Perceiving head movements in news readings: Evidence from web-based auditory vs. audio-visual prominence ratings

Gilbert Ambrazaitis[1], Johan Frid[2], & David House[3]
[1]Linnaeus University, Växjö
[2]Lund University Humanities Lab
[3]KTH (Royal Institute of Technology)

Abstract

Previous research has shown that visual information can be integrated in the perception of prominence, but the available evidence stems mostly from controlled experimental settings, often making use of synthetic stimuli. The present study provides evidence from spontaneously produced head gestures that occurred in Swedish television news readings. Materials were rated for word prominence by 85 adult volunteers in a between-subjects design (audio-visual vs. audio-only ratings) using a crowd-sourcing approach. Accented words accompanied by a head movement were perceived as more prominent than accented words lacking a head movement. Crucially, the difference in perceived prominence level between words with and without a head gesture was found to be larger in an audio-visual rating condition compared to an audio-only condition. The results suggest that visual prominence signals are integrated in speech processing even in a relatively uncontrolled, naturalistic setting, such as watching the news.

Introduction

Spoken language is essentially an audio-visual, or multimodal phenomenon, comprising audible, acoustic information and visible, kinematic information, both concerning articulatory (e.g., lip) movements (Dohen & Loevenbruck, 2009; Scarborough et al., 2009) as well as gestures, produced, for instance, with the limbs, the fingers, the torso, the head, or the eyebrows (e.g., Kendon, 2004; McNeill, 2005). In particular, speech and gesture have been shown to converge in the production of prominence, as pitch accents and stressed syllables are regularly co-produced and temporally aligned with gestures (e.g., Alexanderson et al., 2013; Ambrazaitis et al., 2020; Esteve-Gibert & Prieto, 2013; Esteve-Gibert et al., 2017; Leonard & Cummins, 2011; Loehr, 2012; Swerts & Krahmer, 2010, Yasinnik et al., 2004).

Following a standard account of gesture classification going back to McNeill (1992), we can distinguish between iconic, metaphoric, deictic, and beat gestures (or rather, dimensions of gestures, e.g., Shattuck-Hufnagel & Prieto, 2019), where beat gestures are assumed to signal prominence. In this study we focus on head movements and their role as prominence cues in speech perception. We thus refer to these movements as beat gestures or ‘head beats’.

Previous studies have shown that visually perceived beat gestures are integrated in speech perception in various ways (e.g., Al Moubayed et al., 2011; Wang et al., 2013). For instance, seeing a head movement may improve speech intelligibility (Al Moubayed et al., 2011). Furthermore, there is evidence suggesting that visually perceived gestures can contribute to perceived prominence (e.g., Krahmer & Swerts, 2007; Prieto et al., 2015). However, to our knowledge, most studies have been restricted to experimental settings typically using stimuli where the audio and video are presented separately or are non-congruent (e.g., Dohen & Loevenbruck, 2009; House et al., 2001; Krahmer & Swerts, 2007; Scarborough et al., 2009) or where carefully controlled synthetic stimuli are used (House et al., 2001; Prieto et al., 2015).

Although results from experimental settings are informative, they should be validated by means of testing spontaneously produced gestures from ecologically valid settings. A recent example for this approach is the study by Jiménez-Bravo & Marrero-Aguiar (2020), who collected audio-only and audio-visual prominence ratings for a sample of spontaneous speech taken from a Spanish television talent show. Their results showed significantly more prominence marks in the audio-visual than in the audio-only condition.

The present study continues this line of research asking how spontaneously produced gestures (albeit in a very special genre: news readings) contribute to prominence perception. To this end, we collected prominence ratings using a web-based set-up and a crowd-sourcing approach, where participants rated the words in a sample of news readings from home, via their personal computer or mobile phone, almost as if they actually watched the news. Thus, not only the material used, but also the rating situation can be ascribed a high level of ecological validity. While in Jiménez-Bravo & Marrero-Aguiar's (2020) data, head gestures were combined with manual gestures in a majority of the cases, our data (news readings) generally lack manual gestures, which enables us to focus strictly on the role of head beats.

