THE TIMING OF SPEECH-ACCOMPANYING GESTURES WITH RESPECT TO PROSODY

Yelena Yasinnik [1], Margaret Renwick [2] & Stefanie Shattuck-Hufnagel [1]

[1] Speech Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2] Wellesley College, Wellesley, MA USA
stef@speech.mit.edu

ABSTRACT

The hypothesis that some of the hand and head movements produced during speaking are timed with respect to the prosodic structure of the utterance was tested, by comparing the timing of separately-labelled video and sound files from videotaped lectures by three male speakers of English. Word boundaries and prosody were labelled in the sound files using the ToBI system to specify location of pitch-accented words and syllables, and of intonational phrase boundaries. Gestures were labelled in the video files using a two-way distinction between DISCRETE movements (characterized by a sudden stop suggesting target attainment) and more CONTINUOUS movements. Gesture times were expressed in terms of frame location in a 30-frames-per-second video display, and the video frames corresponding to target attainment for the discrete gestures (2 speakers) or to onset and offset for all gestures (1 speaker) were aligned with the prosodic markings in the sound files. Preliminary analysis suggests that a) discrete gestures may be timed with respect to pitch accented syllables (or possibly the prominence-related constituents they define), and b) gestures which span the boundaries between adjacent intonational phrases may indicate larger structural groupings. If ongoing studies of additional utterances and speakers confirm these results, it will provide evidence that speech planning models need to generate a speaking plan and a gesturing plan in tandem.

BACKGROUND

The nature of speech-accompanying gestures has long been a focus of intense interest; one aspect of this interest is the timing of the gestures with respect to the prosody of their accompanying utterances. For example, earlier studies raised the possibility that manual gestures of the type defined by McNeil (1992) as 'beats' are timed to occur with speech rhythms. Other published studies on the relationship between prosody and gestures have focused on facial expression, specifically eyebrow movement, and its correlation to the f0 (pitch) track. Cavé et al. (1996) found that 71% of rapid eyebrow movements corresponded to rises in speakers' f0 tracks. Keating et al. (2003 and p.c.) also found a correlation between f0 height and eyebrow height. Flecha-García (2001) investigated eyebrow movements' distribution across conversation types and concluded that their unequal distribution suggests a linguistic, communicative function for eyebrow movements.

This issue has recently become more accessible to quantitative investigation with the advent of widely (although not universally) accepted tools for transcribing speech prosody, such as the ToBI system (Beckman and Elam 1997). This study tests the more general hypothesis that speakers time their gestures with respect to the prosodic constituent structure and prominence

patterns of utterances. Samples of professional lectures by 3 male speakers of English were digitized and labelled for the time of occurrence of both prosodic elements (intonational phrase boundaries and pitch accent prominences) and gestural elements (onsets, offsets and target attainments). The two sets of labels were then aligned with each other, to determine the degree to which prosodic prominences and phrasings were timed with gestures. To the extent that such alignments are found, we can infer that models of the production planning process should provide a mechanism for mutual timing between non-speech gestures and prosody.

METHOD

Database

Samples were excised from commercially available videotapes of 3 academic lectures. All 3 lecturers were male; one appeared to be a speaker of Australian English (M1au) and two to be speakers of American English (M2am, M3am). The video samples were transferred to a MacIntosh computer for labelling of the speech-accompanying gestures of the head, hand and eyebrows using iMovie, which permits both real-time playback and frame-by-frame playback with a display of the individual frame number. The corresponding sound files were transferred to a Unix system for ToBI labelling. The sample for Speaker M1au was approximately 5 minutes long, for speaker M2am approximately 7.5 minutes long, and for speaker M3am approximately 8 minutes long (although results for only 1 min 50 seconds are discussed here.)

Labelling

Prosody labelling

The prosody of the spoken utterances was labelled using the ToBI system; aspects of the labels used here include pitch accent locations (i.e. the words and syllables marked with intonational phrase-level prominence), and intonational phrase boundaries (Beckman and Pierrehumbert 1986). Experienced labellers aligned the words in a ToBI .word file with the wave form (including the symbol PAU for each perceptually-noticeable silence), and transcribed the pitch accents and boundary-related tones of the utterances in .tones and .breaks files, as they listened to the sound files and viewed time-aligned f0 tracks which had been created using the xwaves algorithm. The .words, .breaks and .tones files provide a time-stamp for each word or pause ending, tonal element and intonational phrase boundary that is transcribed. A different prosody labeller transcribed each of the three speech samples.

