Acoustic features of multimodal prominences

Do visual beat gestures affect verbal pitch accent realization?

Ambrazaitis, Gilbert; House, David

Published in:
Proceedings of The 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)

2017

Document Version:
Publisher's PDF, also known as Version of record

Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization?

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

[1]Centre for Languages and Literature, Lund University, Sweden
[2]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
gilbert.ambrazaitis@ling.lu.se, davidh@speech.kth.se

Abstract

The interplay of verbal and visual prominence cues has attracted recent attention, but previous findings are inconclusive as to whether and how the two modalities are integrated in the production and perception of prominence. In particular, we do not know whether the phonetic realization of pitch accents is influenced by co-speech beat gestures, and previous findings seem to generate different predictions.

In this study, we investigate acoustic properties of prominent words as a function of visual beat gestures in a corpus of read news from Swedish television. The corpus was annotated for head and eyebrow beats as well as sentence-level pitch accents. Four types of prominence cues occurred particularly frequently in the corpus: (1) pitch accent only, (2) pitch accent plus head, (3) pitch accent plus head plus eyebrows, and (4) head only. The results show that (4) differs from (1-3) in terms of a smaller pitch excursion and shorter syllable duration. They also reveal significantly larger pitch excursions in (2) than in (1), suggesting that the realization of a pitch accent is to some extent influenced by the presence of visual prominence cues. Results are discussed in terms of the interaction between beat gestures and prosody with a potential functional difference between head and eyebrow beats.

Index Terms: audio-visual prosody, multimodality, Swedish, news speech, co-speech gestures

1. Introduction

1.1. Research question

In spoken language, words are made prominent by means of prosody for various, e.g. information-structural or expressive, reasons [1]. 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 (cf. 1.1) [2][3][4][5][6][7]. It has, moreover, been shown that visual and verbal prominence cues may co-occur in various constellations (such as 'pitch accent plus head beat' or 'pitch accent plus head and eyebrow beat') [8][9], suggesting that we need to distinguish between verbal-only, visual-only, and possibly different multimodal prominence types or combinations, which might to some extent serve different communicative functions [10]. However, few systematic studies have been carried out to investigate the acoustic-phonetic realization of multimodal prominences. In this paper we ask the question: Do accompanying beat gestures in some way affect the realization of a pitch accent? Answering this question would add to our understanding of the interaction of verbal and visual prominence cues, and more generally, of gesture-speech integration. This paper presents an exploratory study on acoustic properties of multimodal prominences based on a corpus of read news from Swedish television.

1.2. Verbal prosodic prominence in Swedish

Unlike so-called intonation languages such as English, German or Dutch, Swedish is a pitch-accent language, making use of pitch contrasts at the lexical level. In particular, Swedish has a binary distinction between two word accents (Accent 1 and Accent 2), two different pitch accents assigned to words by means of lexical/ morphological rules. In addition, words can be highlighted at the sentence level, just as in English or German. Stockholm Swedish exhibits a well-established phonological distinction between the non-focal, accented realization of a word and a focal realization of a word. Note that the notion *"focal accent"* does not strictly relate to the information-structural notion of focus; it is rather synonymous with *sentence accent*. There exist various phonological interpretations of these basic patterns [11][12], but for the purpose of this study it is most relevant to note that, in both Accent 1 and 2, the focal accent is characterized by an additional rising pitch movement. It is important to understand that while the non-focal vs. focal accents represent two different phonological prominence levels, no difference in prominence is generally assumed between the two word accent categories (Accent 1 vs. 2) [12].

1.3. Audio-visual integration in prosodic prominence

It has been shown that beat gestures can facilitate both speech production [13] and speech processing [14][15]. A growing body of evidence suggests that hand, head and eyebrow movements are aligned with pitch accents in speech and in this way contribute to the production and perception of prosodic prominence [16][17][18][8][19][20].

