10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Word prominence ratings in Swedish television news readings: effects of pitch accents and head movements

Gilbert Ambrazaitis[1], *Johan Frid*[2], *& David House*[3]

[1]Linnaeus University, Växjö, Sweden, [2]Lund University Humanities Lab, Sweden,
[3]KTH (Royal Institute of Technology), Stockholm, Sweden

gilbert.ambrazaitis@lnu.se, johan.frid@humlab.lu.se, davidh@speech.kth.se

Abstract

Prosodic prominence is a multimodal phenomenon where pitch accents are frequently aligned with visible movements by the hands, head, or eyebrows. However, little is known about how such movements function as visible prominence cues in multimodal speech perception with most previous studies being restricted to experimental settings. In this study, we are piloting the acquisition of multimodal prominence ratings for a corpus of natural speech (Swedish television news readings).

Sixteen short video clips (218 words) of news readings were extracted from a larger corpus and rated by 44 native Swedish adult volunteers using a web-based set-up. The task was to rate each word in a clip as either non-prominent, moderately prominent or strongly prominent based on audio-visual cues. The corpus was previously annotated for pitch accents and head movements.

We found that words realized with a pitch accent and head movement tended to receive higher prominence ratings than words with a pitch accent only. However, we also examined ratings for a number of carefully selected individual words, and these case studies suggest that ratings are affected by complex relations between the presence of a head movement and its type of alignment, the word's F0 profile, and semantic and pragmatic factors.

Index Terms: multimodal speech perception, multimodal prominence, audiovisual prosody

1. Introduction

The act of speaking engages the entire body. Posture changes, movements of the head, eyebrows and facial features, and gestures of the hands and arms typically co-occur with speech. Studying the synchronization between speech and gesture has played an important role in building theories of human communication which approach speech and gesture production as arising from a common generation process [1] and [2]. The specific relationship between speech prosody and gesture is an area which has attracted considerable attention, particularly as prosody and gesture can have similar functions [3]. One of the more important functions shared by prosody and gesture involves the signaling of prominence.

Terken and Hermes [4] define prosodic prominence as a linguistic entity which "stands out from its environment by virtue of its prosodic characteristics" with the primary acoustic prosodic properties being amplitude, duration and F0. The acoustic signaling of prosodic prominence often coincides or co-occurs with a beat gesture generally defined as a rapid movement of a hand, finger, the head or the eyebrows. It is well established that beat gestures contribute to prosodic prominence [5], [6], [7], [8], [9], [10], and [11] but we still do not have a clear understanding of how the various acoustic and visual cues contribute to the perception of prominence or of their relative importance.

In earlier work on the relationship between prosodic prominence and gesture, we have concentrated on analyzing the co-occurrence of pitch accents with head nods and eyebrow movement. We have chosen an ecologically valid speech genre (news reading) which is nonetheless restricted in a way that encourages head and eyebrow movement as gestural signals for prominence [12]. Previous studies on the perception of prominence have been largely aimed at investigating audio speech cues [13], [14], [15], and [16]. Studies including audiovisual cues have generally been restricted to specially designed experimental settings typically using stimuli where the audio and video are presented separately or are non-congruent [17], [18], [19], and [20], or where carefully controlled synthetic stimuli are used [21], [22], and [23].

The present study builds on our previous work on investigating the interrelationships among multiple audiovisual dimensions (head, eyebrows, pitch accents) for signaling prominence in speech production [12]. Here we address the need for perceptual ratings to complement and verify our findings concerning production. Using a subset from the previously analyzed material (news reading obtained from Swedish Television), we have conducted a pilot study with a two-fold purpose: First, we aim to validate the methodological set-up by means of testing if words co-occurring with a pitch accent are actually perceived as much more prominent than words having no pitch accent – which would be expected if the rating task works successfully.

Second, we aim to obtain a preliminary answer as to whether words co-occurring with head nods and pitch accents combined are perceived as having higher levels of prominence than words co-occurring with only a pitch accent. To this end, we present the results of two analyses: an overall quantitative approach involving the entire dataset of 218 words by five speakers, combined with two case studies of 11 selected words by two speakers.