Methods

Sixteen short video clips from Swedish television news broadcasts were rated, in a between-subjects design, by 44 participants in an audio-visual condition, and 41 participants in an audio-only condition. Ratings were collected using a web-based set-up. Each word was to be rated as either non-prominent, moderately prominent, or strongly prominent, by means of clicking the word in question until the desired prominence level was encoded though a specific colour (see below for details).

The audio-visual 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. The clips comprise speech of five different speakers (news anchors) and were taken from a larger corpus (see Ambrazaitis & House, 2017) that had previously been annotated for head movements (binary absence/presence decision per word), as well as for so-called 'big' pitch accents in Swedish (Myrberg & Riad, 2015) also known as the 'sentence accent' or the 'focal accent'.

Data collection

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 below. 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.

The rating task

Each clip was rated using a GUI including a video-player (in the audio-visual condition) or an audio-player (in the audio-only condition), an orthographic representation of the text of the clip, as well as a *Nästa* 'Next' button. 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/audio. During this initial presentation, the rating buttons (incl. the orthographic representations) were hidden. Participants in the audio-visual condition were instructed to carefully look at the video during this double screening. This was done in order to 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/audio 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.

Participants

Volunteers were recruited via social media and e-mail. They were offered a (digital) cinema ticket for their participation. A total of 85 adult native Swedish volunteers participated in the study: 44 in the audio-visual condition, and 41 in the audio-only condition. All raters were native Swedish adult volunteers with no reported hearing impairment and normal or corrected sight. They were encouraged to conduct the rating in a silent surrounding.

Analysis

The collected prominence ratings were analysed in two steps. First, four heuristic measures were calculated per rater in order to explore overall rating behaviour as a function of rating condition (audio-visual vs. audio-only). These were:

(1) $\bar{x}_{\mathrm{w}}$ (average word prominence): sum of all ratings (for all 218 words) divided by number of words (218)
(2) %W_{pr} (percentage of prominent words): number of all words that were marked either moderately or strongly prominent, divided by number of words (218), multiplied by 100
(3) %W_{st} (percentage of strong words): number of all words marked strongly prominent, divided by number of words (218), multiplied by 100
(4) %$\mathrm{W}_{\mathrm{st/pr}}$ (relation between prominence levels): number of all words marked strongly prominent, divided by number of words marked either moderately or strongly prominent, multiplied by 100

The effect of rating condition on these measures was assessed by means of independent samples t-tests.

In a second step, we explored whether the rating condition would affect specifically words that were produced with a head gesture. To this end, all words in our sample were classified as either being realized with a big accent (BA) and a head beat (HB), a BA only, or neither (using the available head beat annotations from Ambrazaitis & House, 2017). Table 1 displays token frequencies for these three categories in the selected data set. We then calculated the following three measures for each rater:

(5) $\bar{x}_{\mathrm{noBA}}$, $\bar{x}_{\mathrm{BA}}$, and $\bar{x}_{\mathrm{BAHB}}$ (average word prominence): sum of all ratings (for all words within each category according to Tab. 1) divided by *n* according to Tab. 1

The measures described in (5) were evaluated by means of linear mixed effects regression models. In particular, we tested how well average word ratings are predicted by (a) the rating condition and (b) the cumulative addition of multimodal prominence markers (+BA+HB). The latter predictor was modelled as a three-level (within-subjects) fixed effect *MMP* (*multimodal prominence*: noBA, BA, BAHB). The former predictor was modelled as a two-level (between subjects) fixed effect *AV_condition* (audio-only, audio-visual). We included an interaction term for *MMP* and *AV_condition,* and *rater* (intercepts only) as a random effect. The full model (*MMP* * *AV_condition* + (1|*rater*)) was then compared to three different reduced models using likelihood ratio tests in order to evaluate the significance of the two predictors *MMP* and *AV_condition* as well as their interaction (see Tab. 2).