There is also evidence suggesting that beat gestures are more likely to occur with *perceptually strong* accents than with *weak* ones: Swerts and Krahmer [8] found in their study of Dutch news readings that the more accented a word was on an auditory scale (no accent, weak accent, strong accent), the more likely the word was to also be accompanied by a head movement, an eyebrow movement or both (most common in the strongly-accented words).

Somewhat at odds with these findings by Swerts and Krahmer [8] are the results from perception experiments by Prieto et al. [21], which suggest that beat gestures (again, head and eyebrow movements) and pitch accents exhibit a kind of

trading relationship in the coding of contrastive focus (vs. information focus), in that one modality is able to compensate for another modality. A possible interpretation of this finding could be that, if a visual gesture is present, then verbal cues for prominence can be potentially weaker in their realization than when only verbal cues are present; something which is clearly not predicted by the results of Swerts and Krahmer [8]. Also, head beats were found to be more informative than eyebrow movements for the identification of contrastive focus [21], which might suggest different roles in prominence cuing for head beats and eyebrow beats (see also [22]).

1.4. Predictions

One prediction derived from our review in the previous section is that a focal pitch accent (in Swedish) is likely to be realized with stronger acoustic prominence cues (such as longer duration or a larger pitch excursion) when accompanied by a beat gesture than when not, since it has been shown that visual cues are more likely to occur with stronger prominences as assessed auditorily [8]. Under this prediction, we could characterize the relation between visual and verbal prominence cues as *cumulative*, where a strong verbal cue seems to attract an additional visual cue (or vice versa). In extension, we could predict (cf. [8]) a cumulative relation between the two visual cues discussed (head and eyebrow beats), implying that we could predict the strength of acoustic prominence to correlate with the number of accompanying visual cues (head/ eyebrows only vs. head and eyebrows combined). An alternative to this *cumulative cue* prediction, however, is a *cue trading* prediction (cf. our discussion of [21] in 1.3), according to which we would expect *weaker* acoustic prominence cues (e.g., shorter durations, smaller F0 excursions) for words accompanied by visual cues.

2. Method

In this study, we make a first attempt to pinpoint possible acoustic effects of accompanying head and eyebrow beat gestures on the realization of focally accented words, focusing on the two acoustic domains most commonly associated with prominence: segmental durations and fundamental frequency (F0). Our tentative approach is to semi-automatically extract two rather rough measures capturing the two dimensions: mean syllable durations (i.e., an estimation of speech rate), and F0 excursion (i.e. range) within a word. The rationale behind this approach was that these measures would be easily extracted based on existing word annotations, without the need for further manual annotations of the F0 contour. A caveat is, of course, that we do explicitly measure the F0 range of the focal-accent related F0 excursion, as the F0 minimum and maximum within a word might also in some cases relate to the preceding word-accent gesture (note that focally accented Accent 2-words are typically double peaked). However, the approach should still provide us with a tentative insight into acoustic effects of accompanying beat gestures.

2.1. Materials

This study is based on audio and video data of 47 brief news readings from Swedish Television (SVT Rapport), comprising 1516 words in total, or about 9 ½ minutes of speech. Each news reading typically contains 1-3 sentences (cf. Table 1). The recordings were retrieved on DVD from the National Library of Sweden (Kungliga Biblioteket).

The corpus includes speech from five news anchors (cf. Table 1). The selection of news anchors was random (only meeting the requirement of including both male and female speakers). The corpus comprises 30 of the 31 news stories used in [9] and [10] (1 was excluded due to a technical problem). That corpus contained speech from four speakers (sp. 1-4 in Table 1), where AN was heavily overrepresented (20 of 31 stories). For the present study, 17 additional stories were added in order to increase the number of stories for speakers 2-4 and to include a fifth speaker.