2. Method

A selection of 16 short video clips from Swedish television news broadcasts was rated by 44 native Swedish adult volunteers with no reported hearing impairment and normal or corrected sight, using a web-based set-up. Each word was to be rated as either non-prominence (no action), moderately prominent, or strongly prominent, by means of clicking the word in question until the desired prominence level was encoded though a specific color (see 2.2.2).

10.21437/SpeechProsody.2020-64

2.1. The audiovisual speech sample

The clips were between 4 and 7 seconds long and contained 13 words on average (218 words in total), ranging from 8 to 19 words. They comprise speech of five different speakers (news anchors) and were taken from a larger corpus [12] that was annotated for head and eyebrow movements (binary absence/presence decision per word), as well as for so-called 'big' pitch accents in Swedish [24] also known as the 'sentence accent' or the 'focal accent'. The big accent (henceforth, BA) consists of a high tonal target (H) added to the preceding lexical pitch accent (a HL, either aligned early in Accent I: HL* or late in Accent II: H*L), that is a two-peaked falling-rising (HLH) pitch accent. In Accent I, the final LH-rise is typically realized largely within the stressed syllable, while in Accent II, it is realized in the post-stress, or – in compounds, which generally receive Accent II – in the secondary stress syllable. F0 measurements of the HL-fall and the LH-rise for each word are available [25].

2.2. Data collection

Volunteers were recruited via social media and e-mail. They were offered a cinema ticket for their participation. They were encouraged to conduct the rating in a silent surrounding. A session, including instructions and questionnaire, took approximately 17 minutes on average, ranging from 8 to 45 minutes. The actual test took 10 minutes on average, ranging from 5 to 22 minutes.

2.1.1 The set-up /rating procedure.

Data collection was performed using a custom-made web page implemented in javascript, jQuery and the jQuery Simple Presentation plugin. We used the HTML5 software solution stack, particularly making use of the `<video>` tag, which facilitates web-based video playback considerably. The web page guided the participant through an instruction phase and a training phase. Then, the data collection proper consisted of 16 rating tasks (16 clips), described in detail in 2.2.2. The order of clips to be rated was randomized for each participant. When the test was finished, all the data was sent to a sheet in Google docs.

2.2.2 The rating task

Each clip was rated using a GUI including a video-player, an orthographic representation of the text of the clip, as well as a *Nästa* 'Next' button (see Fig. 1). The text was presented word-by-word in equally-sized boxes. The boxes were to be used as buttons for the prominence rating: A click with the mouse (or the touch screen) changed the colour of the box, which would turn YELLOW (prominence level 1) after one click, RED (prominence level 2) after another click, and neutral again after a third click.

A clip presentation always started with a still video and a 'Start' button. When that button was clicked, the clip was played automatically two times, without any break in between and without the option to pause the video. During this initial presentation, the rating buttons (incl. the orthographic representations) were hidden. Participants were instructed to carefully look at the video during this double screening. This was done in order ensure that the participants' first impression of the clip and its prominence relations would be based on the full audio-visual input. After this initial phase, the text buttons along with usual video playing controls appeared. The participant was then free to play the video again as often as necessary, making use of pausing or playing smaller parts if desired, and to rate all words using the text buttons. When satisfied, the participant clicked the 'Next' button to reach the next clip.

Figure 1: *Screenshot of the GUI.*

2.2.3 Participants

Forty-four adult native Swedish volunteers participated in the study (27 female, 16 male, one without specifying sex or gender; 42 years old on average, ranging from 23 to 73 years).

2.3. Analysis

Two separate analyses were performed in order to explore possible effects of pitch accents and head movements on word prominence ratings. The first is a systematic, quantitative attempt to determine overall effects of the mere presence of a pitch accent or a head movement on a given word, based on our existing annotations (see 2.2). To this end, the 218 words included in the rating task were classified as either being realized with a BA (Big Accent) and a HB (Head Beat), a BA only, or neither. Table 1 displays token frequencies for these three categories in the selected data set.

