Accentual falls and rises vary as a function of accompanying head and eyebrow movements

Gilbert Ambrazaitis[1] and David House[2]
[1]Department of Swedish, Linnaeus University, Växjö, Sweden
[2]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

Abstract

In this study we examine prosodic prominence from a multimodal perspective. Our research question is whether the phonetic realization of accentual falls and rises in Swedish complex pitch accents varies as a function of accompanying head and eyebrow movements. The study is based on audio and video data from 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 results suggest a tendency for a cumulative relation of verbal and visual prominence cues: the more visual cues accompanying, the higher the pitch peaks and the larger the rises and falls.

Introduction

Previous research on co-speech gestures and audio-visual prosody strongly suggests that prosodic prominence is indeed an audio-visual, or multimodal, phenomenon: 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., Kendon, 1980, McClave, 2000).

It has, moreover, been shown that visual and verbal prominence cues may co-occur in various constellations (Swerts & Krahmer, 2010; Loehr, 2012; Ambrazaitis & House, 2017) and that beat gestures are more likely to occur with *perceptually strong* accents than with *weak* ones: Swerts and Krahmer (2010) 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 pitch accents in a corpus of Stockholm Swedish news readings. Our research question is whether the phonetic realization of accentual rises and falls in such accents varies as a function of accompanying beat gestures by the head and the eyebrows.

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 (Myrberg & Riad, 2015), as opposed to the *non-focal* or *small* accent (Myrberg & Riad, 2015). Crucially, the distinction between Accent 1 and Accent 2 is encoded at both levels. According to Bruce's (1977) seminal analysis, 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, cf. Bruce, 1977) 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. Bruce, 1977; Myrberg & Riad 2015; Wetterlin et al., 2007), 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 (Myrberg & Riad, 2015). In this brief report, we simplify the debate by referring to the tones defining the word accents as *accentual*, and the subsequent rise (the H- tone) as the *sentence accent rise*, (cf. Bruce, 1977).

The choice of complex (or big) pitch accents in Stockholm Swedish as an object of study thus introduces 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 fall at the accentual level, (b) the rise at the sentence level, or (c) both components of a complex pitch accent in 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.

Method

The present study is based on audio and video data from 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 (Sloetjes & Wittenburg, 2008) and Praat (Boersma, 2001). In a first step of annotation, the presence 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 (Fleiss, 1971), and turned out fair to good (BA: $\kappa = 0.77$; HB: $\kappa = 0.69$; EB: $\kappa = 0.72$). For the purpose of this study, the analysis focuses on three conditions:

- (i) words 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)
- (iii) words with BA co-occurring with 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). We treat conditions (i-iii) above with the baseline condition added as a four-level independent variable (or fixed factor) in our analysis (see Results section) and refer to this variable as *multimodal prominence cluster* (henceforth, **MMP**).

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** – absolute peak height of the accentual fall (HL*/H*L) in Hz
- **II** – absolute peak height of the sentence accent rise (H-) in Hz
- **III** – range of the accentual fall (HL*/H*L) in semitones
- **IV** – range of the sentence accent rise (H-) in semitones
- **V** – highest peak in word (= either HL*/H*L or H-)
- **VI** – largest range in word (= either accentual fall or sentence accent rise)
- **VII** – the difference between H- and the preceding HL*/H*L in semitones.

Results and discussion

The results reveal slight effects of the factor **MMP** (cf. above) on measures **I-VI**. Figures 1-3 display the results for measures **III**, **IV**, and **VI** as an example; the corresponding illustrations for measures **I, II** and **V** provide a similar picture. No effect of MMP is observed for measure **VII**.

These results suggest 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** (cf. Fig 3). All dependent variables were analyzed by means of linear mixed effects models assuming three fixed factors:

- *MMP* (cf. above)
- *speaker sex*
- *word accent* (Accent 1, Accent 2),

In addition, *speaker* was included as a random effect.

According to a likelihood ratio test for each of the seven dependent variables (i.e. measures **I-VII**), the effect of MMP was significant for measures **V** (p=.015*) and **VI** (p=.042*) suggesting a correlation of verbal and visual prominence cues that is reflected in both components of complex pitch accents in Stockholm Swedish – the accentual fall and the sentence accent rise.