Table 1: *Materials: number of news stories included per speaker (i.e. news anchor); f/m=female/male.*

Speaker		**Materials**		
no.	ID	stories	words	minutes:seconds
1	AN (m)	19	608	3:48.5
2	SL (f)	6	165	1:05.9
3	PE (m)	8	265	1:52.4
4	KS (f)	6	206	1:23.2
5	FS (m)	8	272	1:28.9
	total	*47*	*1516*	*9:38.9*

2.2. Annotations

The material was transcribed, segmented at the word level, and annotated for focal accents (henceforth, FA), head beats (HB) and eyebrow beats (EB) using ELAN [23] [24]; word segmentations were adjusted using Praat [25] and re-imported in ELAN prior to doing the annotations.

The annotation scheme was simple in that only the presence vs. absence of the three prominence markers (FA, HB, EB) markers was judged upon. That is, no time-aligned annotations were done and hence, no decisions had to be made upon temporal onsets and offsets of the HB and EB movements. A word was annotated for bearing a (HB or EB) movement in the event that the head or at least one eyebrow rapidly changed its position, roughly within the temporal domain of the word.

A focal accent was annotated when a rising F0 movement corresponding to the focal H- tone in the Lund model of Swedish prosody [12] was recognizable in the F0 contour (cf. 2.2); note that this F0 movement was expected in the stressed syllable for Accent 1 words, while later in the word, surfacing as a second peak, in Accent 2 words. Praat [25] was used for inspecting the F0 contour. Focal accents were annotated with access to the audio channel, an F0 display, and the word segmentations, but without access to the video display. Note that our annotations of focal accents represent a phonological annotation, and not a perceptual assessment of different prominence levels as carried out in the Swerts and Krahmer study [8]. Phonological prominence does, however, have implicit perceptual relevance in Swedish.

Principles for HB and EB annotations differed slightly between the older part of the corpus (30 files, cf. 2.1) and the new part (additional 17 files), as follows: For the older part, annotations of HB and EB were done with full access to the audio- and video channels, as well as a display of the word segmentations. The rationale behind this decision was that annotations were made directly with reference to words, which was most feasible with both graphic and auditory reference to the words involved. For the additional news stories, HB were annotated as before, while EB were annotated in a second

step, by another annotator, without access to the audio channel and without prior listening. For EB this was judged feasible as our previous annotations had revealed rather few instances of EB in this kind of data [9].

The first 30 files were annotated (FA, HB, EB) by three annotators, independently of each other. Inter-rater reliability was tested using Fleiss' κ [26], and turned out fair to good (FA: κ = 0.77; HB: κ = 0.69; EB: κ = 0.72). Prior to analyses, our three-fold annotations were converted to a single, consensus (i.e. majority) rating for each word. The additional 17 stories were labelled by two additional annotators, where annotator 1 labelled HB and annotator 2 labelled first EB (without access to the audio, see above), and then FA (with audio access only).

2.3. Measurements and data analysis

Our previous studies (on a subset of the present data) have revealed that four (of seven possible) combinations of FA, HB, and EB seem to occur particularly frequently in our corpus: FA, FA+HB, FA+HB+EB, and HB. That is, FA and HB may occur without any other of the three cues, but this rarely happens for EB. Also, EB+HB tend to occur together with a FA, and FA+EB (without HB), is avoided, too. Table 2 below displays frequencies of occurrence of the four primary clusters for the present corpus.

Our analysis focuses on acoustic features of words annotated as either FA, FA+HB, FA+HB+EB, and HB. This enables us to test whether acoustic features of focally-accented words differ depending on accompanying beat gestures – either HB alone or HB+EB. We have in our annotations not distinguished between non-focally accented words and completely de-accented words. For that reason, we cannot include a baseline consisting of non-focally accented words without beat gesture. However, the HB-only category provides us with an auxiliary baseline of non-focally-accented (and, possibly, in some cases, completely de-accented) words (although accompanied by a HB).

Word durations and word-level F0 ranges (in semitones) were extracted automatically using the Praat script Prosody Pro [27], based on our manual word segmentations. In order to avoid unnecessary F0 analysis errors, F0 calculation was performed in the time-domain based on 'pulses' automatically determined by Praat, which we manually corrected using ProsodyPro [27]. The script also applies a smoothing algorithm removing minor spikes from F0 curves.

