator grant, number O4/IN3/I568, from Science Foundation Ireland to the second author. Thanks are due to two anonymous reviewers whose constructive criticisms substantially improved this manuscript.

Introduction

Gestures accompany speech in almost all speaking situations, whether the conversational partner is spatially present or not (Goldin-Meadow, 1999). Their function remains a subject of vigorous debate, although strong cases have been made that gesturing contributes to both the production of speech (Krauss et al., 1995), and its perception (McNeill, 1992). Irrespective of which view is taken, much of the argumentation about the interaction of gesture and speech makes the fundamental assumption that their parallel streams are tightly coordinated in time (Cassell et al., 1999; Wachsmuth, 1999). This assumption, while deeply intuitive and important, has proven hard to substantiate, not least because gestures are not easily described in terms of some atomistic vocabulary. This contrasts with speech, for which we have an abundance of putative atomistic descriptions employing such units as phonemes, syllables, intonational events, etc. Gesture taxonomies and associated notation schemes have been proposed (Kendon, 1980; McNeill, 1992; Kita et al., 1997), but there is far less agreement across researchers than in the study of speech, and it seems probable that the domain of gesture is intrinsically open and heterogeneous in a way that the speech domain is not.

In gesture studies, *beat gestures* have long been recognized as a well defined sub-type of gesture, but they have attracted relatively little formal experimental attention, compared to iconic or deictic gestures (Kelly et al., 2008). Also called 'batons', these gestures are emphatic in presumed purpose, and exhibit relatively little structural variation. A typical beat gesture consists of two phases: an extension phase and a later retraction phase. This may be as small as the waggle of a finger, or it may be a movement of the whole arm and torso. Compared to iconic or metaphoric gestures, there is little meaningful content to a beat gesture, except for its strength and, crucially, its time of occurrence. These two features permit the tight integration of beats into the continuous stream of speech, thereby ostensibly "reveal[ing] the speaker's conception of the narrative discourse as a whole" (McNeill, 1992, p. 15). Beat gestures produced along with speech have been found to modulate brain activity in listeners (Hubbard et al., 2009).

One way of interpreting a beat gesture is as a rhythmical pulse that coincides with, and hence emphasizes, a rhythmical pulse in the speech stream. An initial problem with this interpretation is that rhythm in speech is, itself, difficult to define or operationalize (Dauer, 1983; Cummins, 2009). Another challenge arises because the identification of rhythmic pulses in speech is not a straightforward matter, as the moment at which a pulse is presumed to be felt (the P-centre) does not stand in simple correspondence to any single articulatory or acoustic event (Scott, 1993; de Jong, 1994). Many speech rhythm studies have taken the onset of a stressed syllable to be a rhythmically salient locus. Trying to locate a beat around the syllable onset then becomes essentially the job of P-centre estimation (Morton et al., 1976; Scott, 1993; Cummins and Port, 1998). Kendon (1980) suggested that the extension phase of a beat gesture would coincide with, or slightly precede, the onset of a stressed syllable. De Ruiter (2009) noted that although the issue of temporal synchronization of the two modalities is problematic, the onset of a gesture tends to precede the onset of its lexical affiliate, normally by less than one second.

In the quantitative assessment of synchrony, the researcher must identify points in both speech and gesture that are to be compared. Along with the P-centre, the pitch accent has been suggested as a possible target within speech with which gestures may be coordinated (Roth, 2002; Loehr, 2004). The pitch accent itself has parts, but usually the peak is taken as a point of reference. A simple beat gesture likewise comprises an extension phase, a point of maximum extension (sometimes called the *apex*) and a retraction phase. Little attention has been paid in the literature to

the fine details of synchronization between these points in the gesture and speech streams.

Loehr (2004) proposed that the apex of the gestural stroke, the point of maximum extension, seems to consistently co-occur with a pitch accent, often the intonation peak of the stressed syllable. McClave (1994) observed a tendency for the downbeat (the stroke, or extension phase) of beat gestures to co-occur with the nuclei of tone groups and with the stressed syllable in multisyllabic words. Birdwhistell (1970), in his pioneering work on kinesics, hypothesized a fixed relationship between some kinesic point in gesture and the speech intonation contour. Tuite (1993) made a claim for a regular rhythmic kinesic pulse underlying both the production of speech and gestures. Within speech this kinesic pulse is manifest as a peak in intonation, or pitch accent. In gesture the kinesic pulse corresponds to the extension phase of the gesture. Tuite cites examples from his own corpus in which many of the gestures produced were beat gestures. He observed that the extension phases of these beat gestures occur at regular intervals within speech. The extension phase of a beat gesture will often slightly precede the intonational peak in speech. It appears that this finding was largely based on impressionistic transcription, and despite the efforts of many within the field of speech rhythm research there is yet to be found any event in speech that occurs in isochronous series (e.g. Dauer, 1983). Further evidence of a coupling between body movement and prosodic aspects of speech has been demonstrated in studies of co-speech facial movements. Cavé (1996) identified a correlation between rapid eyebrow movements and rises in F0 tracks. This finding was echoed by Keating (2003) when it was demonstrated that a rise in F0 height mirrored rises in eyebrow height.

