Speech and manual gesture coordination in a pointing task

Jelena Krivokapic,[1, 2] *Mark K. Tiede,* [2] *Martha E. Tyrone,* [2,3] *Dolly Goldenberg* [2,4]

[1]University of Michigan, USA
[2]Haskins Laboratories, USA
[3]Long Island University Brooklyn, USA
[4]Yale University

jelenak@umich.edu, tiede@haskins.yale.edu, tyrone@haskins.yale.edu, dolly.goldenberg@yale.edu

Abstract

This study explores the coordination between manual pointing gestures and gestures of the vocal tract. Using a novel methodology that allows for concurrent collection of audio, kinematic body and speech articulator trajectories, we ask 1) which particular gesture (vowel gesture, consonant gesture, or tone gesture) the pointing gesture is coordinated with, and 2) with which landmarks the two gestures are coordinated (for example, whether the pointing gesture is coordinated to the speech gesture by the onset or maximum displacement). Preliminary results indicate coordination of the intonation gesture and the pointing gesture.

Index Terms: gestural coordination, manual gestures, EMA, motion capture, data collection methods

1. Introduction

This paper examines the multimodal expression of prosodic structure by investigating how speech and manual gestures are coordinated under prominence. Prominence is manifested in temporal and tonal properties. For English, under prominence, acoustic segments and articulatory gestures lengthen (Turk & Sawusch 1997, Cambier-Langeveld & Turk 1999, Cho 2006), and prominent words are associated with pitch accents (patterns in F0, such as falling or rising pitch, e.g., Pierrehumbert & Hirschberg 1990). These pitch modulations can be understood as tone gestures and they are systematically coordinated with supraglottal constriction gestures (Gao 2008, Mücke et al. 2012).

In addition to speech gestures, communication is also manifested through body gestures. Their salience has been argued to be an essential component of the communication system (e.g., Kendon 1972, 2004, McNeill 1985, 2005, Bernadis & Gentilucci 2006), and neural evidence supports this view (Özyürek, Willems, Kita, & Hagoort 2007, Willems, Özyürek, & Hagoort 2007). In an early linguistic analysis of the relationship between prosodic structure and body gestures, Yasinnik, Renwick, & Shattuck-Hufnagel (2004) found that discrete gestures (movements that have a sudden stop which indicates that they reached their target) tend to occur with pitch-accented syllables. Further studies have examined the relationship between manual movements (beat gestures, finger tapping, or pointing gestures) and prominent syllables or words (e.g., Yasinnik, Renwick, & Shattuck-Hufnagel 2004, Swerts & Krahmer 2010, Mendoza-Denton & Jannedy 2011, Esteve-Gibert & Prieto 2013), and the relationship between eyebrow movement and head nods and prominence (e.g., Hadar, Steiner, Grant, & Rose 1983, Krahmer & Swerts 2007, Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson 2004). These studies suggest that body gestures and prosodic prominence are timed to one another, and thus provide evidence for a link between these two modalities. Evidence for a close relationship between body gestures and prominence also comes from a small set of studies which find that body gestures lengthen under prominence (Kelso, Tuller, & Harris 1983 and Parrell, Goldstein, Lee, & Byrd 2014 for finger tapping; Krivokapic, Tiede, & Tyrone 2015 for pointing gestures). Finally, a number of studies have examined the precise temporal coordination between manual gestures and speech, specifically asking whether there is systematic coordination between them, and which component of speech is coordinated with which component in the manual gestures (McClave 1994, Loehr 2004, Rochet-Capellan et al. 2008, Roustan & Dohen 2010, Leonard & Cummins 2011, Esteve-Gibert & Prieto 2013; see Wagner et al. 2014 and Esteve-Gibert & Prieto 2013 for an overview). While these studies differ in their findings, the most comprehensive studies have identified the gesture apex (i.e., the displacement maximum) as the target of coordination in manual gestures (e.g., Leonard & Cummins 2011, Esteve-Gibert & Prieto 2013). There is less research examining the landmarks that could be the target of coordination in speech, and no stable pattern has been identified so far (see Roustan & Dohen 2010, Leonard & Cummins 2010), although the strongest evidence points to the peak of the pitch accent (Esteve-Gibert & Prieto 2013) as the point of coordination with manual pointing gestures.