The extracted word durations were used to calculate mean syllable durations based on a count of canonical syllables for each word. Raw data (F0 range; mean syllable durations) were analyzed by means of fitting linear mixed models using the lmerTest package [28] in R [29], with prominence type as a four-level factor (FA, FA+HB, FA+HB+EB, HB), and speaker (i.e. news anchor) as a random factor.

3. Results

Our acoustic analysis comprises 501 data points per measure (mean syllable duration; F0 range), as it is based on 501 annotated words, i.e. all words annotated as either FA, FA+HB, FA+HB+EB, or HB in our corpus of 47 news readings. Table 2 displays the absolute frequencies of the four labels as well as their distribution across speakers (i.e., the five new anchors).

Table 2: *Distribution of four prominence categories (=combinations of labels FA, HB, EB) that occurred particularly frequently in our corpus of 47 read news stories (5 speakers, i.e. news anchors; f/m=female/ male). Absolute frequencies.*

Speaker		**Prominence label**				
no.	ID	FA	FA+HB	FA+HB+EB	HB	*total*
1	AN (m)	62	71	29	36	*198*
2	SL (f)	26	23	3	9	*61*
3	PE (m)	42	24	15	14	*95*
4	KS (f)	37	19	9	6	*71*
5	FS (m)	62	8	0	6	*76*
	total	*229*	*145*	*56*	*71*	*501*

Mean values of the two acoustic measures extracted for the labelled words are displayed in Table 3 for the four prominence categories, pooled across all five speakers. Each of the measures is further illustrated and discussed in the following subsections.

Table 3: *Mean values for mean syllable duration and F0 ranges for the four prominence categories (FA, FA+HB, FA+HB+EB, HB), pooled across all five speakers.*

Measure	**Prominence level**			
	FA	FA+HB	FA+HB+EB	HB
mean syll. dur. [ms]	252	250	246	214
F0 range [semitones]	9.25	10.64	10.29	7.22

3.1. Speech rate (mean syllable duration)

Our duration measurements reveal a somewhat lower speech rate (i.e. a longer mean syllable duration) for words spoken with a focal pitch accent (all three labels involving FA) than for words associated with a head beat only (label HB). This is clearly reflected both in the mean syllable duration (Table 3) as well as in the distribution of values as shown by the boxplot in Figure 1. Focal accents have been shown to be realized in multiple acoustic dimensions, among them the durational domain [30]. However, these results also suggest that the realization of a focal accent is independent of accompanying beat gestures, as there are not any longer mean syllable durations observed for the multimodal prominence clusters FA+HB and FA+HB+EB as compared to the verbal-only FA; on the contrary, the (non-significant) trend is rather the opposite, i.e.: shorter mean syllable durations for the multimodal clusters, most clearly seen in the median for FA+HB+EB (Fig. 1).

These observations are supported by a linear mixed model fit, showing that only HB differs significantly from FA (df=488.200; t=-2.663; p=.008**), while neither FA+HB (df=454.700; t=-.213; p=.831) nor FA+HB+EB (df=472.400; t=-.298, p=.766) differ significantly from FA.

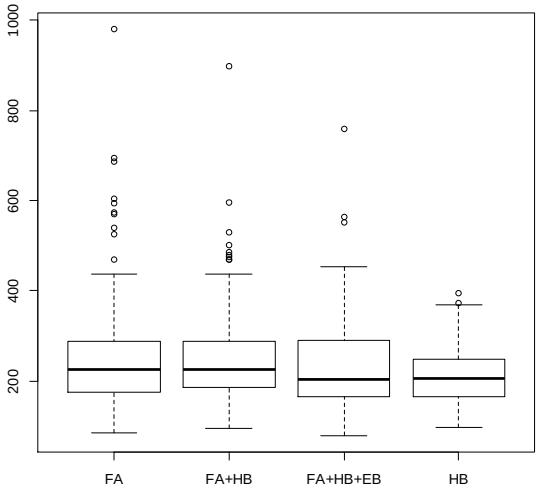

Figure 1: *Syllable durations (in milliseconds) for the four prominence categories (FA, FA+HB, FA+HB+EB, HB), pooled across all five speakers.*