Treffner et al. (2008) varied the relative timing between a beat gesture produced by an avatar, and the associated words "put the book there now". They found that the relative timing of the gesture influenced the location of perceived emphasis. Specifically, in their experiment, the relative timing of an animated gesture was systematically moved to all possible locations within the phrase "put the book there now", after phonetic cues to prominence and focus had been removed. Based on their findings, they suggest that, to appear natural, a beat gesture should be aligned such that its mid-point is "synchronous [with] or even preceed[ing] the acoustic body of the word uttered" (p. 54). The speech employed was highly unnatural in both timing (words were separated by gaps of equal size) and intonation (the F0 contour was flat throughout). Krahmer and Swerts (2007) likewise found that beat gestures contribute to the perceived prominence of words.

In the present work we present two studies which seek to contribute to the understanding of the temporal relation between beat gesture and speech. The first is a perceptual study, in which the sensitivity of viewers to altered temporal alignment of visual (gesture) and auditory (speech) streams is probed. The second study examines variability in the temporal relation between various reference points in each stream, in an attempt to identify the most stable temporal relations between them.

Experiment 1: Perception of the Temporal Relationship Between Gesture and Speech

A first experiment sought to establish the sensitivity of observers to the temporal relation between speech and an accompanying beat gesture.

Figure 1: Two frames from one stimulus video, illustrating the neutral position (left) and the maximal extension of a single beat gesture (right).

Methods

Three short videos were prepared, each of a speaker (FC) reading a short fable. For most of the reading, the reader was filmed standing against a white background, with his left arm by his side, and his right arm against his chest. Three words that are naturally accented were selected within each text, and the reader executed a beat gesture to coincide, in as natural a manner as possible, with each of the three words. Sample frames illustrating both the neutral position and a beat gesture are shown in Fig. 1. Audio was simultaneously recorded using an external microphone and solid-state digital recorder. The production of a clap by the reader allowed subsequent alignment of audio and video tracks.

From the three recordings, nine stimulus sets were prepared, one set for each beat gesture, and all stimuli within a set were based upon the same relatively short video excerpt. In the video excerpt, the beat gesture was centered within a 3 second window in which the speaker was visible. Before and after this 3 second window the speaker was completely masked with a black screen. The audio was continuous and was not masked in any way. A minimum of 5 seconds of speech (including pauses) preceded the visible window. The amount of speech audible before and after the visible window varied due to the contingent nature of the phrase within which the accented words were contained. A rectangular mask covered the speaker's head throughout so that no lip movement or other facial movements were available as cues to asynchrony. The body of the speaker was visible from the waist up.

Within each set, nine stimuli were prepared from the same video excerpt. The central stimulus had synchronized audio and video. For the others, the audio was time shifted by amounts of 200, 400, 600, or 800 ms. We refer to stimuli in which the gesture preceded the accompanying sound as *gesture leads*, and when the gesture was late, we refer to the stimuli as *lags*. Lead and lag values were chosen based on informal pilot work.

