Are torso movements during speech timed with intonational phrases?

Stefanie Shattuck-Hufnagel[1], *Pei Lin Ren*[1] *and Elizabeth Tauscher*[2]

[1] Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA USA [2] Wellesley College, Cambridge MA USA

sshuf@mit.edu, pelire@mit.edu, etausch@wellesley.edu

Abstract

It is well known that speakers often move their arms, hands, heads and parts of their faces in conjunction with their speech. Recent studies indicate that these movements are often temporally aligned with the accented syllables in the spoken utterances they accompany, forming a kind of gestural analogue to the accentual aspect of phrase-level prosody. But less attention has been given to the question of how body movements might relate to the other major aspect of phrase-level prosody: word grouping or phrasing. In this study we examined the temporal alignment of torso movement with the intonational phrases identifiable in short speech samples from two speakers selected from a corpus of academic lecture videos. Results show that a) the two speakers differ substantially in their torso movements during speech, b) the speaker who showed the most frequent use of left-right movement of the shoulders timed these movements to coincide to a notable extent with his intonational phrases, and c) torso movements are sometimes timed in other ways, such as during silences between spoken phrases, possibly in conjunction with a change in topic. Overall, these observations support the hypothesis that both of the grammatically-significant aspects of phrase-level prosody, i.e. prominence and phrasing, can have analogues in the organization of body movements that accompany speech production. This strengthens the view that speech production planning involves the coordination not only of gestures within the vocal tract with each other and with prosodic structure, but also of gestures and movements of other parts of the body, whose contribution to the communicative act merits further exploration.

Index Terms: gesture, intonational phrases, speech-accompanying movement, speech production planning

1. Introduction

Theories and observations of the gestures that accompany (and in some cases even replace) spoken utterances have clearly demonstrated that these movements are often systematically related to the speech they accompany, and play a significant role in the communication process (Kendon 1975, 2004, McNeil 1979, 1996, 2005 and many colleagues). These works have provided a strikingly-useful set of categories for the analysis of such gestures, largely based on their relation to (and potential contribution to) the meaning of the speech. To some extent, however, the question of how the gestures and speech prosody might be planned together has been looked on as a less pressing question.

This question is part of a larger issue, which concerns the nature of the cognitive representations that govern all aspects of a communicative act. Recent decades have seen the emergence of explicitly-articulated linguistic grammars of phrase-level prosody (Liberman and Prince 1977, Selkirk 1984, Pierrehumbert 1980, Beckman and Pierrehumbert 1986, Nespor and Vogel 1986, inter alia), and these advances have made it possible to ask whether the representations of prosodic prominence and grouping proposed by these grammatical theories play a significant role in the cognitive process of speech production planning. Resulting studies have clearly shown that the answer to this question is yes; for example, many heretofore mystifying patterns of phonological alternation and phonetic variation in spoken utterances have been shown to reflect the influence of phrase-level prosodic grouping and accent, rather than the more traditionally considered morphosyntactic structures of the sentences these utterances represent (Feirrera 1993, Shattuck-Hufnagel et al. 1994, Dilley et al. 1996, Fougeron and Keating 1997, Cho and Keating 2001, Jun 2005 and many others). Keating and Shattuck-Hufnagel (2002) have proposed a conceptual model of speech production planning in which an abstract representation of the prosodic structure for a particular utterance of a planned sentence undergoes gradual and increasingly detailed specification of its morphosyntactic and lexical content, so that the interaction of the morphosyntactic/lexical information with the prosodic structure determines the surface phonetic form of the utterance.

Such a line of thinking leads quite naturally to the question of whether prosodic structure also governs the production of speech-accompanying gestures, and a number of studies have addressed this question for phrase-level pitch accents (Keating et al. 2003, Swertz and Krahmer 2006, Loehr 2004, Yasinnik et al. 2004, Renwick et al. 2004). These studies offer support for the hypothesis that certain gestures are timed with respect to the prosody of their spoken utterances, in particular with respect to the prominent words and syllables in speech. But less attention has been paid to the hypothesis that prosodic phrasing structure governs body movements as well. This paper addresses the question of whether the movements of the body sideways in space are also prosodically governed, reflecting the intonational phrasing structure (rather than the accentual structure) of the speaker's utterances.