The variation in results across studies and the ambiguity regarding the point of alignment in the speech signal is likely to be due in part to the imprecise measurements available to researchers. Previous studies have generally relied on acoustic and video data (e.g., Esteve-Gibert & Prieto 2013), or on kinematic data for body gestures and external oral gestures such as jaw movement (e.g., Rochet-Capellan et al. 2008), or on kinematic properties of speech and constrained body gestures, such a finger tapping (Parrell et al. 2014). However, for a full investigation of the coordination of speech and body gestures, kinematic data are needed for both. In the current study, we use a novel method to simultaneously collect audio, kinematic body movement and speech articulation data and investigate the coordination of pointing gestures and speech under prominence. This allows us to evaluate the following two questions: 1) which speech gesture is the pointing gesture

coordinated with, in other words, is it the tone gesture, as has been claimed (e.g., Mendoza-Denton & Jannedy 2011, Esteve-Gibert & Prieto 2013), or is it the consonant or vowel gesture, which, due to the limitations in previous work, could not be fully examined; and 2) which landmark of the pointing gesture is coordinated with which landmark of the speech gesture? The answer to these questions will allow for an examination of the coordination principles between manual gestures and speech. Based on our understanding of coordination of speech gestures, we expect that the pointing gesture will be coordinated with the tone or the vowel gesture, and that it will not affect the coordination of consonant and vowel gestures (Gao 2008, Mücke, Nam, Hermes, & Goldstein 2012). This prediction is also supported by results from previous studies, which suggest that the apex of the pointing gesture (maximum displacement of the finger) will be timed to the peak of the tone gesture (Esteve-Gibert & Prieto 2013).

2. Methods

2.1. Stimuli and participants

The experiment manipulated stress (levels: stress on the first and stress on the second syllable) and phrase-initial boundary (levels: word, ip, and IP boundary). Three sentences varying in phrase-initial boundary strength (word, ip, and IP) were constructed, with a target word following the boundary. There were two target words differing in stress (*MIma* and *miMA*, with stress on the first and second syllable, respectively), yielding a total of six sentences (see Table 1 for the set with *miMA* as the target word; the sentences with *MIma* were identical except for the target word). These target words were constructed so as to keep the segments in the two stress conditions identical. Participants read the sentences twelve times, for a total of 72 productions (2 stress *x* 3 boundary *x* 12 repetitions). The sentences were pseudo-randomized in blocks of six sentences each. The sentences were presented on a computer screen and 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 the target word *miMA*. The boundary is before the target word.

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.

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

2.2. Data collection

Audio recordings, vocal tract gestures, and body movements were recorded concurrently. Audio was recorded at 22050 Hz with a directional microphone, synchronized with vocal tract gestures tracked in 3D at 100 Hz using a Northern Digital WAVE electromagnetic articulometer (EMA; Berry, 2011). Three sensors were placed on the tongue (tip, body, and dorsum), one on the lower incisors to track jaw movement, and an additional three placed on the upper incisors and mastoid processes were used as references to correct for head movement. A motion capture system (Vicon; Oxford, UK) was used to record body movement. This system uses infrared sensitive cameras and a visible-light camera to track the 3D movement of reflective markers synchronized with video, both at a sampling rate of 100 Hz (Tyrone et al., 2010). Nineteen motion capture markers were taped near the lips and eyebrows, on the arms and hands, including one marker on each index finger, and on the forehead and nose (for head reference alignment with EMA). Data from the Vicon, video streams, audio, and EMA were temporally aligned through cross-correlation of the 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 each speaker's occlusal plane.