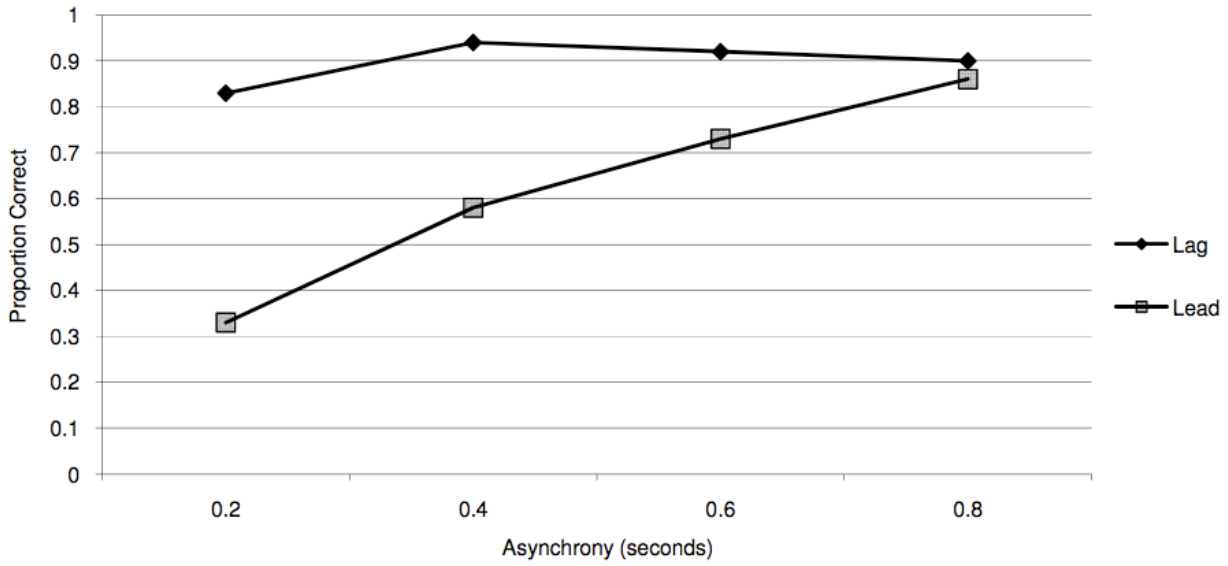

Figure 2: Proportion of correct responses as a function of stimulus asynchrony.

On a single trial, subjects were presented with two videos embedded side by side in a single window. They were instructed to play each once, first the left and then the right. One video always had the central, unaltered, stimulus, while the other was selected from the set of 8 altered experimental stimuli. Assignment of altered or unaltered stimulus to the right or left was randomized. Upon viewing both videos, instructions called for the subject to select the one they perceived as being "out of sync" or "unnatural".

Subjects first completed a short practice of 15 minutes without feedback. They then completed a single randomized block of 72 trials (4 asynchrony levels * lag or lead * 9 stimulus sets). A second full block was completed 24 hours later. Eight subjects took part in the experiment (six males and two females, aged between approximately 18 and 30 yrs). The subjects were volunteers, drawn from the undergraduate and postgraduate population at the University College Dublin, who replied to advertisements and canvassing emails. All were native speakers of English and reported no hearing impairment. All subjects had normal or corrected-to-normal vision. Participants received nominal remuneration for their time. Experimental procedures were approved by the University College Dublin Research Ethics Committee for the Life Sciences.

Results

None of the subjects reported having any difficulty in completing the experiment. For each subject, the proportion of correct responses was ascertained, and results are shown in Fig. 2. It is evident that subjects were able to detect asynchrony consistently across all levels of asynchrony in the gesture lag condition but not in the gesture lead condition. The difference in subject performance is most notable at the lowest levels of asynchrony of 200 and 400 ms, where there is a marked relationship between the differential in performance and the associated amount of asynchrony. At the 600 ms levels subjects performed slightly better in the lag condition than in the lead condition. At the 800 ms level, performance was similar across the two conditions.

A repeated measures ANOVA on the arcsin transformed data was performed with within subject factors of direction (lead vs lag), asynchrony (4 levels) and session. P-values were corrected using the conservative Geisser-Greenhouse adjustment to degrees of freedom. As there was no main effect or interaction involving session, the data from the two sessions were combined, and an RM analysis with factors of direction and asynchrony revealed main effects of direction ($F(1,117)=50.7$, $p < .001$), and asynchrony ($F(1,117)=46$, $p < .001$), and a significant interaction ($F(1,117)=29.9$, $p < .001$). A subsequent by-word analysis revealed no differences between the 9 stimulus sets employed. Wilcoxon matched pairs signed ranks tests were conducted, comparing the arcsin transformed proportions for both leads with lags at each asynchrony level separately, with no correction for multiple testing. Leads differed significantly from lags for asynchronies of 200 ms ($V = 0.2, p < .001$) and 400 ms ($V = 2.5, p < .001$), but not for 600 ms or 800 ms.

The most prominent feature of the results is the evident asymmetry between the condition in which gesture leads speech and when it lags, or occurs later. When the gesture is later than it should be, subjects have no difficulty in spotting the asynchrony, even at the smallest lag of 200 ms. When the gesture is early, on the other hand, subjects require a considerably greater degree of asynchrony in order to notice the experimental manipulation.

