Phonetic realization of Swedish complex pitch accents as a function of accompanying head and eyebrow movements

Gilbert Ambrazaitis[1] *& David House*[2]

[1]*Linnæus University, Växjö, Sweden*
[2]*KTH (Royal Institute of Technology), Stockholm, Sweden*

In this study we are examining prosodic prominence from a multimodal perspective. Our research question is whether the phonetic realization of tonal and intonational prominence varies as a function of accompanying head and eyebrow movements (*beat gestures,* cf. below). More specifically, we explore the realization of the tonal and intonational components of so-called *focal or big accents* in a Stockholm Swedish corpus of television news readings.

Swedish makes use of pitch contrasts at the lexical level, distinguishing between two so-called word accents (Accent 1, Accent 2). Orthogonal to this word accent contrast, many varieties of Swedish, including the Stockholm dialect studied here, distinguish between two pitch-related phonological prominence levels, where the higher level has been commonly referred to as a *focal* accent or more recently as a *big* accent [1], as opposed to the *non-focal* or *small* accent [1]. Crucially, the distinction between Accent 1 and Accent 2 is encoded at both levels. According to Bruce's seminal analysis [2], the big accent can be conceived of as a complex pitch accent composed of the tonal configuration for the word accent (Accent 1 or 2) and a following high tone H (the *sentence accent* in his analysis [2]) which is realized as a rise in pitch from the accentual L (HL* in Accent 1, H*L in Accent 2). Although the details of tonal representation of Swedish word accents, as well as the question of the lexicality of tones involved, is much debated (e.g. [1][2][3]), there is a certain consensus on the compositional nature of big accents, as well as the assumption that the tonal components of a big accent relate to different prominence levels and different domains in the prosodic hierarchy of Swedish [2]. In this abstract, we simplify the debate by referring to the tones defining the word accents as *tonal*, and the subsequent rise (the H tone; i.e. the sentence accent [2]) that distinguishes small from big accents as *intonational*.

Previous research on co-speech gestures and audio-visual prosody strongly suggests that prosodic prominence is indeed an audio-visual, or multimodal, phenomenon, as pitch accents (verbal prominence cues) are frequently accompanied by movements of the hands, the head and certain facial areas (visual cues), also referred to as beat gestures, e.g. [4][5]. It has, moreover, been shown that visual and verbal prominence cues may co-occur in various constellations [6][7] and that beat gestures are more likely to occur with *perceptually strong* accents than with *weak* ones: Swerts and Krahmer [7] found in their study of Dutch news readings that the more accented a word was on an auditory scale, the more likely the word was to also be accompanied by a head movement, an eyebrow movement or both. Hence, we might predict a *cumulative relation* of verbal and visual prominence cues, i.e. a positive correlation between the acoustic strength of a pitch accent (e.g. in terms of segmental durations, F0 peak height or F0 ranges) and accompanying beat gestures. In this study, we test this prediction for the special case of complex (big, cf. above) pitch accents in Stockholm Swedish, thereby adding a prosodic-phonological dimension to our research question: Do we find a cumulative relation between the occurrence of head and eyebrow movements and (a) the tonal, (b) the intonational, or (c) both components of a big accent in Stockholm Swedish? Answering this question would add to our general understanding of gesture-speech integration, and more specifically of the interaction of visual and verbal prominence cues.

The present study is based on audio and video data of 60 brief news readings from Swedish Television (SVT Rapport), comprising 1936 words in total, or about 12 minutes of speech from five news anchors (two female, three male). The material was transcribed, segmented at the word level, and annotated for big accents (henceforth, BA), head beats (HB) and eyebrow beats

(EB) using a combination of ELAN and Praat. In a first step of annotation, the presence vs. absence of BA, HB and EB was judged upon on a word-basis. About half of the materials (30 files) were annotated by three labelers independently of each other. Inter-rater reliability was tested using Fleiss' κ, and turned out fair to good (BA: κ = 0.77; HB: κ = 0.69; EB: κ = 0.72). For the purpose of this study, the analysis focuses on three conditions: (i) words produced with a BA only (i.e. without a beat gesture: 276 tokens in our material), (ii) words with BA co-occurring with a HB (BA+HB: 178 tokens) and (iii) words with BA co-occurring with both HB and EB (BA+HB+EB: 73 tokens). In a second step of annotation, tonal targets were labelled for all 527 tokens of interest: (H+)L* H- in case of Accent 1 and H*+L H- in case of Accent 2 (where H- is the sentence accent tone). In addition, as a baseline condition, tonal targets were labelled for a random selection of 102 non-focally accented words (small accents: 52 Accent 1, 50 Accent 2). Based on these tonal annotations, seven measures were derived to capture different aspects of the phonetic realization of the accentual fall (HL* or H*L respectively), and the sentence accent rise (H-): I/II – absolute peak height of HL*/H*L and H- (2 measures), III/IV – range of accentual fall and H- rise in semitones (2 measures), V – highest peak in word (= either HL*/H*L or H-), VI – largest range in word (= either accentual fall or H- rise), and VII – the difference between H- and accentual peak (HL*/H*L) in semitones.

The results reveal slight effects of the multimodal-prominence cluster (henceforth, MMP, i.e. conditions i-iii above plus the baseline), on measures I-VI, suggesting a tendency for a cumulative relation of verbal and visual prominence cues: i.e., the more visual cues accompanying, the higher the pitch peaks and the larger the rises and falls. These effects are strongest for the combined measures V and VI. All dependent variables were analyzed by means of linear mixed effects models assuming three fixed factors: *MMP* (cf. above), *speaker sex*, and *word accent* (Accent 1, Accent 2), as well as *speaker* as a random effect. According to a likelihood ratio test for each of the dependent variables, the effect of MMP was significant for measures V (*p*=.015*) and VI (*p*=.042*) suggesting an interaction of verbal and visual prominence cues that is reflected in both components of complex pitch accents in Stockholm Swedish – here referred to as the *tonal* fall and the *intonational* rise. The results are discussed both in the light of a proposed outline of a model of multimodal prominence production, and with reference to prosodic domains assumed in Swedish phonology.

References

[1] S. Myrberg and T. Riad, "The prosodic hierarchy of Swedish," *Nordic Journal of Linguistics*, vol. 23, no. 2, pp. 115-147, 2015.

[2] G. Bruce, "Swedish Word Accents in Sentence Perspective," Lund: Glerup, 1977.

[3] A. Wetterlin, E. Jönsson-Steiner, and A. Lahiri, "Tones and loans in the history of Scandinavian," T. Riad and C. Gussenhoven (eds.), *Tone and Tunes Volume 1*. Berlin; New York: Mouton de Gruyter, 2007.

[4] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication,* pp. 207-227 The Hague: Mouton, 1980.

[5] E. McClave, "Linguistic functions of head movements in the context of speech," *Journal of Pragmatics,* 32, pp. 855-878, 2000.

[6] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology. Journal of the Association for Laboratory Phonology* 3, pp. 71-889, 2012.

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