During the experiment, the participant was seated facing a computer monitor and a confederate co-speaker. The participant was asked to read the sentences that appeared on the monitor as if reading a story to someone and to point at a picture of a doll (named *MIma* or *miMA*) while producing the associated target word. The doll was always in the same position on the screen. A paper dot was attached near the participant's knee, serving as the resting position for the pointing finger (cf. Rochet-Capellan et al. 2008). The day before each experiment the participants had a brief training session where they learned the novel words (*MIma, miMA*) and familiarized themselves with the task and stimuli. During the experiment, the co-speaker monitored the sentence productions and asked participants to repeat incorrectly read sentences as necessary.

2.3. Data analysis

All utterances were checked for the targeted prosodic structure using the Tone and Break Indices labeling system (Beckman & Ayers Elam 1997). All targeted ip and IP boundaries were produced as IP boundaries.

The movement data were labeled using a semi-automatic labeling procedure (mview; Haskins Laboratories) for the consonant, vowel, and pointing gestures. The target words were *MIma* and *miMA*. For the two bilabial consonants, the gesture was labeled on the lip aperture trajectory (LA, the Euclidean distance between upper and lower lip markers) and for the two vowels, on the tongue dorsum (TD) vertical displacement trajectory. The pointing gesture was labeled on the trajectory of the index finger of the dominant hand. For each gesture, the following temporal landmarks were identified using velocity criteria: gesture onset, target, maximum constriction (or finger displacement), offset, and peak velocity of the closing and opening movement or pointing and returning movement (see Figure 1). F0 was labeled in Praat (Boersma & Weenink 2015). The pitch accent on the target word was L+H*. The onset of the L tone could only be reliably identified in a few cases, and it is therefore not included in the analysis (though it is included in Figures 2 and 4 showing the data for the individual sentences). The onset and target of the pitch movement for the H* part of the pitch accent were labeled manually by identifying the turning point (the low valley indicating the onset of the H*), and the end of the F0 rise was labeled as the target. (This labeling

corresponds to the view of tone gestures as developed within Articulatory Phonology; see Gao 2009, and Mücke et al. 2012 for pitch accents within Articulatory Phonology in comparison to the Autosegemental-metrical-approach). In cases where the L part of the pitch accent could be labeled, the turning point for the L tone (the peak in F0 just before the lowering of F0 for the L tone starts) was identified. For the current analysis, we examined only a subset of the kinematic landmarks (see below). These were identified based on the expectation that coordination occur between onsets of gestures or between targets of gestures and based on the examination of the gestural scores. Future work will examine additional landmarks.

Our analysis focuses on the constriction forming movement of the first consonant, the first and second vowel gesture, and the forward movement of the pointing gesture (i.e., the movement of the finger from the resting position towards the picture of the doll). To examine the coordination of the pointing gesture with the consonant, vowel, and tone gestures, we calculated the intervals from pointing gesture onset (FingOns) to: first consonant onset (C1Ons), first vowel gesture onset (V1Ons), and H tone gesture onset (ToneOns). We also calculated the intervals from the pointing gesture maximum displacement (FingMaxC) to tone gesture maximum constriction (ToneMaxC), first vowel maximum constriction (V1MaxC), and second vowel maximum constriction (V2MaxC). The interval from L tone onset (LOns) to FingOns was also calculated, but since it was available for only one sentence, it is included for illustrative purposes only, not in the statistical analyses. The durations of

