# A KINEMATIC ANALYSIS OF PROSODIC STRUCTURE IN SPEECH AND MANUAL GESTURES

Jelena Krivokapic,[1, 2] Mark K. Tiede, [2] Martha E. Tyrone [2,3]

[1]University of Michigan, [2]Haskins Laboratories, [3]Long Island University Brooklyn
jelenak@umich.edu, tiede@haskins.yale.edu, Martha.Tyrone@liu.edu

## ABSTRACT

Two experiments examining the effects of prosodic structure on the kinematic properties of speech and manual gestures are presented. Experiment 1 investigated the effects of prosodic boundaries, stress and their interaction on manual, oral, and intonation gestures (their duration and coordination). Experiment 2 investigated the effects of different types of prominence (deaccented, narrow focus, broad focus, contrastive focus) on oral constriction, intonation and manual gestures (duration and coordination). We recorded speech audio, vocal tract gestures (using electromagnetic articulometry) and manual movements (using motion capture). To the best of our knowledge, this is the first study to examine kinematic properties of body movement and vocal tract gestures concurrently. Preliminary results focus on the effects of prosodic structure on gesture duration and show that 1) manual and oral gestures are longer phrase-initially than phrase-medially and 2) manual and oral gestures lengthen under phrase-level prominence. [Supported by NIH].

**Keywords**: prosody; speech production; manual gestures; articulatory gestures; prosodic lengthening.

## 1. INTRODUCTION

Prosodic structure is marked by temporal and tonal properties. At boundaries, acoustic segments become longer and articulatory gestures become larger, longer, and less overlapped (e.g., [4, 6, 15, 25]). The effects are graded, such that higher boundaries show more lengthening (e.g., [4, 6, 8, 25]). The overall effects of prominence are less well understood, but prominent syllables also exhibit lengthening (e.g., [7, 10, 13, 19]). Tonally, prosodic structure is marked by rising and falling pitch on prominent syllables and phrase-finally.

In addition to acoustic and articulatory manifestations of prosodic structure, there is evidence of a relationship between prosodic structure and body movement (gesturing). Thus gesturing is timed to prominent syllables (e.g., [11, 18, 21, 22, 23, 26]), and this timing can be influenced by prosodic boundaries ([14]). The exact nature of the relationship between prosodic structure and body gesturing is not well understood, however. The first goal of this study is to address some basic questions regarding this relationship. We present two experiments. Experiment 1 examines the effect of prosodic boundaries, stress, and their interaction on oral, intonation and manual gestures. Experiment 2 examines the effects of varying degrees of prominence (deaccented, narrow focus, broad focus, contrastive focus) on these gestures. Working within the framework of Articulatory Phonology ([3 ff.]), these two experiments allow us to examine 1) the effect of prosodic structure (boundaries and prominence) on the duration of vocal tract gestures and manual gesturing (specifically pointing gestures) and 2) the coordination of oral gestures, manual gestures and intonation gestures, and how prosodic structure affects this coordination. We test the hypothesis that prosodic control extends beyond the vocal tract ([21, 26]). Under this hypothesis, for question 1, pointing movements are expected to be affected by prosodic structure and to show phrase-initial temporal lengthening which increases with boundary strength, in parallel with oral constriction gestures. The effects of prominence are less clear, but we expect that overall there will be an increase in gesturing duration from the deaccented condition to contrastive focus (as has been observed in [9, 12, 19]). This question is the focus of the current study. Further analyses to test the second question are ongoing. The prediction is that manual gestures are coordinated with intonation and/or oral constriction gestures in-phase or anti-phase, but that they will not affect how vowel and consonant gestures are coordinated with each other, since manual gestures are not part of the lexical representation of the word ([16, 20]).

To address these questions, kinematic data are needed for both vocal tract gestures and manual gesturing. For this purpose, we recorded speech and body movement concurrently, using electromagnetic articulometry (EMA) and motion capture. To the best of our knowledge, this is the first study to record the kinematics of vocal tract gestures and body movements simultaneously; accordingly, a second goal of the study is to demonstrate the feasibility of the data collection method.