Table 1: *Frequencies of occurrence of words with 'big accent' (BA), with BA and a 'head beat' (BAHB), and without BA (noBA).*

noBA	BA	BAHB	total
148	22	48	218

Mean prominence ratings were calculated for each rater and each of the three word categories (noBA, BA, BAHB) by dividing the sum of raw prominence ratings for all tokens in a category by *n* according to Table 1.

A comparison of rater means for these categories should provide us with an overall impression of rater behavior with respect to the presence or absence of pitch accents and head movements. However, a drawback with this overall approach is that it does not consider factors such as the realization of a BA in terms of pitch range, duration, or other typical prominence-related acoustic features, or the realization and alignment of a head movement, not to speak of top-down effects due to semantic or pragmatic factors.

In order to shed some preliminary light on these issues, our second analysis consists of two case studies of 11 carefully selected words from two of the speakers (one female: Katarina, one male: Pelle). To this end, for each speaker, the available words labelled BA and BAHB were inspected focusing on phonological-prosodic and F0 characteristics (using available

measurements from [25]) in order to identify a selection as controlled as possible. For Katarina, we were able to identify a set of 5 words sharing the feature of being di-syllabic, initially-stressed Accent II-words, with crucial variations regarding the presence of head movements and the realization of the big accent (Tab. 2). For Pelle, the available set consisted of 6 words with a similar variation, where all words were controlled as being compounds, which in Swedish implies Accent II and two lexical stresses (Tab. 3).

For these case studies, prominence ratings of individual words were normalized by subtracting the rater-specific mean rating of all words from the rating of the word in question.

3. Analysis 1: Overall effects of pitch accents and head movements

3.1. Results

Figure 2 displays the results from the first analysis (see 2.3). It shows that words uttered without a 'big accent' (BA) were rated low overall, while words with a BA tended to receive considerably higher prominent ratings, where words with HB tended to receive higher ratings than words with BA only. This effect of the presence of BA or HB is highly significant according to a repeated-measures ANOVA ($F[2;86]=285.80$; $p<.001$ after Greenhouse-Geisser correction; Sphericity violated according to Mauchly's test).

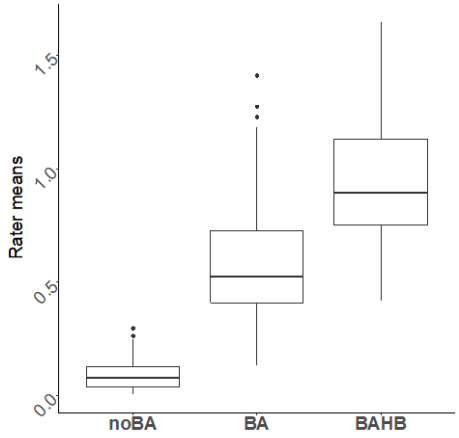

Figure 2: *Boxplots of prominence ratings (rater means) for words realized without any 'big accent' (noBA), with a 'big accent' only (BA), or a 'big accent' and a 'head beat' (BAHB).*

3.2. Discussion

The results of analysis 1 are clearly in line with the expectations formulated in the introduction, which provides us with two insights, or preliminary conclusions.

First, it would seem that our rating task exhibits a certain level of validity, as it renders clearly higher average prominence ratings for words uttered with a pitch accent (the Swedish 'big accent': often rated as moderately, 1, or strongly, 2, prominent) than for words lacking a 'big accent' (most often rated 0), which is the most plausible outcome.

Second, we might at least tentatively conclude that head movements, as perceived visually, have added to the overall audio-visual percept of prominence. On the one hand, we know that in the present dataset, words realized with a head movement also tend to be realized with larger pitch excursions than words with a big accent only, i.e. without head movement [25]. It might hence be that the difference in prominence ratings obtained for BA words and BAHB words (Fig. 2) is actually explained by acoustic differences rather than the additional visible cue of head movement. On the other hand, the acoustic differences reported for this dataset are moderate.