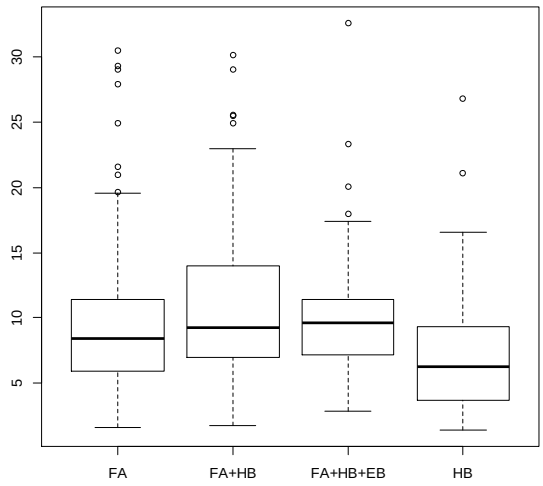

Figure 2: *F0 ranges (in semitones) for the four prominence categories (FA, FA+HB, FA+HB+EB, HB), pooled across all five speakers.*

3.2. F0 range

In contrast to the results for speech rate, our F0 range data do indeed suggest that a focal accent realization is somewhat affected by an accompanying beat gesture (in particular, a head beat), as both Table 3 above and the boxplot in Figure 2 suggest that a focally-accented word is spoken with a larger F0 range when accompanied by a head beat (FA+HB or FA+HB+EB) than when not (FA). Furthermore, and in line with the results for durations, focally accented words (FA, FA+HB, FA+HB+EB) are spoken with a larger F0 range than non-focally accented words, marked with a head beat only (HB).

Accordingly, a linear mixed model fit shows that HB differs significantly from FA (df=496.300; t=-3.155; p=.0017**). FA+HB also differs significantly from FA (df=497.000; t=2.346; p=.0194*), supporting the observed beat-gesture-effect just described. However, FA+HB+EB does not differ from FA (df=496.900; t=.711, p=.477). The linear mixed model fit thus suggests that FA+HB is realized with a slightly larger F0 range than the verbal-only FA, while FA+HB+EB is not, as if the addition of an eyebrow beat cancels out or complicates a possible beat-gesture-effect on the realization of a focal accent.

This seemingly contradictory effect of head and eyebrow beats may to some extent be explained in terms of speaker variation. Figure 3 reveals that our observation above (larger F0 excursion for FA+HB, but not for FA+HB+EB) is particularly valid for two speakers (3 and 4), while for speaker 1 (representing a large proportion of the data, cf. Table 1), we observe a tendency towards even larger F0 range in FA+HB+EB.

4. Discussion

The results have revealed a slightly complicated picture, neither clearly supporting the *cue trading* prediction, nor the *cumulative cues* prediction. One aspect of the complication lies in the fact that the results for the two acoustic measures studied yield contradictory results: duration data do not suggest any stronger acoustic prominence cues for words accompanied by beat gestures (rather the opposite, but this trend is not significant), while F0 data do support such a relation. However, another aspect lies in the fact that the results for F0 range seem to suggest different effects for the two visual cues studied (head vs. eyebrow movements), but also speaker-specific behavior: pooled across speakers, only head beats but not eyebrow beats seem to reflect stronger acoustic cues.

The diverging results for duration and F0 data might to some degree be explained in terms of F0 being the stronger (or primary) acoustic prominence cue, with the effects of gesture-speech interplay being too small to affect the (secondary) durational domain.

The picture that emerges, based on the F0 data, is that there can be a difference between head beats and eyebrow beats in their relationship to prominence. Head beats may be a stronger or more favored signal of prominence and therefore could comprise a more unified and simple signal when occurring alone with a focal accent. Here we see a cumulative cue behavior. The addition of eyebrow movements in the cue complex could serve other functions, this also being speaker dependent (cf. [21][22]). This may not be evidence of a trading relationship, but rather of a more complex system of cueing other communicative functions.