An unexpected feature of the data became apparent when individual subjects were examined (See Fig. 3). For some subjects, most notably Subject RY, performance at the shortest leads (gesture before speech) was substantially lower than chance (50%). This means that these subjects consistently selected the stimulus in which the gesture was slightly earlier as being more natural or unmarked. A binomial test showed the proportion correct to be significantly lower than chance for three subjects in the 0.2 sec lead condition.

Discussion

These data suggest the presence of a clear asymmetry in the perception of asynchrony between gesture and speech. Subjects were quick to recognize lags (late gestures) even when the temporal manipulation was as small as 0.2 sec. For early gestures, on the other hand, detection of the experimental manipulation was considerably worse at the shorter asynchronies (0.2 and 0.4 sec). In fact, for three subjects (NM, RS, RY), a gesture was judged as being *more* natural if it was manipulated to appear somewhat earlier than was, in fact, the case. Five of our original 8 subjects were practiced musicians, including all three subjects that displayed a preference for early gestures.

It has been often noted that some gestures regularly precede an associated unit of speech such as a contrastively stressed syllable, even though it may be unclear precisely which features of the gesture are to be considered aligned or associated with which parts of speech (Nobe, 2000; Morrel-Samuels and Krauss, 1992). The highly variable timing data has led some to claim that beats and stresses are not strictly synchronized at all (McClave, 1994). The data presented here suggest an alternative hypothesis, which is that beat gestures stand in an asymmetric temporal relation to some speech event: We might hypothesize an asymmetrical window, aligned with respect to a speech anchor point, which captures the probability of a gesture's time of occurrence. This is illustrated in Fig. 4. Our data are not sufficiently precise to make strong claims about the shape of the window. It is worth noting, however, that the conditions under which the readings and gestures were obtained necessitated a strong awareness on the part of the speaker of the relative timing of the beat gesture. This might conceivably have led to the production of gestures that are more closely linked to their respective anchor points in the speech stream than would generally be the case. If this were so, and the rudimentary model sketched in Fig. 4 is approximately correct,