We cannot resolve this issue in the present paper. However, in the next section we present an attempt to further add to our tentative conclusions by comparing and discussing the ratings obtained for a number of selected words.

4. Analysis 2: case studies

4.1. Results

In Figures 3-4 and Tables 2-3, the words are sorted first with respect to the occurrence of a head movement, and second with respect to the largest F0 range (either fall or rise) measured. For instance, in Figure 3 and Table 2, words 1-3, are realized without a head movement, and for word 1, the largest F0 range measured is 6.83 st which is smaller than 11.38 st for word 2.

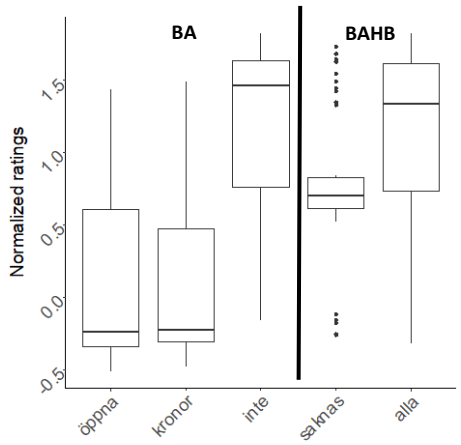

Figure 3: *Boxplots of prominence ratings (rater normalized, see 2.3) for five selected words by female speaker Katarina. See Tab. 2 for word semantics and F0 characteristics.*

Table 2: *Characteristics of five selected words by female speaker Katarina; all words are disyllabic Accent II-words (with initial stress); F0 fall and rise in semitones* (st) *refer to the fall and the rise of the two-peaked 'big accent'.*

Word	Meaning	HB?	F0 fall *(st)*	F0 rise *(st)*
öppna	'to open'	no	0.57	6.83
kronor	Sw. currency	no	11.38	5.78
inte	'not'	no	9.90	14.29
saknas	'is missing'	yes	8.55	6.64
alla	'all'	yes	7.01	10.91

Figure 3 shows that words realized without head movement and small or moderate F0 movements (words 1-2) tend to receive the lowest prominence ratings. Words 4 and 5, which were realized with head movements and with F0 ranges in the same order of magnitude as words 1 and 2, tended to receive considerably higher ratings. For word 3 (no head movement),

the largest F0 ranges were measured and similar results were obtained as for word 5 (head movement, but smaller F0 ranges).

A similar tendency for higher prominence ratings of words with head movements is observed in Figure 4 for the male speaker, but the situation is more complex. Words 1 and 2 are similar with respect to maximum F0 range, but word 2 is perceived as considerably more prominent. Turning to words 3-6 (with head movements), we can observe similar results for words 3-6 although word 3 is realized with a considerably lower F0 range. Furthermore, word 6 tends to be rated slightly lower (when comparing the medians), although it has clearly the largest F0 ranges.

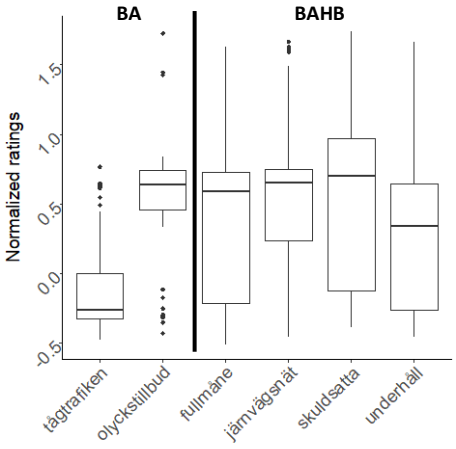

Figure 4: *Boxplots of prominence ratings (rater normalized, see 2.3) for six selected words by male speaker Pelle. See Tab. 2 for word semantics and F0 characteristics.*

Table 3: *Characteristics of six selected words by male speaker Pelle; all words are compound Accent II-words (with initial stress); F0 fall and rise in semitones* (st) *refer to the fall and the rise of the two-peaked 'big accent'.*