Gesture categories.

The initial phase of the research involved finding an aspect of speech-accompanying gestures whose timing with respect pitch accents could be reliably quantified. Preliminary observations by the second author of this paper suggested that, for this purpose, speech-accompanying gestures can be usefully divided into two categories: DISCRETE and CONTINUOUS gestures. The distinguishing characteristic of a Discrete gesture is a “hit”: an abrupt stop or pause in movement, which breaks the flow of the gesture during which it occurs. Hits appear as bouncing, jerky movements, changes in the direction of movement, or as complete stops in movement; they resemble the movements of an orchestra conductor marking each beat with a sharp movement of the baton. If a gesture ends slowly, coming to a gentle halt (slowing down before stopping, rather than speeding up and “crashing” to a stop), it does not contain a hit. Continuous gestures are repetitive movements that are not punctuated by hits; for example,

drawing circles with the hand. In addition to single Discrete gestures, we also observed strings of repeated Discrete gestures, each of which contained a hit. The criterion for such Hybrid gestures was that the articulator's direction of motion or the speaker's hand shape did not change between hits, giving the string an appearance of repeated hits of the same type. All of the hits that occurred in Hybrid gestures are included in the analysis here. Both Discrete gestures' hits and pitch accents are easily pinpointed in time, so comparison of their locations was relatively simple, and allowed for quantifiable comparison of speech events and gestures.

Labelling Discrete gesture hits
M1au's sample was transcribed while listening to the speech, specifying the syllable that seemed to occur with the hit. In contrast, M2am's sample was transcribed without listening, specifying only the frame number that contained the hit. Because movement (of hand, head, etc.) stops during a hit, its video frame is in sharp focus and not blurred by motion, making it relatively easy to identify. These 33ms-long frames provide the opportunity to determine where a hit occurs in relation to the speech. The same person transcribed the samples for speakers M1au and M2am, finding the video frames that contained a Discrete gesture hit.

Labelling Gesture onsets and offsets
Preliminary observations by the first author of this paper suggested that speech-accompanying gestures might group into larger gestural ensembles, raising the possibility that such ensembles correspond to groupings of intonational phrases into larger discourse-related structures. Accordingly, speaker M3am's sample was labelled for the onsets and offsets of hand gestures by a second gesture transcriber using the following criteria:

-Onsets were marked at the frame where hand shape changed, or hand position changed (e.g. wrist turns), or hand location changed (usually accompanied by a blurring of the image of the hand or fingers in the video display).

-Offsets were marked at the frame where the hands stopped moving (this generally corresponded to a clearer image than in the surrounding frames), or at the frame just before the change in hand shape or hand position.

In this speech sample, head and eyebrow gestures always appeared to occur within the time period of a single hand gesture, so they are not reported separately here. The onsets and offsets of Discrete gestures in M2am's sample were also labeled, but these results will not be discussed here.

Alignment
For M2am and M3am, the time alignments of word boundaries, pitch accents and phrase boundaries were transferred from the ToBI label files to an Excel spreadsheet, and aligned with the gesture label files which had been translated from video frame numbers into milliseconds. The resulting displays showed the location, with respect to word boundaries and prosodic elements, of the 33 ms video frame identified as the location of the hit (for Discrete gestures in the M2am sample) or the onset or offset of a gesture (in the M3am sample).

Groupings of gestures for the M3am sample were arrived at in the following way: For each Intonational Phrase (Break Index 3 or 4 in the ToBI transcription system) it was ascertained whether a gesture began in the preceding phrase and continued across the phrase boundary to

a location within the following phrase, or not. Phrases with such boundary-spanning gestures were grouped into ensembles; each of these larger constituents thus ended with an Intonational Phrase for which the gesture did not extend into the speech signal of the following phrase.

As a preliminary analysis to determine whether the gesture-based groupings of intonational phrases defined in this way correspond to larger organizational elements of the discourse, the durations of pauses between IPs were measured. This permitted a test of the hypothesis that the groupings defined by the absence of phrase-boundary-spanning gestures are also separated by longer pauses. Such a finding would support the hypothesis that the groupings of phrases linked by phrase-spanning gestures correspond to higher-level structures.