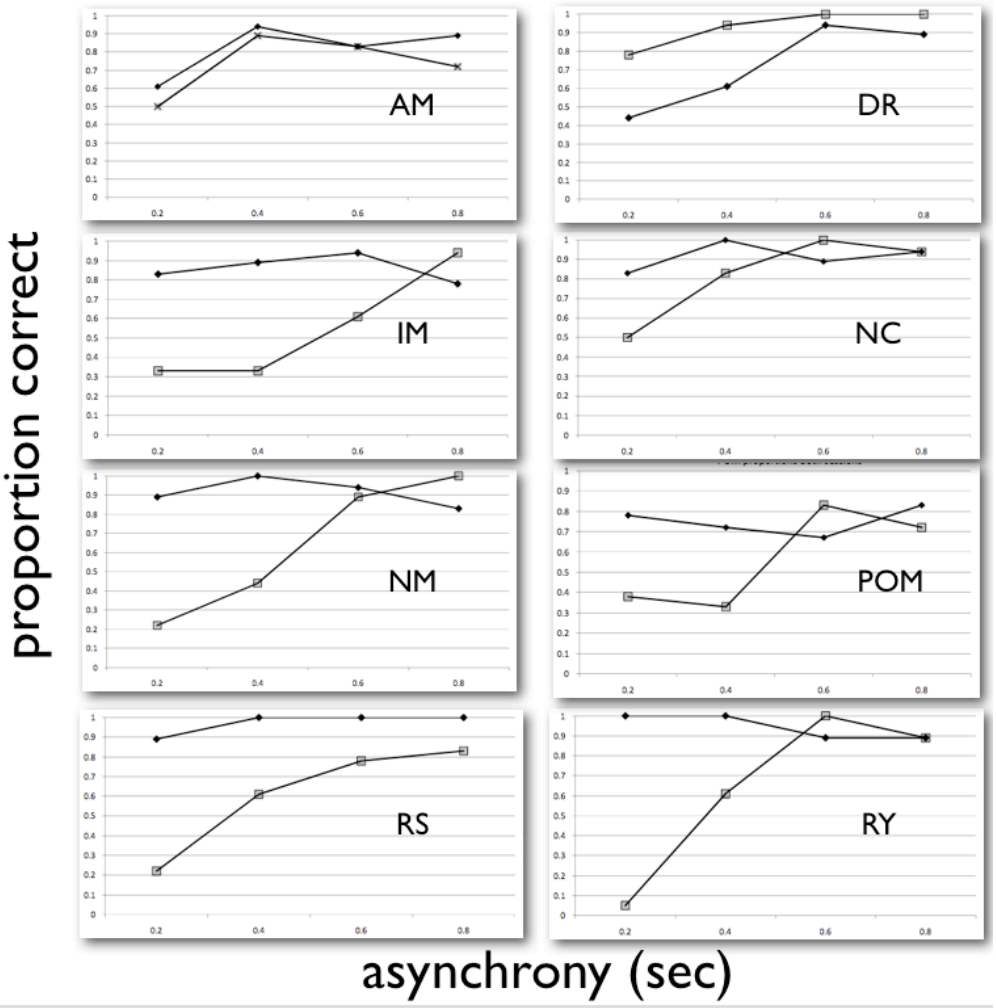

Figure 3: Individual subject data. Proportion of correct responses as a function of stimulus asynchrony.

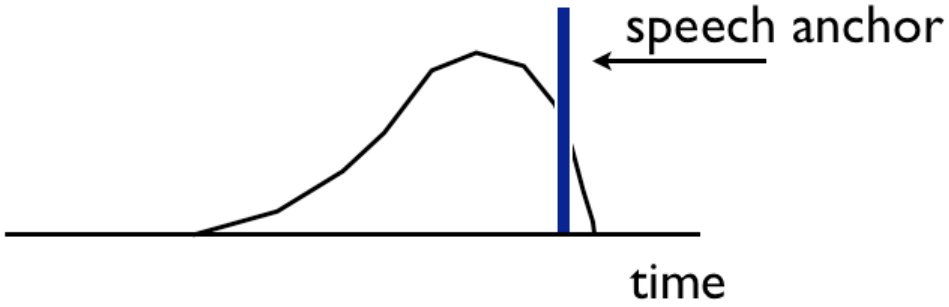

Figure 4: Suggested form for a probability distribution of the timing of a gesture with respect to a speech anchor.

then the tendency of some subjects to regard an earlier gestural placement as more natural than the one actually produced would appear less counter-intuitive.

Our findings are consonant with those of Treffner et al (2008), who noted: "the *perception* of the intended focus of a sentence is strongly influenced by a gesture provided that the gesture is produced prior to or simultaneous with the utterance" (p. 55, emphasis in the original). If we make the not unreasonable assumption that speakers time their gestures to effectively conspire with other prosodic cues in marking prominence, then there appears to be a good match between communicative efficiency and perceptual sensitivity. This then raises the question of exactly how multiple cues produced in parallel streams are yoked together in the service of common communicative goals. It is to this question that we now turn in a production experiment.

Experiment 2: Analysis of the Temporal Relationship Between Gesture and Speech

The quantitative analysis of the temporal relation between gesture and speech must of necessity be based upon the measurement of specific points within both gesture and speech. There is still a great deal of uncertainty about the most appropriate procedures for such measurement. Beat gestures have a highly constrained structure, and so they lend themselves well to quantitative study. In a second experiment, we recorded the three-dimensional movement of the hand and arm as a beat gesture was executed, in order to assess the relative stability of coordination among diverse points in both data streams.

Methods

A Codamotion 3-D motion tracker (Charnwood Dynamics, UK) was employed to obtain movement data. A single LED marker was affixed to the base of the thumb of the reader, who stood while reading. The same three texts were employed, and two readings of each were made. Again, the reader held his arm against his chest except when making each of three beat gestures on prominent syllables chosen beforehand. The 3-D motion tracker provided the position of the sensor in the X (forward-backward), Y (horizontal) and Z (vertical) planes, relative to an origin situated on the floor, near the speaker. The time-varying Euclidian distance from the origin was calculated, and