Word	Meaning	HB?	F0 fall *(st)*	F0 rise *(st)*
tågtrafiken	'railway traffic'	no	10.89	7.92
olyckstillbud	'incident'	no	9.21	11.50
fullmåne	'full moon'	yes	3.37	6.63
järnvägsnät	'rail network'	yes	8.93	12.12
skuldsatta	'indebted	yes	12.23	9.63
underhåll	'maintenance'	yes	14.34	14.58

4.2. Discussion

The results of Analysis 2 are generally in line with those of Analysis 1, suggesting a tendency for higher prominence ratings of words accompanied by a head movement, but less straightforwardly so at first sight.

Word 3 (no head movement) by speaker Katarina is rated as high as words 4-5, but this is likely related to its larger F0 excursions, suggesting that although head movements might tend to correlate with larger F0 movements [25], this is only a tendency and F0 can compensate for a lack of movement. In addition, top-down processes in prominence perception have been observed in previous studies [15] and might also come into play here, as word 3 (*inte*) is semantically/pragmatically essential. Top-down effects, partly pragmatically driven, might also explain the different rating distributions for words 1 and 2 in Figure 4. The two words are taken from the same sentence (*…olyckstillbud i tågtrafiken…* 'incidents in railway traffic'), where word 1 (*tågtrafiken*) is semantically less loaded. Furthermore, *tågtrafiken* is uttered in between two prominent words (*olyckstillbud* and *ökat* 'have increased', realized with BA, HB and eyebrow raises), which might have a degrading effect on its prominence rating.

In order to better understand the ratings obtained for words 3-6 in Figure 4 and how they might relate to the head movement, we re-examined the corresponding video clips by means of visual inspection. Word 3 was clearly realized with a 'double head beat', a phenomenon occasionally observed in these data [26], where each beat was nicely aligned with a stressed syllable (primary and secondary stress on *full* and *må*). We propose that this doubling and the stress-alignment might enhance prominence perception, explaining a tendency for high ratings of this word, despite moderate F0 excursions. Words 4-5 were realized with simple, but likewise nicely aligned head movements. However, for word 6, we found a head movement that did not seem to align to the stressed syllable; it seems to signal confirmation ('nodding') rather than prominence, which might explain the word's tendency for relatively low ratings despite large F0 ranges.

To sum up, results of Analysis 2 provide a more complex picture than those of Analysis 1 as they also take into account the words' F0 patterns (and to some degree, informally, the head movements' alignment patterns) but still show strong evidence for a contribution of visually perceived head movements to prominence perception.

5. Conclusions and outlook

Words realized with a pitch accent and head movement tended to receive higher prominence ratings than words with a pitch accent only. However, our findings in the case studies suggest that prominence ratings can be affected by complex relations between the presence of a head movement and its type of alignment, the word's F0 profile, and semantic and pragmatic factors.

The next step from this pilot study will be a follow-up study testing the same dataset in an audio-only condition (i.e. lacking the video display), with otherwise identical set-up, with a new group of participants. Given that a contribution of the visual modality will be confirmed by the audio-only follow-up, our future agenda can be sketched as follows: We are currently studying correlations between acoustic parameters (F0 and durations, [25]) and visual movements (head and eyebrow movements) in a larger data set from which the current set was taken. Our goal is to add perceptual prominence ratings, obtained audiovisually as in the present pilot, to the entire data set, which we plan to achieve using a crowdsourcing approach (see [27] for details). These ratings, in combination with detailed acoustic measurements and the full gestural annotations (head and eyebrow movements) should enable us to disentangle the individual contributions of various acoustic and visual parameters to prominence in an ecologically valid, if special, data set.

6. Acknowledgements

This work was supported by two grants from the Swedish Research Council (VR-2017-02140 and VR-2013-2003).

7. References

[1] A. Kendon, *Gesture: Visible action as utterance,* Cambridge: Cambridge University Press, 2004.