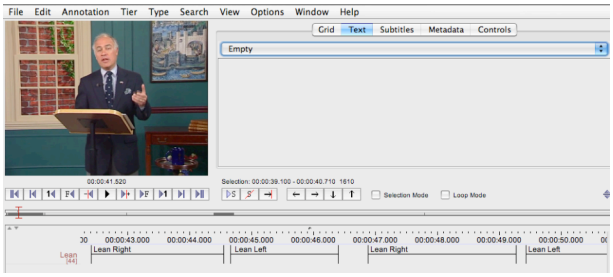
2. Methods

To determine the extent to which the intonational phrases of speech are aligned with the right-to-left sideways movement of the speaker's torso, both the intonational phrasing and sideways movements of video samples from two speakers delivering academic lectures were separately labeled, and then recombined to determine their degree of temporal overlap.

2.1. Video materials and labeling of sideways torso movements

The video materials were selected from a corpus of commercially-available recorded academic lectures; the speakers in this material had been pre-selected for their skill in delivering such lectures. A sample of two minutes of speech was selected for each of two male lecturers. The video file was separated from the accompanying audio file, so that the torso movements could be labeled without listening to the accompanying speech, avoiding any possible effect of the intonational phrasing on the labeler's perception of the torso movements. Each video sample was labeled by an experienced movement labeler for the beginning and end of each sideways movement of the upper torso in space, here called a 'lean', using the ELAN software available from the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands (www.lat-mpi.eu/tools/). An example of such an annotation is shown in Figure 1.

Figure 1: Screenshot of Elan display with 'lean' labels aligned in time.



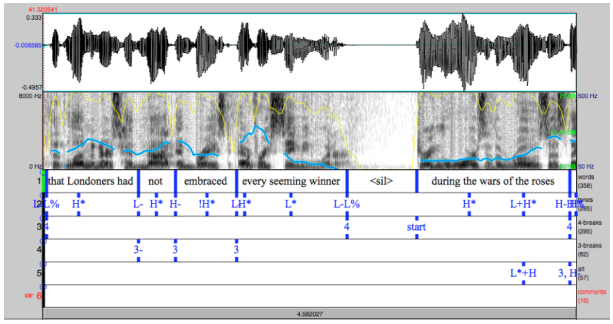
It was observed that speakers usually moved their torso from side to side by shifting their weight from one foot to the other, but sometimes by leaning sideways from the waist or hips; these different types of movement are not distinguished in this report, which simply records the side-to-side movement (horizontal translation) of the shoulders with respect to the background verticals. In addition, some of the apparent sideways movement in relation to the background was caused by the turning of the speaker's body (e.g. in response to a change of the speaker's focus from one camera location to another) rather than by leaning from the waist or shifting weight from one foot to the other; again, this aspect of sideways movement was not distinguished in the labels. It would be of considerable interest to distinguish these different aspects of upper torso movement in future studies, as they may align differently with the structure of the accompanying speech. For example, one of the speakers seemed to rotate slightly just before a strong sideways movement, almost as a preparation movement.

The beginning and end of each sideways torso movement was determined by visually tracking the movement of the speaker's shoulders, which were conveniently displayed against a number of verticals in the video background (e.g. a window with vertical dividers, vertical paneling on the wall and a mounted painting in a frame). In addition, Elan's capacity to advance or reverse the video frame-by-frame and to play the video for a designated temporal region as an independent segment were particularly useful. As a result, it was possible to determine the beginning and end of each translatory movement of the shoulders from left to right or right to left (as seen by the viewer) quite accurately. In many cases the initial and final portions of shoulder movement were considerably slower than the rest of the movement; no effort was made to estimate the changing velocity of the movement in this study, but future studies may find such an analysis useful, since there was some indication that the regions of faster movement aligned more precisely with the spoken phrases than the overall movement did. In some cases, the head was observed to continue moving in the same direction after the shoulders stopped moving, or even to increase its rate of movement, suggesting an additional head gesture not unrelated to the sideways torso movement, but this aspect of the speaker's movement was not included in the labels.