The resuls from this study thus lend *acoustic* support for the *perception*-based prediction of a *cumulative relation* of verbal and visual prominence cues, a prediction we derived from Swerts and Krahmer's (2010) results based on auditory ratings for Dutch.

Figure 1. Boxplot for measure III – range of the accentual fall (HL/H*L) in semitones as a function of the multimodal prominence cluster (MMP); BSL = small accents; BA= big accents; HB=head heat; EB = eyebrow beat.*

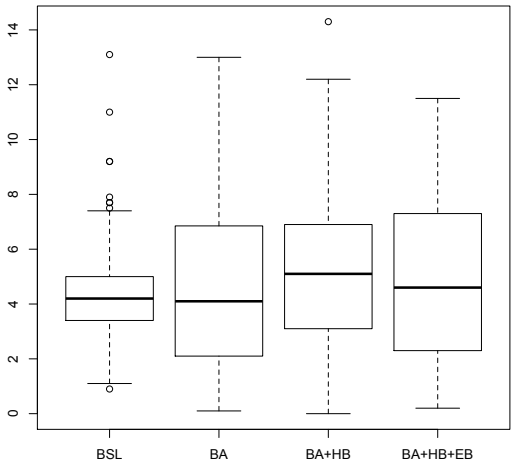


Figure 2. Boxplot for measure IV – range of the sentence accent rise (H-) in semitones as a function of MMP (cf. Fig. 1 for explanations).

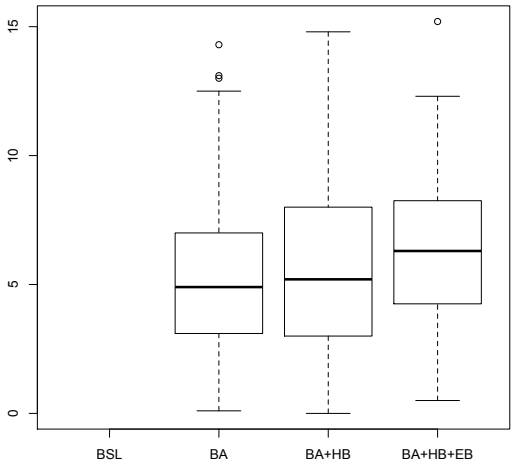
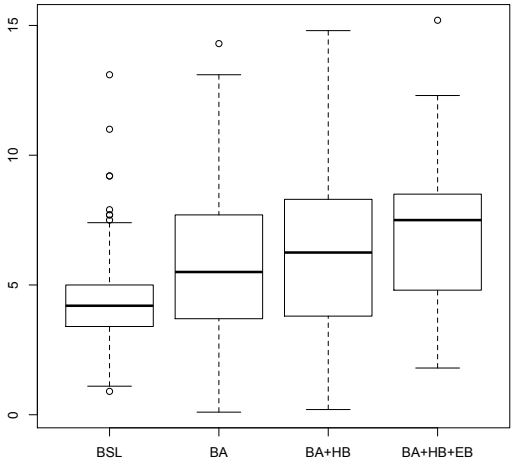


Figure 3. Boxplot for measure VI – largest range in word in semitones as a function of MMP (cf. Fig. 1 for explanations).



At the conference, the results are further 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.

Acknowledgements

We retrieved materials from the National Library of Sweden and received permissions from Swedish Television. We also thank our research assistants Malin Svensson Lundmark, Anneliese Kelterer, and Otto Ewald for assistance with data processing and annotations. This work was supported by the Marcus and Amalia Wallenberg Foundation [MAW 2012.01.03] and the Swedish Research Council [VR 2017-02140].

References

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.

Boersma P (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5: 341-345.

Bruce, G (1977). *Swedish Word Accents in Sentence Perspective*. Lund: Glerup.

Fleiss J (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76: 378-382.

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

Loehr D (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology. Journal of the Association for Laboratory Phonology,* 3: 71-889.

McClave E (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics,* 32: 855-878.

Myrberg S, Riad T (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, 23: 115-147.

Sloetjes H, Wittenburg P (2008). Annotation by category - ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

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.

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