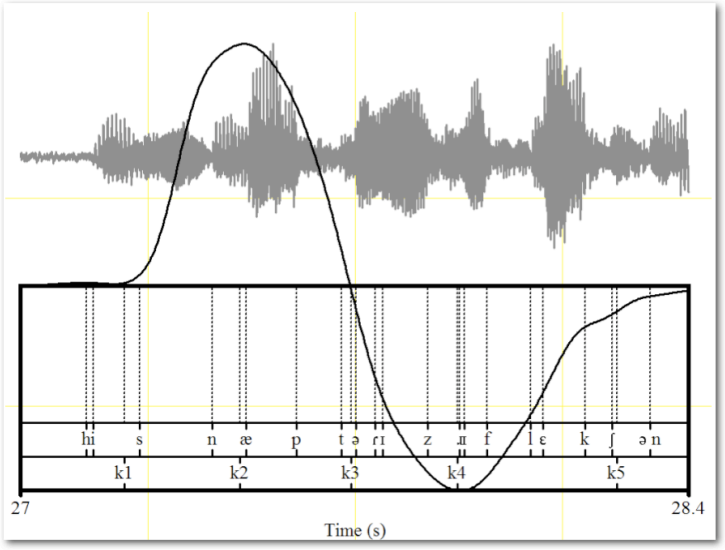

Figure 5: Sample velocity trace of hand movement accompanying the phrase "he snapped at his reflection". Points: k1 = movement onset; k2 = peak velocity of extension phase; k3 = point of maximum extension; k4 = peak velocity of retraction phase; k5 = termination of gesture.

the first difference of this provided a record of velocity as a function of time. All of the kinematic landmarks used for analysis were obtained from the velocity data. Velocity traces were smoothed using a low pass filter in Praat (Boersma 2001) with a cutoff of 30Hz. From these, five kinematic landmarks were identified for each gesture: the onset of movement (k1), the peak velocity of the extension phase (k2), the point of maximum extension of the hand before retraction (k3), the peak velocity of the retraction phase (k4), and the termination of the gesture (k5). Each of these was compared to three landmarks in the speech waveform: the vowel onset of the stressed syllable in each word, the estimated P-centre, and the pitch peak within the stressed syllable. P-centre estimation was done using the method introduced in Cummins and Port (1998), based on the model developed by Scott (1993). This method places a beat, or P-centre, half way through a local rise in the intensity envelope of the filtered waveform, where band pass filtering is first used to eliminate energy below 500 Hz and above 1500 Hz. Estimated P-centres are close to vowel onsets for simple syllables, but tend to occur earlier as consonantal onsets become more complex. An example of a velocity trace for the beat gesture produced on the word 'snapped' is shown in Fig 5.

Results

Fig. 6 presents box plots of the offset of each of the five kinematic landmarks from the three possible speech anchor points. There are 18 data points in each set (3 texts, 3 beats per text, 2 repetitions). Gestures are typically approximately one second long (M: 1.02, s.d. 0.09), with movement onset approximately 300 ms before the onset of the stressed vowel. The point of maximum extension is regularly reached within the stressed syllable, about 200 ms after the vowel onset. By contrast, the velocity maximum of the extension phase of the gesture seems to be well aligned with either the vowel onset or the P-centre. The mean difference between estimated P-centre and vowel onset in

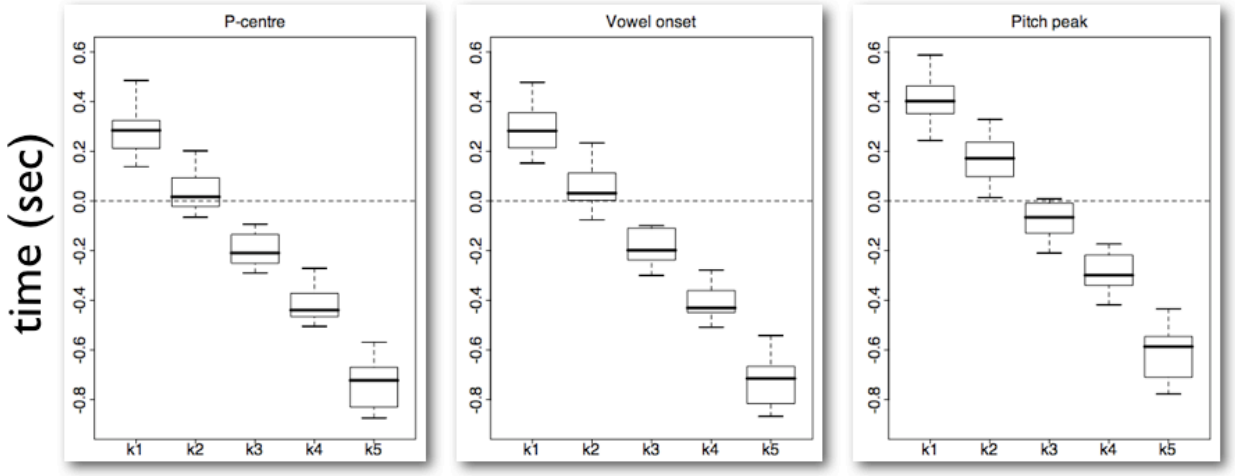

Figure 6: Distribution of the intervals between each of five kinematic landmarks in a beat gesture and three possible speech anchor points. For k1–k5, see previous figure.

this data set is no more than 0.01 sec.