2.2. Speech materials and labeling of intonational phrases

For each sample, the sound file was separated from the video, and labeled independently for intonational phrasing by an experienced ToBI labeler, using the labelling capacity in the Praat software (http://www.fon.hum.uva.nl/praat/). This separation ensured that there was no biasing effect of the torso movement patterns on the perceived phrasing. Such precautions are necessary because of related findings that visual cues may influence the perception of pitch accents in speech (Swerts & Krahmer 2006, Keating et al. 2003, Dohen et al. 2004). Intonational phrase labels included Full Intonational Phrase boundaries (Break Index 4 in the ToBI system) and Intermediate Intonational Phrases (Break Index 3 in the ToBI system); for more information on the ToBI transcription system for American English see Silverman et al. 1992, Petrelli et al. 1994, the ToBI Reference Guide at http://www.ling.ohio-state.edu/~tobi/ame_tobi/, and the ToBI Tutorial at MIT's OpenCourseWare site http://ocw.mit.edu/. The more controversial 2-boundary label, which does not represent an intonationally marked constituent, was also marked, but results for this marker will not be reported here. In addition to the conventional ToBi notation for the end of each intonational phrase constituent, a time marker was added for the start of each phrase, to ensure that the regions of silence or pauses between phrases could be easily identified. An example of such annotation is shown in Figure 2.

Figure 2: A sample prosodic label for the sound file from the video sample shown in Figure 1. Word groups in the Word tier indicate Intermediate Intonational Phrases (iP's; Break Index 3 in the ToBI system, also noted in the 3-breaks Tier). The start and end points of each Full Intonational Phrase (IP) are also labeled in the 4-breaks Tier.

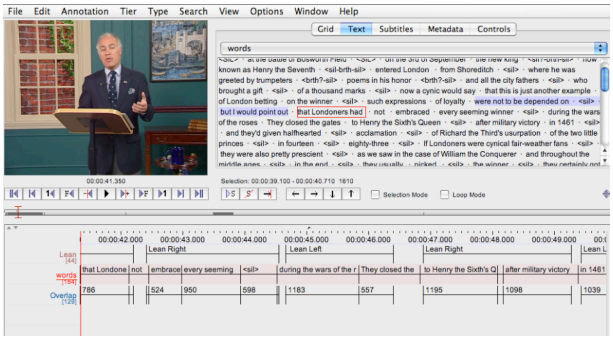


2.3. Alignment of Sideways Torso Movements and Intonational Phrases

To estimate the degree of overlap between the spoken intonational phrases (iPs) and episodes of sideways movement of the speaker's torso, the Praat TextGrid labels for phrasing

were imported into Elan, where they were aligned with the 'lean' labels. The regions of temporal overlap between the two sets of elements were determined, using the appropriate Elan command. An example of such an integrated annotation is shown in Figure 3.

Figure 3: Screen shot of an Elan display showing regions of side-to-side movement of the shoulders in the Leans tier, phrase labels (Intermediate Intonational Phrases corresponding to ToBI break index 3, shown as groupings in the Words tier) and regions of overlap, shown as milliseconds in the Overlap tier. Elan displays the text of the individual phrases in the Words tier in the top right panel of the illustration.



3. Results and Discussion

Results of this analysis have revealed a number of preliminary findings of interest with respect to the timing of torso movements in speech production. First, the two speakers sampled here showed different patterns of side-to-side torso movement (Table 1). For example, in two minutes of speech, Speaker 1 produced 65 phrases and 36 sideways torso movements, while Speaker 2 produced 122 phrases and 16 such torso movements.