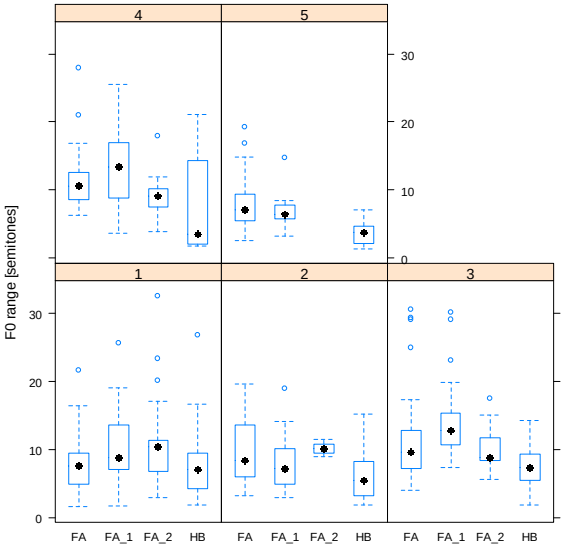

Figure 3: *F0 ranges (in semitones) for the four prominence categories FA, FA+HB, FA+HB+EB, HB (where FA_1 = FA+HB and FA_2 = FA+HB+EB), separately for five speakers (cf. Tab. 1 for speaker IDs).*

A differential behavior of eyebrow vs. head beats is, however, generally in line with the conclusions by Prieto et al. [21] (cf. 1.3), and also with observations concerning the distribution and functionality of head vs. eyebrow beats in Swedish news broadcasts [9] [10]: Eyebrow beats were found much less frequently than head beats, occurring almost always in connection with a head beat, and primarily in connection with semantically loaded words (e.g. denoting a contrast, an emotion, or great value).

Although our results advocate a variant of the cumulative cue prediction, which was derived from [8], our conclusions concerning the differential nature of head and eyebrow beats are less in line with Swerts and Krahmer [8], although their study was also based on news speech (Dutch). Their results rather suggest equivalent functions of the two visual modalities, as each of them alone can mirror a minor degree of prominence, while their combination adds up to a higher degree of prominence. This discrepancy calls for more research taking into account language, culture and genre as potential factors governing the interplay of visual gestures in prominence cuing.

Another task for future studies is to integrate both acoustic measures as in the present study and perceptual prominence ratings as in [8]. Furthermore, additional acoustic dimensions could be included, such as the energy domain [30], as well as more refined and varied measures of F0.

5. Conclusion

The results of this study point in the direction of possible functional differences between eyebrow beats and head beats and their interplay with acoustic prominence. Understanding these differences will be an exciting avenue for continued research.

6. Acknowledgements

We retrieved our materials (television broadcasts) on DVD from the National Library of Sweden (Kungliga biblioteket, KB). We cordially thank Ann-Charlotte Gyllner-Noonan at KB for her kind and patient assistance and Swedish Television (SVT) for relevant permissions. We also thank our research assistants Malin Svensson Lundmark, Otto Ewald, and Anneliese Kelterer for assistance with data processing and annotations, and Joost van de Weijer (Lund University) for statistical advice. This work was supported by the Marcus and Amalia Wallenberg Foundation [grant number MAW 2012.01.03] and the Bank of Sweden Tercentenary Foundation [grant number P12-0634:1].

7. References

[1] D.R. Ladd, *Intonational Phonology* (2nd ed.), Cambridge: Cambridge University Press, 2008.

[2] 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.

[3] D. McNeill, *Hand and mind: What gestures reveal about thought*, Chicago: The University of Chicago Press. 1992.

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

[5] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes," In *Proceedings of Interspeech 2006*, Pittsburg, PA. USA, pp. 1272–1275, 2006.