RESULTS

Results will be summarized for each of the three speaker samples separately, since different methods were used and different questions were addressed in an exploratory way for each.

M1au: Discrete Gesture Hits vs. Pitch Accents

This sample of approximately 5 minutes of speech included 1818 syllables and 622 Pitch Accents (because words sometimes have more than one pitch accent, this total is larger than the number of pitch-accented words). The sample also included 206 Discrete gestures (including both simple and Hybrid types). Hit locations were labelled while listening to the speech. Of the 206 hits, 185 or 90% were transcribed with their hit occurring on a syllable that had been ToBI labelled as pitch-accented. Thus, when a hit did occur, it tended to align with a pitch-accented syllable.

This result is consistent with the hypothesis that speech-accompanying gestures are planned to occur in close synchrony with the prosody of the spoken utterances. However, because the gesture labeller listened to the accompanying speech, there is a possibility that the pitch-accented syllables acted as perceptual attractors, resulting in a bias in the gestural labelling. To address this possibility, the second speaker's sample was labelled for gestures without listening to the speech.

M2am: Discrete Gesture Hits vs. Pitch Accents

This sample of 7.5 minutes of speech contains 1421 words, of which 932 were monosyllables and 489 were polysyllables. There were 769 Pitch Accents and 265 Discrete gestures (simple + Hybrid). Hit locations were labelled by video frame, without listening to the speech. Of the 489 polysyllabic words, 420 were labelled with at least one Pitch Accent while 130 overlapped with the 33 ms video frame that contained the hit associated with a Discrete gesture. Of the 130 hit-aligned words, 117 or 90% also contained a Pitch Accent, consistent for the by-syllable results for speaker M1au.

The picture is somewhat different for the 932 monosyllabic words, of which 280 were pitch-accented and 116 overlapped with the hit frame of a Discrete gesture. Of the 116 hit-aligned words, 75 or 65% also were labelled with a Pitch Accent; 41 were not (although most of these were within 100 milliseconds of a pitch accented syllable and often considerably closer). The fact that a smaller proportion of monosyllabic words than polysyllabic words were both pitch accented and aligned with a Discrete gesture hit raises the possibility that the alignment of the gestures is not with the pitch-accented syllable *per se*, but with another prosodic constituent that groups a strong syllable with the following weak syllables. Some support for this hypothesis is

provided by the following observations: Of the 41 monosyllabic words aligned with hits but not pitch-accented, 8 occurred on a rhythmically strong syllable, and 13 occurred in a weak syllable following a strong syllable (such as the 'me' in 'help me to'). These patterns hint at a role for a foot-like rhythmic grouping as a possible domain of the timing for Discrete gesture hits. An additional preliminary observation that points in this direction is that many the video frames for hits aligned with pitch-accented polysyllabic words begin in the coda of the pitch-accented syllable and overlap with some portion of a following unstressed syllable. Further analysis of the alignment of hits with the specific pitch-accented syllables of polysyllabic words vs. within a foot-like grouping of strong and weak syllables in those words will provide a stronger test of this hypothesis.

M3am: Gesture-defined groupings of Intonational Phrases

This sample included almost 2 minutes of speech/video, with 282 words grouped into 71 Intonational Phrases (ToBI 3's and 4's). It also included 79 individual gestures, defined by the onset and offset criteria described above. The cross-phrase-boundary-gesture grouping criterion grouped the 79 gestures into 9 single-IP and 14 multiple-IP ensembles. The 9 single-IP ensembles were neutral with respect to the hypothesis that gesture-span ensembles correspond to higher-level constituents, since it was not possible to compare within-ensemble vs. between-ensemble inter-phrase pause durations. However, 12 of the 14 multi-phrase ensembles showed the predicted pattern of pause durations, i.e. the pause at the end of the ensemble was longer than the pauses between phrases within the ensemble. The pause durations for one of the 4-phrase ensembles is shown in Figure 1.