Simultaneity of two events is not evidence that the two events are more tightly coupled (i.e. with less variance) than two events that occur at a fixed lag. Evidence that two events exhibit a functional linkage or coupling must come from examining variability. Fig. 7 plots the variance of the interval between each of the five kinematic landmarks and the speech anchors. Regardless of the speech anchor considered, the intervals between the gestural and speech landmarks display a fixed pattern of variability, with the point of maximum extension showing the most consistency in its relative timing with respect to each potential speech anchor. Post-hoc pairwise differences in the variability of any two intervals are not statistically significant.

Discussion

Our data suggest that not all points in the beat gesture are tied to the continuous stream of speech to an equal degree. Variability in relative timing is minimized for the apex of the gesture, irrespective of the speech anchor point examined. The closest speech landmark to the apex is the peak of the pitch accent on the stressed syllable, which agrees with the observations of Loehr (2004) and the suggestions of Roth (2002). As with the previous experiment, some caution is warranted in interpreting these results as the speaker/gesturer was necessarily attending rather more than might usually be the case, to the timing of the beat gestures. It may be that naturally occurring beat gestures exhibit greater variation in timing than seen here.

The highly constrained form of beat gestures makes them particularly suitable for examining the degree to which manual gestures and speech are temporally coordinated. We have presented evidence of an asymmetry in the sensitivity of listeners to the relative timing of beats and speech. We have also shown that the apex of the beat gesture seems to exhibit less temporal variability with respect to speech than any other point within the gesture. Both of these observations are, necessarily, restricted to these particularly simple out-and-then-back beat gestures, and do not

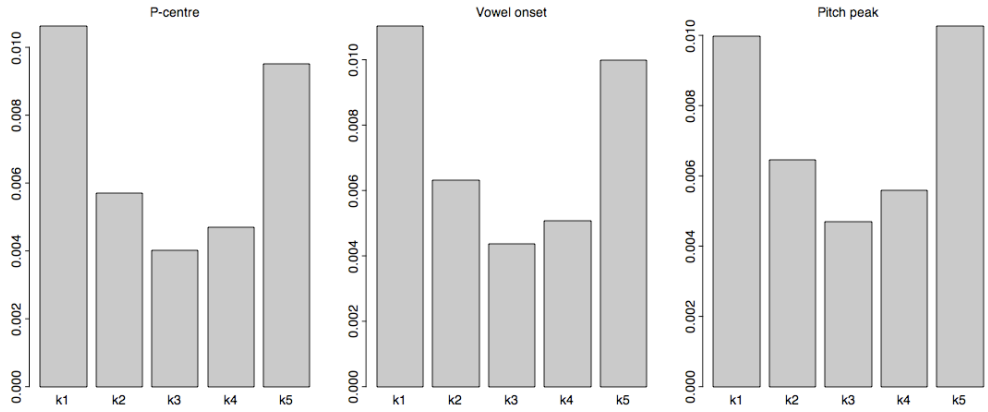

Figure 7: Variance of relative timing of each of the five kinematic landmarks from the beat gestures and the three potential speech anchor points.

allow obvious generalization to the many other, highly variable, gestures accompanying normal speech.

The experimental conditions employed here are, of necessity, highly constrained. Gesturing, however, is normally done unselfconsciously. It might reasonably be objected that the staged context in which the gestures are elicited, together with the speaker's awareness of the importance of gestural timing, make these a poor proxy for naturally occurring spontaneous gestures. The experimental demands also led to relatively large gestures being made. Large beat gestures are not uncommon, but most beat gestures are probably done with a finger or hand, rather than with the whole arm. Several factors suggest that these observations, while not wrong in substance, do not make the present work either uninterpretable or misrepresentative of beat gestures more generally.

The main observation to be made in this context is that the findings here are not surprising. They provide some quantitative substance to observations that, somewhat less formally, are already known. The asymmetry in the perception of gesture-speech timing has been found qualitatively in Treffner et al. (2008) and elsewhere. The relatively tight alignment of gestural apex with pitch accent has been noted before (Loehr, 2004; Roth, 2002), but not assessed quantitatively as here. We might also note that, subjectively, the speaker did not find that the task of beat gesture placement felt in any way artificial, but resembled instead the task of placing stress, focus, or accent in a required point within a sentence.