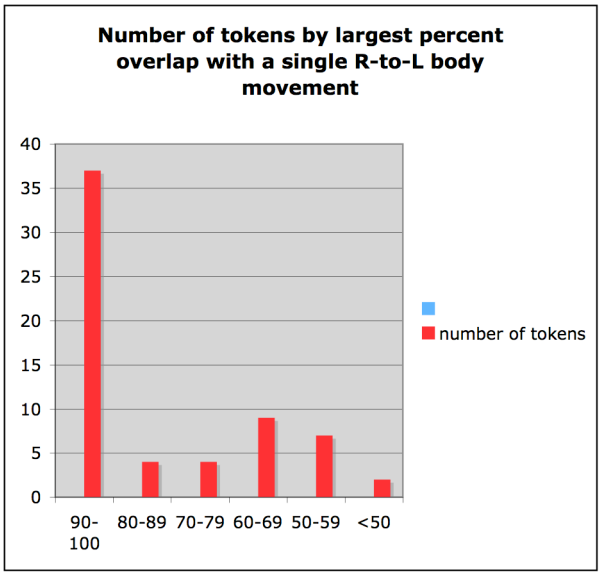
Table 1: Number of Intermediate Intonational Phrases, silences and sideways torso movements ('leans') in 2 minutes of academic lecture style speech from each of two male speakers.

Speaker	#mins	#iPs	#sil	#leans
M1	2	65	35	36
M2	2	122	59	16

These observations are consistent with the hypothesis that different speakers exhibit different amounts of torso-translation during speech, at least in this academic lecturing context; they also indicate substantial differences in prosodic structure, such as habitual phrasing and pausing patterns.

Second, the speaker who exhibited the largest number of torso-translation episodes showed a noticeable degree of alignment between these episodes and the intonational phrases in his speech. Figure 4 shows that the majority of the 65 Intermediate Intonational Phrases in a sample of 2 minutes of speech from this speaker overlapped substantially with a single 'lean'. In some cases a single lean enclosed two or more phrases, and future analyses will determine whether these grouped phrases correspond to a sequence of Intermediate Intonational Phrases (iP, ToBI Break Index 3) combined into a single Full Intonational Phrase (IP, Break Index 4).

Figure 4: The distribution of 65 Intermediate Intonational Phrases into categories by the largest percent overlap with a single sideways torso movement (Speaker 1, 2 minutes of speech). The majority of phrases overlap with a single lean; a few tokens have a largest percent overlap less than 50%, because they overlap with a sequence of 3 different movement labels, e.g. a L-lean, a region of no movement and a R-lean.



Interestingly, in some cases, the torso movement continued through all or most of the pause that followed the intonational phrase, suggesting that this pause was in some sense grouped by the speaker with the preceding phrase. In dialogue contexts, such continued torso movements might help the speaker hold the floor (Jennie Shen, personal communication.) As noted above, in some cases a torso movement continued across two or more intonational phrases, possibly providing a cue to their grouping into a larger prosodic constituent.

Speaker 2, who moved his upper torso less often during speech, also showed a more complex pattern of upper body movement (including forward and backward leans), and his pattern of overlap between torso translation episodes and prosodic phrasing is still being analysed. Preliminary observation showed some interesting patterns of relationship with the speech when he did move his torso. For example, this speaker provided one example in which a substantial torso movement occurred during a pause between two intonational phrases, at a significant moment in the monologue discourse, i.e. when he was introducing a new topic or discourse segment. This same speaker was the one who sometimes showed the 'preparatory shoulder rotation' just before a significant sideways torso movement, almost like a windup before throwing a ball.

An additional observation of interest concerns one token in which the vowel of the word 'not', which was labeled as a single iP, overlapped with a non-movement region, while the onset /n/ and coda /t/ closure overlapped with the preceding and following leans. Earlier work by Yasinnik et al. (2004) and Renwick et al. (2004) suggested that accent-lending hand gestures with sudden sharp end points ('hits') were exquisitely timed with respect to accented syllables in the speech signal,

but unpublished work by Wang (2007) suggested that head hits were likely to be slightly delayed, often into a following weak syllable. This raised the possibility that movements of larger body articulators might be either delayed or less closely timed with spoken prosodic accents than those of smaller articulators which have less mass and momentum. Future analyses of the timing of torso movements will allow us to test this hypothesis with regard to a very large heavy articulator and phrase boundaries. Additional analyses are addressing the question of whether torso movements group silent regions in the speech together with preceding or following phrases, and similarly whether regions of no movement between 'leans' are aligned with spoken phrases in ways that suggest conjunction with the preceding or following 'lean', or whether they function as a separate type of torso positioning pattern.