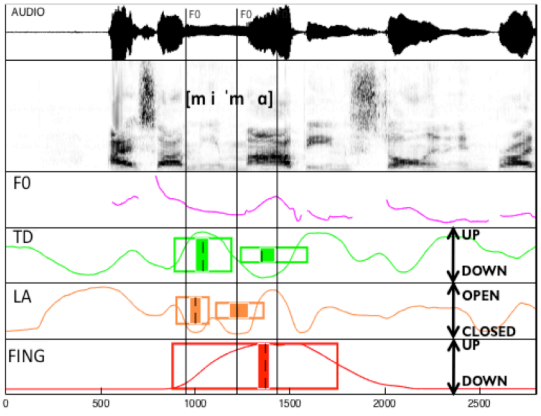

Figure 1. *Sample token with the word miMA, sentence "There are other things. I saw miMA being stolen in broad daylight by a cop." The rectangles show labeled vocal tract and manual gestures. The whole rectangle is the gesture, the filled part represents the gesture nucleus, the dashed line represents the time of maximum constriction/displacement. TD: the tongue dorsum vertical displacement trajectory (vowels [i] and [a]), LA: the lip aperture trajectory (for consonants [m]), FING: left finger vertical displacement trajectory for the pointing gesture, F0: pitch contour for the pitch accent L+H* (tone gesture); labels show the onset of L, onset of H* and target of H*. The pitch labeling was done in Praat and is included here for illustrative purposes only).*

the intervals were *z*-scored by prosodic boundary in order to remove the effects of the boundary, as they were not of interest for this experiment.

We analyzed coordination using a two-factor ANOVA, with the factors stress (first or second) and interval (the six intervals described above), and the dependent variable the duration (of the intervals). We are interested in two properties of the intervals: which interval has the shortest duration and which interval has the smallest variance (see, e.g., Leonard & Cummins 2011). Given that data from only one speaker have been analyzed thus far, variance could not be evaluated statistically, but it is shown in Figure 2.

3. Results and discussion

To get a sense of the data, Figure 2 shows the raw scores. Note that a *positive* duration value indicates that the finger movement event preceded the speech event it is compared to.

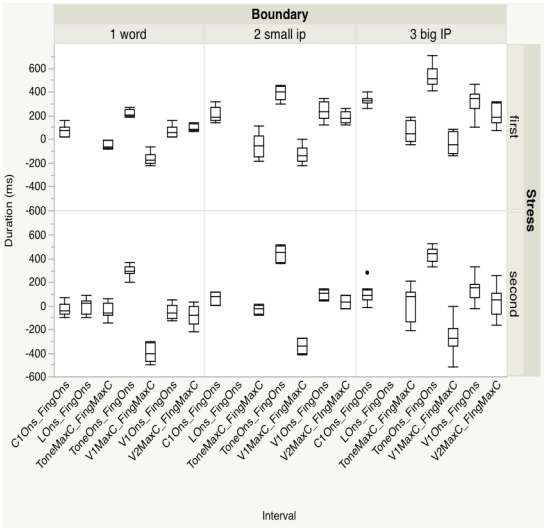

Figure 2*. Duration of intervals, grouped by stress and boundary. The first part of the interval name denotes the first term of the subtraction (e.g., ToneOns_FingOns is calculated by subtracting the time of onset of the pointing gesture from the time of onset of the pitch accent (tone gesture)). Note that LOns_FingOns is given for one sentence which was the only case in which it could be identified a sufficient number of times.*

Figure 4 shows the temporal ordering and the duration of the examined gestures.

The results show a main effect of stress ($F(1, 270)=145.6953$, $p<.0001$), interval ($F(5, 270)=264.1534$, $p<.0001$), and an interaction of the two factors ($F(5, 270)=15.6325$, $p<.0001$). The effect of stress is such that the intervals are longer when stress is on the first syllable (0.27) than when it is on the second syllable (-0.25), reported in *z*-scores. The interaction effect is shown in Figure 3. Tukey HSD comparison of the intervals shows that the shortest intervals are V1Ons_FingOns (stress on second syllable), C1Ons_FingOns (stress on second syllable), ToneMaxC_FingMaxC (stress on first, stress on second syllable), V2MaxC_FingMaxC (stress on second syllable). These intervals are shorter than all the others and not different from each other.