## 2. METHODS

### 2.1. Stimuli and participants

The Experiment 1 stimuli consisted of six sentences varying the phrase-initial boundary (word boundary, ip—intermediate phrase boundary, IP—Intonation Phrase boundary) and stress position (the first or second syllable of disyllabic target words). To keep the segments in the two stress conditions identical, two nonce words were used (the names *MIma* and *miMA*, with stress on the first and on the second syllable, respectively). The sentences were read from a computer screen twelve times for a total of 72 sentences (3 boundary x 2 stress x 12 repetitions). The sentences were semi-randomized in blocks of six sentences. Table 1 lists stimuli for the target word *MIma* (stress on the first syllable). The sentences for the condition with stress on the second syllable were identical except that the target word was *miMA*. Participants were asked to point to the appropriate picture of a doll (named either *miMA* or *MIma*) while reading the target word.

**Table 1**: Stimuli for Experiment 1 for the target word *MIma*. The boundary is before *MIma.*

| Condition | Sentence |
|---|---|
| 1. word | There are other things. I saw MIma being stolen in broad daylight by a cop. |
| 2. ip | Mary would like to see Shaw, MIma, Beebee, and Ann while she is here. |
| 3. IP | There are other things I saw. MIma being stolen was the most surprising one. |

Experiment 2 consisted of four sentences that varied the type of prominence on the target word (deaccented, narrow focus, broad focus, contrastive focus). They were repeated twelve times, for a total of 48 sentences (4 conditions x 12 repetitions). The sentences were semi-randomized in blocks of four sentences. The target word in all sentences was *Bob*. The sentence was always "*Anna wants to see Bob. In the morning if possible*". To elicit the appropriate prominence, the sentence was placed in a question-answer context (the questions are given in Table 2; see [19] for a similar procedure). Participants were asked to point to the picture representing Bob while reading the target word.

The data collected in these experiments are part of a larger study. Two native speakers of American English participated; they were paid for their participation and naïve as to the study's purpose.

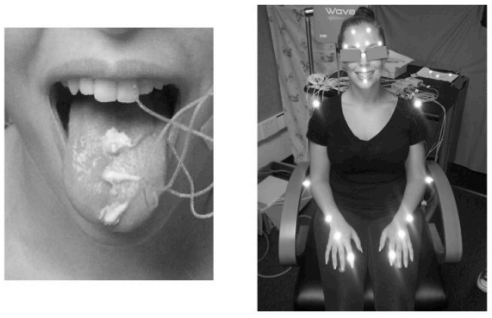
**Table 2**: Context questions for Experiment 2.

| Condition | Context |
|---|---|
| *deaccented* | Does Lenny want to see Bob? |
| *broad* | What is going on? |
| *narrow* | Who does Anna want to see? |
| *contrastive* | Does Anna want to see Mary? |

### 2.2. Data collection

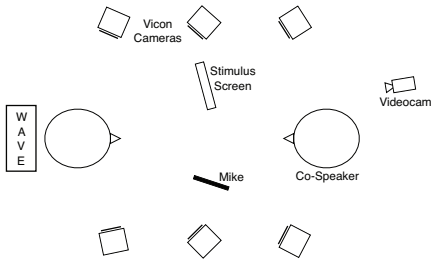
The audio signal, gestures of the vocal tract, and body movement were recorded concurrently. Vocal tract gestures were recorded using an electromagnetic articulometer (EMA; WAVE, Northern Digital), at 100Hz. These data were collected synchronously with the audio signal, which was sampled at 22025Hz ([2]). Body movement was acquired separately using a motion capture system (Vicon; Oxford, UK) which includes 6 infrared sensitive cameras and a visible-light camera, supporting collection of 3D movement data synchronized with video both at a sampling rate of 100Hz ([24]). EMA sensors were placed on the tongue tip, tongue body, tongue dorsum (see Figure 1), on the lower incisors (for jaw movement) and three reference sensors were placed on upper incisors, and left and right mastoid processes (to correct for head movement). Twenty-five motion capture markers were placed on the lips, eyebrows, face, arms and hands, including one marker on each index finger, taped to the participants' skin or to their clothes (Figure 1). In post-processing, data from the concurrently recorded audio, EMA, Vicon and video streams were temporally aligned through cross-correlation of head movement reference data, and trajectories of head-mounted sensors and markers were converted to a coordinate system centered on the upper incisors and aligned with the speaker's occlusal plane. Importantly, the EMA and motion capture systems allow for unrestricted movement, which is necessary for body gesturing and a natural conversational setting.