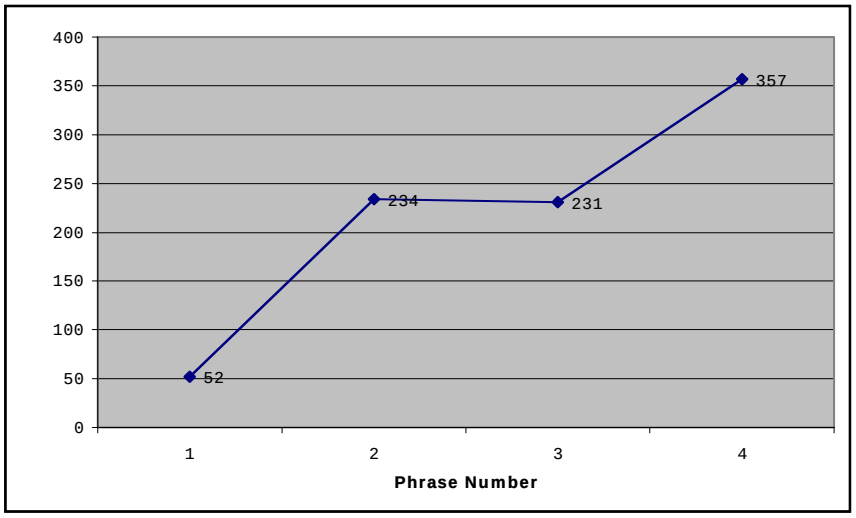

Figure 1. Pause durations after each phrase in a four-phrase ensemble, which is defined by a) the spanning of internal post-boundary pauses by a gesture, and b) the non-spanning of the final post-boundary pause by a gesture.

The remaining two multi-phrase gesture-span-defined ensembles showed a different pattern, in that the pause after the final phrase in the ensemble was not the longest one. However, re-examination of these tokens revealed that, in both cases, there was a sequence of gestures: one gesture ended in the long mid-ensemble pause and another gesture began immediately

afterwards, within the same pause. Since there was no time delay between the two gestures, they were originally labelled as a single gesture spanning the pause between the two intonational phrases. If these two gestures are regarded as separate, as indicated by the clear change in direction of motion of one or both hands, then there is no gesture spanning this long between-phrase pause. As a result, it meets the definition of an ensemble-final boundary, and these two multi-phrase ensembles correspond to the pattern of the other 12, i.e. their longest pause occurs after the final (non-gesture-spanned) phrase boundary.

DISCUSSION

The results described above provide preliminary support for the hypothesis that, when speakers plan utterances, they plan the accompanying gestures to occur in close temporal alignment with the prosodic structure of the speech. The observations are in agreement with other recent studies indicating facial gestural timing alignment with f0 height, and they extend these results to gestures made by the head and hands, while providing information about the phonological (i.e. prominence-lending) character of the associated f0 patterns and indicating a role for Intonational Phrase groupings as well. Questions about the generality of such timing relationships across different speakers, different speaking circumstances and different languages remain to be explored. But the results point strongly to the need for speech production planning models which incorporate a mechanism for aligning movements with other bodily movements during spoken utterances.

ACKNOWLEDGMENTS

This work was supported in part by NIH/NIDCD grant RO1 DC00075, the Keith North Fund at the Speech Group at MIT, and MIT's Undergraduate Research Opportunities Program. Discussions with Anna Esposito and Diane Brentari are gratefully acknowledged, as is technical assistance and support from Helen Hanson, Majid Zandipour and especially Dr. William Renwick of Miami University (Ohio). All errors are of course the responsibility of the authors.

REFERENCES

Beckman, M.E. and Pierrehumbert, J.B. (1986), Intonational structure in Japanese and English. *Phonology Yearbook 3*, 255-309

Beckman, M.E. and Elam, Gayle Ayers (1997), Guidelines for ToBI labelling. Available from ling.ohio-state.edu/~tobi

Cave, C., I Guaitella, R. Bertrand, S. Santi, F. Harlay & R. Espesser (1996), About the relationship between eyebrow movements and f0 variations. *Proceedings of the ICSLP*, (pp. 2175-2179). Philadelphia, PA, USA.

Flecha-García, M. L. (2001), Facial Gestures and Communication: what induces raising eyebrow movements in Map Task dialogues. Retrieved April 23, 2004 from http://www.ling.ed.ac.uk/~pgc/archive/2001/marisa2001.pdf

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., Bernstein, L. (2003), Optical phonetics and visual perception of lexical and phrasal stress in English. *Proceedings of the ICPhS* (pp. 2071-2074). Barcelona, Spain.

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