Multimodality in prominence production and its sensitivity for lexical prosody

Gilbert Ambrazaitis[1] and David House[2]
[1]*Linnaeus University, Växjö,* [2]*KTH (Royal Institute of Technology), Stockholm, Sweden*

Prominence is essentially a multimodal phenomenon, as verbal prominence markers such as pitch accents are frequently accompanied by so-called beat gestures, typically produced by the hands, the head, or the eyebrows [1-3] and visual beats seem to add to perceived prominence [4-6]. However, our understanding of gesture-speech integration is still far from complete. One important issue currently under discussion is whether the verbal and the visual modality tend to interact in a compensatory fashion in prominence production, where less verbal effort is required when a visual beat is added or vice versa, or whether multimodal prominence is rather cumulative in nature, with pitch accents and beat gestures reinforcing each other [3-7]. Another unresolved question is how visual beats interact with lexical-prosodic structure. So far, previous studies have only more or less implicitly suggested a link between beats and lexical prosody, as gesture strokes have been shown to associate and temporally align with lexically stressed syllables [2][8-9].

The present study addresses these two issues by examining the phonetic realization of pitch accents in Swedish as a function of accompanying beat gestures by the head and the eyebrows, as well as of the lexical-prosodic structure of the accented words. Concerning the latter, Swedish exhibits lexical stress and tone, which are both connected to the language's well-known "word accents" (Accent 1, Accent 2; henceforth, A1, A2): Tone is lexically specified only in A2 single-stress words, while it is post-lexical in A1 as well as in A2 words with additional secondary stresse(s) (compounds and some derivations) [10]. We can thus distinguish between three lexical-prosodic categories: (i) simplex stress (A1), (ii) simplex stress + tone (A2) and (iii) compound stress (A2). In addition, Swedish distinguishes between two tonal prominence levels, as we can distinguish between a "small" (HL – realized HL* for A1 and H*L for A2) and a "big" accent (HLH – realized (H)L*H or H*LH) (terminology adopted from [10]).

Our study is based on 12 minutes of audio and video data (1936 words) from five Swedish television news anchors (two female). The material was annotated for big accents (BA), head beats (HB) and eyebrow beats (EB) (Fleiss' kappa: $\kappa_{BA} = 0.77$; $\kappa_{HB} = 0.69$; $\kappa_{EB} = 0.72$), as well as for tonal targets. The realization of the falling (HL) and the rising part (LH) of big accents (fall/rise range in semitones) was examined in three lexical-prosodic conditions (see i-iii above), comparing three (multimodal) constellations of prominence markers: (a) words produced with a BA only (without a beat gesture), (b) words with BA co-occurring with a HB (BA+HB) and (c) words with BA and both HB and EB (BA+HB+EB).

The results suggest a slight tendency for a positive correlation (i.e. a cumulative relation) between the presence of visual beats and the realization of the BA-rise (Fig. 1a). For the accentual fall, the situation is less clear and more complex (Fig. 1b). For both measures, the results suggest a strong interaction between lexical prosody and the constellation of prominence markers. For instance, the combination of a head and an eyebrow beat, but not a head beat alone, seems to induce an enlarged BA-rise in simplex stress words (either A1 or A2), while for the compound stress words, an enlarged pitch range is observed even with the addition of a head beat alone (Fig. 1a). Crucially, the interaction between the multimodal prominence constellation and the lexical-prosodic condition is significant both for the BA-rise *($p=.017$)*[i], and the accentual fall *($p<.001$)*, while there is no significant main effect of the multimodal prominence constellation on either of the two measures. The results are in line with the idea of an essentially cumulative relation between verbal and visual prominence markers in production, at the same time suggesting an interaction of the two modalities that is sensitive for lexical prosody, an issue that needs to be further examined in future research.

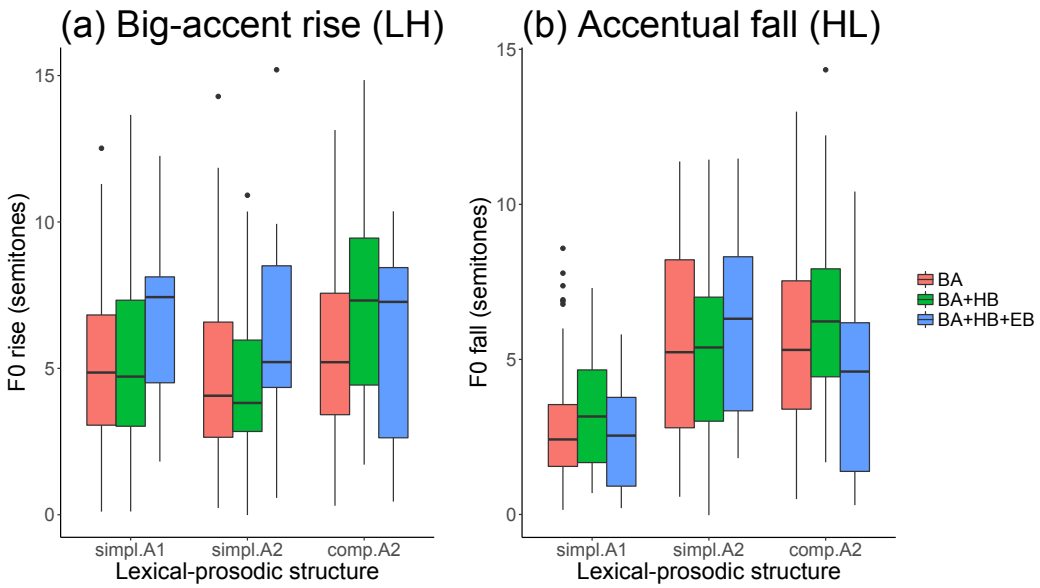

Figure 1. *Boxplots for the BA-rise (a) and the preceding accentual fall (b) measured in semitones as a function of the multimodal prominence constellation (colors) and the lexical-prosodic condition (x-axis: simplex stress (Accent 1), simplex stress + tone (Accent 2), compound stress (Accent 2)); n_{BA}=276, n_{BA+HB} =178, $n_{BA+HB+EB}$=73.*

[1] 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.

[2] Loehr, D. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Journal of the Association for Laboratory Phonology* 3, 71-89.

[3] Ambrazaitis, G., & House, D. 2017. Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication* 95, 100-113.

[4] 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.

[5] Dohen, M., & Loevenbruck, H. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech* 52, 177–206.

[6] Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. 2015. Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics* 49(1), 41-54.

[7] 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. *Proceedings of Speech Prosody 2010* (Chicago, IL, USA).

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

[9] 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* 56 (3), 850-864.

[10] Myrberg, S., & Riad, T. 2015. The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics* 23(2), 115-47.

[i] Tested by means of likelihood-ratio tests based on linear mixed regression models with the multimodal prominence constellation (3 levels), lexical-prosodic structure (3 levels), and speaker sex (2 levels) as fixed factors, assuming random intercepts and slopes for speaker.