**Figure 1**: Gesture tracking. EMA sensors on the left and motion capture markers on the right.



Participants were seated with a clear view of a monitor (for stimulus presentation) and a confederate co-speaker (Figure 2). During the experiment, sentences appeared on the monitor; participants read them and pointed with their dominant index finger at pictures as they read the associated target words. A green paper dot was affixed close to the participant's knee to serve as the resting position for the pointing finger (similar to [22]). The co-speaker was monitoring the productions and prompting participants to re-read incorrectly produced sentences as necessary. The day before each experiment, participants had a brief training session to familiarize themselves with the task, stimuli, and novel words.

**Figure 2**: Experimental setup.



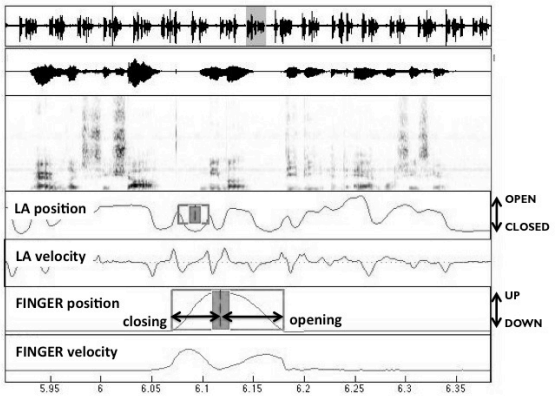

### 2.3. Prosodic verification

During post-processing the utterances were checked for appropriate prosodic structure. Boundaries were analysed using the Tone and Break Indices labeling system ([1]). The anticipated ip and IP boundaries in all sentences in Experiment 1 were realized as IP boundaries and are therefore in the remainder of the paper referred to as IP1 (the original ip condition) and IP2 (the original IP condition). In Experiment 2, the target words were never fully deaccented (the pointing gesture may have interfered; see [17]). This condition was excluded from the analysis.

### 2.4. Data analysis

The data were labeled using a semi-automatic labeling procedure (mview; Haskins Laboratories). The target words were *MIma*, *miMA* (Exp. 1), and *Bob* (Exp. 2). The gestures were labeled on the following trajectories: Lip aperture (LA, the Euclidean distance between upper and lower lip trajectories) for the labial consonants, the tongue dorsum vertical displacement trajectory for the vowels, F0 for the intonation gestures, and the right index finger trajectory for the pointing gestures. For each of these gestures, we identified the following temporal landmarks using velocity criteria (see Figure 3): gesture onset, target, maximum constriction, offset, and peak velocity of the opening and closing movement. Focusing here on the effects of prosodic structure on gesture duration, the following variables are of interest for the LA and pointing gestures: 1) duration of the constriction closing movement (from onset to maximum constriction), 2) duration of the constriction opening movement (from maximum constriction to gesture offset). For the pointing gesture, the closing movement is the movement towards the picture, and the opening movement is the movement returning to the rest position.

**Figure 3**: Labeling example for *Mary would like to see Shaw, MIma, Beebee, and Ann while she is here*. The identified landmarks shown here are: onset (left edge of the box), target (left edge of the shaded box), maximum constriction (dashed line), offset (right end of the box).




We examined the effects of prosodic boundaries (Exp. 1) on the duration of the movements closest to the boundary (constriction closing movement) of the pointing gesture and of the LA gesture for the first consonant in *miMA* and *MIma*. These movements are expected to show the strongest effect of the boundary ([5]). We also tested the effect of the movement further away from the boundary (opening movement) for the LA and pointing gestures. To examine the effect of prominence (Exp. 2), we tested the duration of the opening movement of the first consonant (C1), and the closing movement of the second consonant (C2) in the target word *Bob* (which are related to the nucleus vowel), the closing C1 and opening C2 movement, and both closing and opening movements of the pointing gesture.