Of particular interest is the lack of a stress effect on the intervals between landmarks of the tone gesture and of the finger gesture. As can be seen in Figure 3, and as confirmed by the Tukey HSD comparison, there is no effect of stress on the timing relations between tone and finger gestures, but there is a stress effect on the timing relations between the oral gestures and finger gestures, indicating that the intervals between tone and finger gestures are not just the shortest, but also the most stable intervals. Note also that the LOns_FingOns interval, as can be seen in Figure 2, compares in length to the shortest intervals. While not conclusive, these findings suggest a stable temporal relation between the manual pointing gesture and the tone gesture, providing further support for the hypothesis that manual gestures are coordinated with pitch (Yasinnik et al. 2004, Mendoza-Denton & Jannedy 2011, Esteve-Gibert & Prieto 2013). However, without analyses of variance and further analysis of the L gesture, the landmarks and properties of coordination cannot be established.

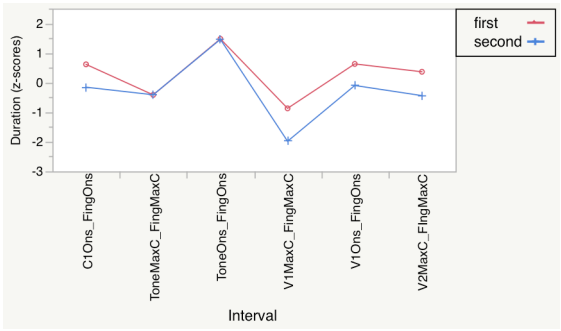

Figure 3*. The interaction of stress and interval effect on the duration of the intervals (the lags between temporal landmarks of the gestures).*

4. Conclusions

We examined the coordination of body and speech gestures. Specifically, we examined which speech gesture (vowel, consonant, or tone gesture) the manual pointing gesture is coordinated with. The second question we examined is what the kinematic landmarks of coordination are. Preliminary results suggest that the pointing gesture is coordinated with the tone gesture. The exact coordination landmarks will be investigated in more detail.

5. Acknowledgements

We are grateful to Mandana Seyfeddinipur, Argyro Katsika, Anna Stone, Lauren McGee, and Douglas Whalen for their help. We would also like to thank Stefanie Shattuck-Hufnagel for her continuing inspiration and invaluable discussions about speech, prosody, and body gestures in particular. This work was supported by NIH DC002717 to Douglas Whalen and NIH DC-012350 to Mark Tiede.

