This is the accepted version of a paper presented at 1st International Multimodal Communication Symposium (MMSYM 2023), Barcelona, Spain, April 26-28, 2023.

Citation for the original published paper:

Web-based, audio-visual prominence ratings of Swedish news reading materials: Effects of head movements, rating condition, and hardware
In:

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-125669
**Web-based, audio-visual prominence ratings of Swedish news reading materials:**

**Effects of head movements, rating condition, and hardware**

Gilbert Ambrazaitis, *Linnaeus University, Växjö, Sweden*

Johan Frid, *Lund University Humanities Lab, Sweden*

David House, *KTH (Royal Institute of Technology), Stockholm, Sweden*

Although prominence is increasingly recognized as an essentially multimodal phenomenon, relatively little is known about how spontaneous gestures in naturally occurring speech relate to perceived prominence (Jiménez-Bravo & Marrero-Aguiar, 2020). To study these relations, we need to collect prominence ratings for large amounts of ecologically valid audio-visual speech data efficiently, and to this end, we are currently developing a web-based rating set-up (Ambrazaitis et al., 2019; 2020; 2022). In our prototype rating task, 16 clips from Swedish television news (218 words in total) are to be rated by volunteers using a web interface. In our GUI, the text is displayed below the video player. Each word is to be rated as either non-prominent, moderately prominent (yellow), or strongly prominent (red), by clicking on the word button until the desired prominence level is encoded through a specific color. Participants have so far been free to use a mobile phone or a computer, and headphones or loudspeakers.

In an initial analysis, we showed that overall rating behavior is affected by hardware choices and screen size (Ambrazaitis et al., 2019). We later also collected data using the same materials in an audio-only condition (Ambrazaitis et al., 2022), and, instead of only measuring overall rating behavior, we compared ratings for words uttered with or without sentence-level pitch accents and head movements of any type (Ambrazaitis et al., 2020; 2022). Figure 1a displays results based on 85 raters, showing, first, a clear effect of the occurrence of accents and head movements, which is highly significant (model comparison: $\chi^2=322.37$, $df=2$, $p<.001$). The higher ratings for ‘accent plus head’ compared to ‘accent’ (in both conditions) are explainable, as accents with head movement are often realized stronger acoustically than accents without (Ambrazaitis & House, 2022). The plot also suggests a slight (but not significant) effect of the rating condition (and an interaction with the occurrence of accents and head movements).

In a next step, which will be presented in detail at the symposium, we tested a revised set-up where only a section of the video was shown, displaying only the speaker (i.e., the newsreader), but not the studio background (usually presenting various illustrations), which, we hypothesize, might attract some attention and account for the relatively weak effects of the presence of head movements seen in Figure 1a. So far, we have collected ratings from 29 subjects using this new rating condition. Our preliminary analysis does not reveal any
significant difference between the two video-rating conditions. However, it once again reveals an effect of the rating device used (mobile phone vs. computer screen). In Figure 1b, we have subsumed the two video conditions (with or without displaying the studio background) but excluded all raters that used a mobile phone ($n=20$). This plot, again, suggests an interaction between rating condition and the occurrence of accents and head movements, which this time turns out significant ($\chi^2=4.92, df=1, p=.027$). This preliminary result suggests that words with head movements tend to be rated more prominent if the movement is seen by the rater, as would be expected, but that a considerable number of raters are required to generate this effect, and that a similar rating set-up, using a computer screen instead of a mobile phone, should be used.

**Keywords:** audio-visual perception; multimodal perception; prominence; pitch accent; head movement; beat gesture; crowdsourcing; web-based

**Figures**

Figure 1. Boxplots of average prominence ratings collected in an audio-only ($n_{\text{audio}}=41$) and in an audio-visual condition with (a) $n_{\text{video}}=44$ including raters using either computers or mobile phones and (b) $n_{\text{video}}=53$ including only raters using a computer screen, but two different video display conditions subsumed (44+29-20, see text).

**References**