## 3. RESULTS

Effects of prosodic structure are presented here for the first speaker. A one way ANOVA shows a main effect of prosodic boundaries on the LA closing movement, and post-hoc analysis (Fisher's PLSD) shows that these gestures are shorter in duration at word boundaries than at IP2 boundaries which are in

turn shorter than at IP1 boundaries. For the LA opening movement (further from the boundary), a one way ANOVA shows an effect of the boundary and Fisher's PLSD shows that gestures at word boundaries are shorter than gestures at IP1 and IP2 boundaries. For the closing movement of the pointing gesture (toward the picture), the main effect of boundary is also significant, and Fisher's PLSD shows that pointing gestures at word boundaries are shorter than at IP1 and IP2 boundaries. There was no effect on the finger pointing gesture opening movement (returning to the resting position). Results are shown in Table 3. (The focus here is on boundaries, so the data for the two stress conditions were *z*-scored and pooled and a one way ANOVA conducted. Note that the results are essentially the same as for a two way ANOVA (factors boundary and stress), where there were some effects of stress, but no boundary *x* stress interactions).

**Table 3**: Results of boundary analyses. Means (std. err.).in *z*-scores, ANOVA, Fisher's PLSD.

| *LA closing movement* | |
|---|---|
| Word = -0.64 (0.19)<br>IP1= 0.56 (0.19)<br>IP2= -0.02 (0.20)<br>$F(2,61)=10.0183$, $p=.0002$ | Word, IP1: $p<.0001$<br>Word, IP2: $p=.0293$<br>IP1, IP2: $p=.0398$ |
| *LA opening movement* | |
| Word = -0.44 (0.15)<br>IP1= 0.25 (0.14)<br>IP2= -0.2 (0.15)<br>$F(2,61)= 6.0397$, $p=.0041$ | Word, IP1: $p=.0012$<br>IP1, IP2: $p=.0329$ |
| *Pointing gesture closing* | |
| Word = -0.51 (0.20)<br>IP1=0.07 (0.19)<br>IP2=0.31 (0.20)<br>$F(2,62)=4.2851$, $p=.0182$ | Word, IP1: $p=.0443$<br>Word, IP2: $p=.0060$ |

Turning to prosodic prominence, a one way ANOVA showed a main effect of prominence on the closing and on the opening movement of C1 (C1 is the first [b] in the target word *Bob*), on the closing movement of C2 (C2 is the second [b] in *Bob*), and on the opening but not on the closing movement of the pointing gesture. Fisher's PLSD show that the manual opening and C1 closing movement are longer in the contrastive than in the broad and narrow conditions, and the C1 opening movement is longer in the contrastive than in the broad condition. C2 closing movement is longer in the broad than in the narrow condition. Results are shown in Table 4.

## 4. DISCUSSION AND CONCLUSIONS

We have demonstrated that EMA and motion capture methods can be used to simultaneously collect acoustic speech data, vocal tract gestures, video, and body movement. This approach allows precise analyses of gestures and gesturing and their coordination.

**Table 4**: Results of prominence analyses. Means (std. deviation) in ms., ANOVA, Fisher's PLSD.

| *C1 closing (LA)* | |
|---|---|
| Broad: 86.36 (9.24)<br>Contrast.: 118.46 (12.14)<br>Narrow: 92.73 (13.48)<br>$F(2,32)=25.4129$, $p<.0001$ | Contr.-broad, $p<.0001$<br>Contrastive-narrow, $p<.0001$ |
| *C1 opening (LA)* | |
| Broad: 134.55 (9.34)<br>Contrast.: 156.92 (22.87)<br>Narrow: 145.46 (15.72)<br>$F(2,32)=4.9797$, $p=0.0131$ | Contrastive-broad, $p=0.0035$ |
| *C2 closing (LA)* | |
| Broad: 117.5 (4.63)<br>Contrast.: 110.83 (15.64)<br>Narrow: 102.5 (7.07)<br>$F(2,25)= 3.5435$, $p = 0.0442$ | Broad-narrow: $p= 0.0136$ |
| *Pointing gesture opening* | |
| Broad: 515.65 (12.61)<br>Contrast.: 569.57(56.05)<br>Narrow: 525.91 (31.98)<br>$F(2,32)= 6.4707$, $p = 0.0044$ | Contr.-broad $p<.0021$<br>Contr.-narrow, $p<.0107$ |

Preliminary results show an effect of prosodic structure such that oral gestures become longer at boundaries and under prominence, in accordance with previous findings (e.g., [4, 19]). We also found evidence that durations of pointing gestures are affected by prosodic structure. The effect is parallel to the effect on oral constriction gestures, namely, manual movement lengthens phrase-initially and under prominence. The effects are not identical to the effects of prosodic structure on oral constriction gestures. For example, for prominence, there is only an effect on the opening movement, rather than—like in the LA gestures—also on the closing movement. While the implications of this discrepancy can only be assessed once the coordination of the pointing gesture to the oral constriction gesture is known, this finding is surprising. However, the fact that there is an effect and the overall similarity of the prosodic effect on manual and oral gestures lend further support to the hypothesis that control of prosodic structure extends beyond the vocal tract.