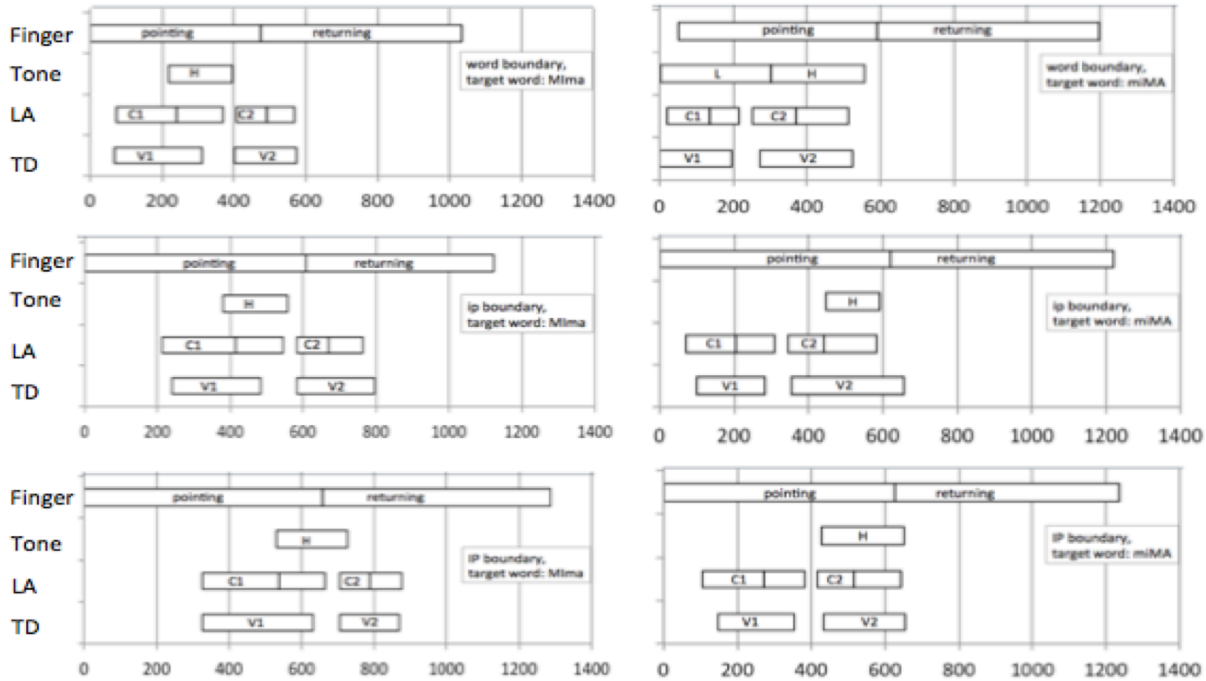

Figure 4. *Gestural scores (in ms) of the target words. Boxes denote duration of the finger gesture, the H part of the L+H* pitch accent (and L in one sentence), consonants (C1 and C2) and vowels in the target words (MIma and miMA). Vertical lines indicate maximum constriction/displacement. The release of V1 (from maximum constriction to the end of the gesture) often overlapped with the constriction forming movement of V2 (from V2 onset to V2 maximum constriction). Because of this, only the constriction forming part of the vowel gestures are shown.*

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

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

[4] Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.22, retrieved on October 8, 2015 from http://www.praat.org/.

[5] Cambier-Langeveld, T. & Turk, A. (1999). A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics*, 27, 171-206.

[6] 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.* Berlin/New York: Walter De Gruyter, pp. 519-548.

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

[8] 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.

[9] Gao, M. (2008). *Mandarin tones: an Articulatory Phonology account.* Ph.D. Dissertation, Yale University.

[10] Hadar, U., Steiner, T. J., Grant, E.C., & Rose, F.C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*. 1983, 26, 117-129.

[11] Kelso, J. A. S., Tuller, B., & Harris, K. (1983). A "dynamic pattern" perspective on the control and coordination of movement. In: MacNeilage, P. (Ed.). *The production of speech*. NewYork: Springer-Verlag, pp. 138-173.

[12] Kendon, A. (1972). Some relationships between body motion and speech. In: Seigman, A. & B. Pope (Eds.). *Studies in dyadic communication*. Oxford: Pergamon Press, pp. 177-210.

[13] Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

[14] Krivokapić, J., Tiede, M., & Tyrone, M. (2015). A kinematic analysis of prosodic structure in speech and manual gestures. *Proceedings of ICPhS 2015.*

[15] Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26, 1457-1471.

[16] Loehr, D. (2004). *Gesture and intonation*. Ph.D. Dissertation, Georgetown University.

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

[18] McNeill, D. (1985). So you think gestures are non-verbal?, *Psychological Review*, 92, 350-371.

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

[20] McNeill, D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

[21] Mendoza-Denton, N., & Jannedy, S. (2011). Semiotic layering through gesture and intonation: A case study of complementary and supplementary multimodality in political speech. *Journal of English Linguistics*, 39, 265-299.

[22] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15, 133-138.

[23] 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, C., Kühnert, B. (eds.), *Consonant clusters and structural complexity*. Berlin/New York: Mouton de Gruyter, 205-230.

[24] Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, 19, 605-616.

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

[26] Pierrehumbert, J., & Hirschberg, J. (1990). The Meaning of Intonation in the Interpretation of Discourse. In: *Intentions in Communication*. Edited by P. Cohen, J. Morgan & M. Pollack. Cambridge, MA: MIT Press, pp. 271-311.

[27] 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.

[28] Roustan, B. & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: temporal coordination and effects on the acoustic and articulatory correlates of focus. In: *Proceedings of Speech Prosody 2010*, 11-14 May 2010.

[29] 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.

[30] Turk, A. E., & Sawusch, J. R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, 25, 25-41.

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

[32] Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.

[33] Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, 17, 2322-2333.

[34] 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.