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

Finally, the following difference measure was calculated for each rater:

(6) $\bar{x}_{\Delta\mathrm{HB}} = \bar{x}_{\mathrm{BAHB}} - \bar{x}_{\mathrm{BA}}$

Just as for measure (1-4), the effect of the rating condition on measure (6) was assessed using an independent samples t-test.

All modelling was done in R (R Core Team, 2012) using the lmer function from the lme4 package (Bates et al., 2012). R^2-values were obtained using the function r.squaredGLMM from the MuMIn package (Barton, 2020), and χ^2-tests and t-tests were performed using the t.test and the anova function from the stats package (R Core Team, 2012).

Results and discussion

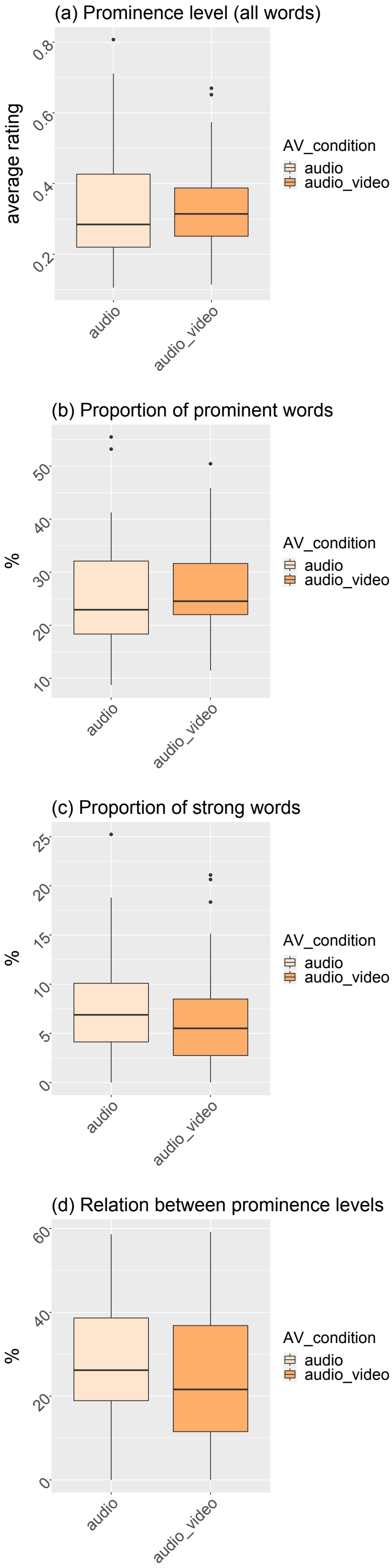
Figure 1 displays the results for the four heuristic measures defined to explore overall rating behaviour. For none of these measures did the results suggest an effect of the rating condition (audio-only vs. audio-visual; $\bar{x}_{\mathrm{w}}$: t = .11, df = 83, p = .91; %W_{pr}: t = -.41, df = 83, p = .69; %W_{st}: t = .98, df = 83, p = .33; %$\mathrm{W}_{\mathrm{st/pr}}$: t = 1.30, df = 83, p = .20). That is, raters did, on average, not assign significantly more or higher prominence ratings to words in the audio-visual condition compared to the audio-only condition. A possible interpretation of these results is that visual information did not significantly add to perceived prominence. However, it might also be the case that our four heuristic measures were too coarse to capture effects of visual prominence cues. If, for instance, a visible head beat in fact makes a word more prominent, but the effect is rather small, then it might not have a significant impact on the overall word prominence average (taken across words with and without head beats).