## 5. ACKNOWLEDGMENTS

The authors thank Mandana Seyfeddinipur, Dolly Goldenberg, Argyro Katsika, and Doug Whalen for their help. This work was supported by NIH grant DC002717 to Doug Whalen.

## 6. REFERENCES


[1] Beckman, M. E., Ayers Elam, G. 1997. Guidelines for ToBI labelling. Version 3.0, unpublished ms. (available online at: http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf).

[2] Berry, J. J. 2011. Accuracy of the NDI wave speech research system. *Journal of Speech, Language, and Hearing Research* 54, 1295-1301.

[3] Browman, C. P., Goldstein, L. M. 1986. Towards an Articulatory Phonology. *Phonology Yearbook* 3, 219-252.

[4] Byrd, D., Saltzman, E. 1998. Intragestural dynamics of multiple phrasal boundaries. *Journal of Phonetics* 26, 173-199.

[5] Byrd, D., Saltzman, E. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 31, 149-180.

[6] Cho, T. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *Journal of the Acoustical Society of America* 117, 3867-3878.

[7] Cho, T. 2006. Manifestation of prosodic structure in articulation: Evidence from lip kinematics in English. In: Goldstein, L. (ed.), *Laboratory Phonology 8: Varieties of phonological competence.* New York: Walter De Gruyter, 519-548.

[8] Cho, T., Keating, P. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics* 29, 155-190.

[9] Cooper, W. E., Eady, S. J., Mueller, P. R. 1985. Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* 77, 2142–2156.

[10] de Jong, K. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America* 97, 491–504.

[11] de Ruiter, J. P. A. 1998. *Gesture and speech production*. Unpublished Ph.D. Dissertation, Radboud University, Nijmegen.

[12] Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R., Lotts, D. W. 1986. Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech* 29, 233–251.

[13] Edwards, J., Beckman, M., Fletcher, J. 1991. The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America* 89, 369-382.

[14] Esteve-Gibert, N., Prieto, P. 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, 850-864.

[15] Fougeron, C., Keating, P. 1997. Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America* 101, 3728-3740.

[16] Katsika, A., Krivokapic, J., Mooshammer, C., Tiede, M., Goldstein, L. M. 2014. The coordination of boundary tones and their interaction with prominence. *Journal of Phonetics* 44, 62-82.

[17] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57, 396-414.

[18] McNeill, D. 1992. *Hand and mind.* Chicago: University of Chicago Press.

[19] Mücke, D., Grice, M. 2014. The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics* 44, 47-61.

[20] Mücke, D., Nam, H., Hermes, A., Goldstein, L. 2012. Coupling of tone and constriction gestures in pitch accents. In: Hoole, P., Pouplier, M., Bombien, L., Mooshammer, Ch., Kühnert, B. (eds.), *Consonant clusters and structural complexity*. Berlin/New York: Mouton de Gruyter, 205-230.

[21] Parrell, B., Goldstein, L., Lee, S., Byrd, D. 2014. Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics* 42, 1-11.

[22] Rochet-Capellan, A., Laboissière, R., Galván, A., Schwartz, J. L. 2008. The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language and Hearing Research*, 51, 1507-1521.

[23] Swerts, M., Krahmer. E. 2010. Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics* 38, 197-206.

[24] Tyrone, M. E., Nam, H., Saltzman, E., Mathur, G., Goldstein, L. 2010. Prosody and movement in American Sign Language: A task-dynamics approach. *Proc. Speech Prosody 2010* Chicago, 100957, 1-4.

[25] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P. J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91, 1707-1717.

[26] Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S. 2004. The timing of speech-accompanying gestures with respect to prosody. *Proceedings of From Sound to Sense* Boston, 97-102.