Beat gestures are notably lacking in any overt semantic or propositional content. It has been suggested that representational gestures with overt semantic associations are more likely to occur when discussing action related topics, particularly when it is necessary to imagine or simulate the action in question (Hostetter and Alibali, 2008). Indeed, Hostetter found that the proportion of representational gestures increased, and the proportion of beat gestures decreased, as the strength of simulated action was greater (Hostetter, 2008). This being so, it seems less objectionable that beat gestures should be studied in a highly constrained experimental context, as these circumstances will, in all probability, lack the kind of speaker engagement found to produce representational gestures at the expense of beat gestures.

The links between basic bodily movement and speech run deep. The ubiquity of gesturing when speaking, even among the blind has often been noted (Goldin-Meadow, 1999; McNeill, 1992).

Speech is, of course, an exquisite motor skill, evocatively described by Stetson as "movement made audible" (Stetson, 1951). Speech and manual movement share extensive brain mechanisms, and often exhibit linked pathologies (Iverson and Thelen, 1999). For example, Mayberry et al (1998) noted that gestures tend to freeze during a stuttering event and resume once the dysfluency has passed. The present study adds to this growing body of work that insists that speech is properly understood as a thoroughly embodied activity, in which both speakers and listeners are manifestly influenced by the physical instantiation of the act of communication in bodies.

References

Birdwhistell, R. (1970). *Kinesics and Context: Essays on Body Motion Communication.* University of Pennsylvania Press.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

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

Cavé, C., Guaïtella, I., Bertr, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and f0 variations. In *Proc. ICSLP96*, pages 2175–2179.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica*, 66(1–2):15–28.

Cummins, F. and Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2):145–171.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62.

de Jong, K. (1994). The correlation of P-center adjustments with articulatory and acoustic events. *Perception and Psychophysics*, 55(4).

de Ruiter, J. (2000). The production of speech and gesture. In Mc Neill, D., editor, *Language and Gesture: Window into Thought and Action.* Cambridge University Press.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419–429.

Hostetter, A. B. (2008). *Mind in motion: The Gesture as Simulated Action framework.* PhD thesis, The University of Wisconsin – Madison.

Hostetter, A. B. and Alibali, M. W. (2008). Visible embodiment: gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3):495–514.

Hubbard, A. L., Wilson, S. M., Callan, D. E., and Dapretto, M. (2009). Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30(3):1028–1037.

Iverson, J. M. and Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 6(11–12):19–40.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., and Bernstein, L. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pages 2071–2074.

Kelly, S. D., Manning, S. M., and Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, 2(4):569–588.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M. R., editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton, The Hague.

Kita, S., van Gijn, I., and van der Hulst, H. (1997). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 23–35. Springer-Verlag, Bielefeld, Germany.

Krahmer, E. and Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414.

Krauss, R., Dushay, R., Chen, Y., and Rauscher, F. (1995). The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology*, 31(6):533–552.

Loehr, D. (2004). *Gesture and Intonation*. PhD thesis, Georgetown University.

Mayberry, R. I., Jaques, J., and DeDe, G. (1998). What stuttering reveals about the development of the gesture-speech relationship. *New Directions for Child Development*, 79:77–87.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23(1):45–66.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Morrel-Samuels, P. and Krauss, R. M. (1992). Word familiarity predicts the temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:615–623.

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, 83:405–408.

Nobe, S. (2000). Where do most spontaneous representational gestures actually occur with respect to speech. In Mc Neill, D., editor, *Language and Gesture*, pages 186–198. Cambridge Univ Press.

Roth, W. (2002). From action to discourse: The bridging function of gestures. *Cognitive Systems Research*, 3(3):535–554.

Scott, S. K. (1993). *P-centers in Speech: An Acoustic Analysis*. PhD thesis, University College London.

Stetson, R. H. (1951). *Motor Phonetics*. North-Holland, Amsterdam, 2nd (1st ed. 1928) edition.

Treffner, P., Peter, M., and Kleidon, M. (2008). Gestures and phases: The dynamics of speech-hand communication. *Ecological Psychology*, 20(1):32–64.

Tuite, K. (1993). The production of gesture. *Semiotica*, 93(1-2):83–106.

Wachsmuth, I. (1999). Communicative rhythm in gesture and speech. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 277–289. Springer-Verlag London, UK.