[2] D. McNeill, *Gesture and Thought*, Chicago: University of Chicago Press, 2005.

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

[4] J. Terken, and D. Hermes, "The Perception of Prosodic Prominence," in M. Horne (ed.), *Prosody: Theory and Experiment. Text, Speech and Language Technology,* vol 14, Dordrecht: Springer, 2000.

[5] Y. Yasinnik, M. Renwick, and S. Shattuck-Hufnagel, "The timing of speech-accompanied gestures with respect to prosody," *Proceedings of From Sound to Sense*, MIT, Cambridge, MA, pp. 97-102, 2004.

[6] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English," *Speech Communication,* 52 (2010), pp. 542-554, 2010.

[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.

[8] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Lang. Cogn. Processes,* 26, pp. 1457-1471, 2011.

[9] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Lab. Phonol.: J. Assoc. Lab. Phonol.,* 3, pp. 71-89, 2012.

[10] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements," *J. Speech Lang. Hear. Res.,* 56 (3), pp. 850-864, 2013.

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

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

[13] G. Fant, A. Kruckenberg, J. Liljencrants, and S. Hertegård, "Acoustic-phonetic studies of prominence in Swedish," *TMH-QPSR,* 41(2-3), pp. 1-52. 2000.

[14] A. Eriksson, E. Grabe, and H. Traunmuller, "Perception of syllable prominence by listeners with and without competence in the tested language," *Proceedings Speech Prosody 2002*, Aix-en-Provence, pp. 275–278, 2002.

[15] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, 110, pp. 425-452, 2010.

[16] D. Arnold, P. Wagner, and B. Möbius, "Obtaining prominence judgments from naïve listeners -- Influence of rating scales, linguistic levels and normalization," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association,* Portland, Oregon, Proceedings, pp. 2394-2397, 2012.

[17] E. Krahmer, and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language,* 57 (3), pp. 396–414, 2007.

[18] M. Swerts, and E. Krahmer, "Facial expressions and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics,* 36 (2), pp. 219-238, 2008.

[19] R. Scarborough, P. Keating, S. L. Mattys, T. Cho, A. Alwan, and E. T. Auer, "Optical phonetics and visual perception of lexical and phrasal stress in English," *Language and Speech*, 52, pp. 135–175, 2009.

[20] M. Dohen and H. Loevenbruck, "Interaction of audition and vision for the perception of prosodic contrastive focus," *Lang and Speech,* 52 (2009), pp. 177-206, 2009.

[21] D. House, J. Beskow, and B. Granström, "Timing and interaction of visual cues for prominence in audiovisual speech perception," In *Proc of Eurospeech 2001*, Aalborg, Denmark, pp. 387-390, 2001.

[22] S. Al Moubayed, J. Beskow, B. Granström, and D. House, "Audio-Visual Prosody: Perception, Detection, and Synthesis of Prominence," in A. Esposito, A. M. Esposito, R. Martone, V. C. Müller, and G. Scarpetta (eds), *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, Lecture Notes in Computer Science, vol 6456, Springer, Berlin, Heidelberg, 2011.

[23] 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.

[24] S. Myrberg and T. Riad, "The prosodic hierarchy of Swedish," *Nord. J. Linguist*., 38 (2), pp. 115-147, 2015.

[25] G. Ambrazaitis, and D. House, "Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization?" *Proceedings of the 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*, Stockholm, Sweden, 2017.

[26] A. Kelterer, G. Ambrazaitis, and D. House, "Head beats as pitch-accompanying visual correlates of primary and secondary lexical stress: evidence from Stockholm Swedish compounds," *Proc. TAL2018, Sixth International Symposium on Tonal Aspects of Languages,* Berlin, Germany, pp.124-128, 2018.

[27] G. Ambrazaitis, J. Frid, and D. House, "Multimodal prominence ratings: effects of screen size and audio device," in *6th European and 9th Nordic Symposium on Multimodal Communication*, University of Leuven, Belgium, pp 2-3, 2019.