[6] S. Alexanderson, D. House, and J. Beskow, "Aspects of co-occurring syllables and head nods in spontaneous dialogue," In *Proc. of 12th International Conference on Auditory-Visual Speech Processing* (AVSP2013). Annecy, France, 2013.

[7] P. Wagner, Z. Malisz, and S. Kopp. "Gesture and speech in interaction: An overview," *Speech Communication* 57, pp. 209-232, 2014.

[8] 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.

[9] G. Ambrazaitis, M. Svensson Lundmark, and D. House, "Multimodal levels of promincence: a preliminary analysis of head and eyebrow movements in Swedish news broadcasts," In M. Lundmark Svensson, G. Ambrazaitis, and J. van de Weijer (Eds.), *Proceedings of Fonetik 2015*, pp. 11-16, Lund University, Sweden, 2015.

[10] G. Ambrazaitis and D. House, "Multimodal levels of prominence - The use of eyebrows and head beats to convey information structure in Swedish news reading," In *Seventh Conference of the International Society for Gesture Studies* p. 310, Paris, 2016

[11] T. Riad, "Scandinavian accent typology," In Å. Viberg (Ed.), *Special issue on Swedish. Sprachtypologie und Universalienforschung (STUF)*, 59, pp. 36-55, 2006.

[12] G. Bruce, "Components of a prosodic typology of Swedish intonation," In T. Riad and C. Gussenhoven (Eds.), *Tones and Tunes – Volume 1: Typological Studies in Word and Sentence Prosody*, Berlin; New York: Mouton de Gruyter, pp. 113-146, 2007.

[13] C. Lucero, H. Zaharchuk, and D. Casasanto, "Beat gestures facilitate speech production," *Proc. of the 36th Annual Conference of the Cognitive Science Society*, Austin, TX, pp. 898-903, 2014.

[14] E. Biau and S. Soto-Faraco, "Beat gestures modulate auditory integration in speech perception," *Brain & Language*, 124, pp. 143-152, 2013.

[15] L. Wang and M. Chu, "The role of beat gesture and pitch accent in semantic processing: An ERP study," *Neuropsychologia*, 51, pp. 2847-2855. 2013.
[16] Y. Yasinnik, M. Renwick and S. Shattuck-Hufnagel, "The timing of speech-accompanied gestures with respect to prosody," In *Proc. of From Sound to Sense*, MIT, Cambridge, MA, pp. 97-102. 2004.
[17] D. McNeill, *Gesture and thought*, University of Chicago Press, Chicago. 2005.
[18] M. L. Flecha-Garcia, "Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English," In *Proceedings of CogSci-2007*, Austin, Texas, USA, p. 1753, 2007.
[19] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes* 26, pp. 1457−1471, 2011.
[20] 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.
[21] P. Prieto, C. Puglesi, J. Borràs-Comes, E. Arroyo, and J. Blat, "Exploring the contribution of prosody and gesture to the perception of focus using an animated agent," *Journal of Phonetics* 49(1), pp. 41-54. 2015.
[22] D. House, J. Beskow, and B. Granström, "Timing and interaction of visual cues for prominence in audiovisual speech perception," In *Proceedings of Eurospeech 2001*, Denmark: Aalborg, pp. 387–390, 2001.
[23] H. Sloetjes and P. Wittenburg, "Annotation by category – ELAN and ISO DCR," In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008.
[24] ELAN. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, http://tla.mpi.nl/tools/tla-tools/elan/
[25] P. Boersma and D Weenink, "Praat: doing phonetics by computer [Computer program]," version 5.4.01, http://www.praat.org/, 2014.
[26] J. Fleiss, "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, 76(5), pp. 378-382, 1971.
[27] Y. Xu, "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France*, pp. 7–10, 2013.
[28] lmerTest. Tests in Linear Mixed Effects Models. A. Kuznetsova, P. Bruun Brockhoff, and R. H. Bojesen Christensen, https://CRAN.R-project.org/package=lmerTest
[29] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2015.
[30] M. Heldner, *Focal accent – F0 movements and beyond*, Ph. D. thesis, Umeå University, 2001.