4. Conclusions

This study explores the relationship between movements of the body during speech and the intonational phrasing structure of the speech. Although the alignment in time between these two kinds of elements was not absolute, there is evidence that in this sample the two streams of behavior are integrated. This finding offers some support for the hypothesis that speech production planning for prosodic structure and body-gesture movements is coordinated, and raises interesting questions about how this coordination is achieved, and how it serves the communicative process.

5. Acknowledgments

We gratefully acknowledge the support of MIT's Undergraduate Research Program for this research, and the participation of students in MIT's Freshman Advisory Seminar on Prosody, 2008 and 2009.

6. References

Beckman, M. and Pierrehumbert, J. (1986), Intonational Structure in Japanese and English. *Phonology Yearbook* 3, pp. 255–309

Cho, T. and Keating, P. (2001) Articulatory strengthening at the onset of prosodic domains in Korean. *Journal of Phonetics* 28:155-190

Dilley, L., Shattuck-Hufnagel, S. and Ostendorf, M. (1996), Glottalization of vowel-initial syllables as a function of prosodic structure, *Journal of Phonetics* 24, 423-444

Dohen, M., Loevenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, 44, 155-172

Ferreira, F. (1993). The creation of prosody during sentence production. *Psychological Review*, *100*, 233-253.

Fougeron, C. and Keating, P.A. (1997), Articulatory strengthening at edges of prosodic domains, *Journal of the Acoustical Society of America* 101, 3728-3740

Jun, S.-A. (2005), *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford: Oxford University Press

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. *Proceedings of the Fifteenth International Conference on Spoken Language Processing,* 2071–2074

P. Keating, T. Cho, C. Fougeron, and C. Hsu (2003), Domain-initial articulatory strengthening in four languages. In *Phonetic Interpretation (Papers in Laboratory Phonology 6),* edited J. Local, R. Ogden, R. Temple, Cambridge University Press, pp. 143-161

Keating, P.A. and Shattuck-Hufnagel, S. (2002), A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics,* **101**, pp. 112-156

Kendon, A., Harris, R.M. and Key, M.R. (1975), *Organization of Behavior in Face-to-Face Interaction.* Mouton.

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

Liberman, M. and Prince, A. (1977), On Stress and Linguistic Rhythm, *Linguistic Inquiry* 8, 249-336.

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

McNeill, D. (1979), *The Conceptual Basis of Language.*

McNeill, D. (1996), *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press.

McNeill, D. (2005), *Gesture and Thought.* University of Chicago Press.

Nespor, M. and Vogel, I. (1986/2007). *Prosodic phonology*. Dordrecht: Foris Publications

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* PhD. Dissertation, Massachusetts Institute of Technology.

Pitrelli, J., Beckman, M. and Hirschberg, J. (1994), Evaluation of Prosodic Transcription Labeling Reliability in the Tobi Framework. *Proceedings of the International Conference on Spoken Language Processing.* Yokohama, Japan.

Renwick, M., Yasinnik, Y. and Shattuck-Hufnagel, S. (2004), The timing of speech-accompanying gestures with respect to prosody, *Journal of the Acoustical Society of America* 115, 2397.

Sekirk, E.O. (1984), *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge, Mass.

Shattuck-Hufnagel, S., Ostendorf, M. and Ross, K. (199X), Stress shift and early pitch accent placement in lexical items in American English, *Journal of Phonetics* 22, 357-388

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg (1992), ToBI: a standard for labelling English prosody. *Proceedings of ICSLP* 92. 867-870

Swerts M. and Krahmer, E. (2006), The importance of different facial areas for signalling visual prominence. In *Proceedings of the International Conference on Spoken Language Processing* (Interspeech 2006), Pittsburgh

Wang, J. (2007), *Relation of gestures to prosody*. Unpublished MIT MS. submitted to fulfill MIT's Advanced Undergraduate Program requirement.

Yasinnik, Y., Renwick, M. and Shattuck-Hufnagel, S. (2004), The timing of speech-accompanying gestures with respect to prosody. In *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 11-13 June 2004, Cambridge, MA. <http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Friday%20Posters/FA-Yasinnik-STS-MAC.pdf>