Therefore, in a second step, we distinguished between words produced with and words produced without an accompanying head beat. Figure 2 (next page) displays the distribution of average prominence ratings per multimodal prominence constellation (*MMP*) for the audio-only and the audio-visual condition. It can be seen – for both rating conditions – that words with big accent and head beat tend to receive higher prominence ratings than words with a big accent only, and words lacking a big accent are generally rated low. This predictive value of *MMP* is highly significant (Tab. 3). Furthermore, Table 2 shows that our models account for up to 83.2% of the observed variability in prominence ratings, whereby the contribution of *MMP* is decisive: Models reach an $R^2{}_c$ of around 83% only if *MMP* is included, and irrespective of other factors or interactions present.

That is, words accompanied by a head beat tend to be rated stronger even if no visual information on the head beat is available (in the audio-only condition). This is well in line with the results from the (audio-only) ratings obtained by Swerts & Krahmer (2010), suggesting that words with head beats are also produced with stronger acoustic prominence cues. A tendency for larger accentual big-accent rises as a function of accompanying head (and eyebrow) beats has indeed been observed in an ongoing study (Ambrazaitis & House, submitted).

The rating condition, however, had no predictive value in the linear mixed models (Tab. 2 & 3). Notably, the results did not even reveal a significant interaction between *AV_condition* and *MMP*, although Figure 2 suggests a tendency for such an interaction: For BA words, the plot suggests a tendency for slightly lower ratings in the audio-visual condition, while when a head beat is present (BAHB), a slight trend for higher ratings is seen when the visual modality is available.

Figure 1. Boxplots for the four measures of overall prominence rating behavior comparing the audio-only and the audiovisual condition: (a) $\bar{x}_w$, (b) %W_{pr}, (c) %W_{st}, (d) %$W_{st/pr}$ (see 1-4 for explanations).



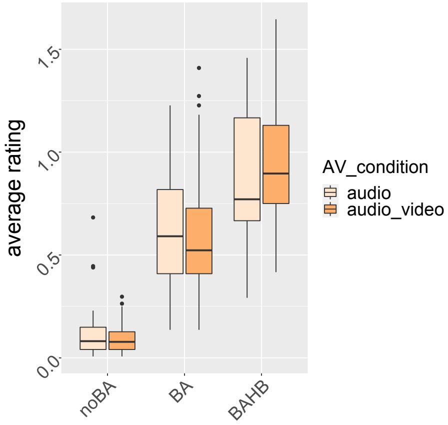


Figure 2. Boxplots of average prominence ratings collected in an audio-only and in an audio-visual condition for words realized without any big accent, with a big accent only, or a big accent and a head beat ($\bar{x}_{noBA}$, $\bar{x}_{BA}$, and $\bar{x}_{BAHB}$).

Table 2. Model fit for all full and reduced models measured using R^2. R^2_m = marginal R^2 measuring the amount of variation described by the fixed factors; $R^2{}_c$ = conditional R^2 measuring the amount of variation described by the entire model including random-effects factor; '' denotes the interaction between factors, as oppose to '+'.*

	Model	R^2_m	$R^2{}_c$
Full	MMP * AV_con. + (1|rater)	.635	.832
Red.1	MMP + AV_con. + (1|rater)	.632	.828
Red.2	MMP + (1|rater)	.632	.828
Red.3	AV_con. + (1|rater)	.000	.000

Table 3. Results of likelihood ratio tests comparing full and reduced models.

Model comparison	Effect tested	χ^2	*df*	*p*
Full vs. Red.1	*interaction*	3.93	2	.14
Red.1 vs. Red.2	*AV_con.*	.01	1	.92
Red.1 vs. Red.3	*MMP*	322.37	2	.000

Figure 3 scrutinizes this relation between BA and BAHB words, displaying the mean differences in prominence between BAHB and BA words. The figure suggests that a slightly greater distinction is made between words with and without head beats in the audio-visual compared to the audio-only condition, and this trend is significant (t = -2.5459, df = 83, p = .01). This result suggests that the visual perception of head beats indeed adds to perceived prominence.

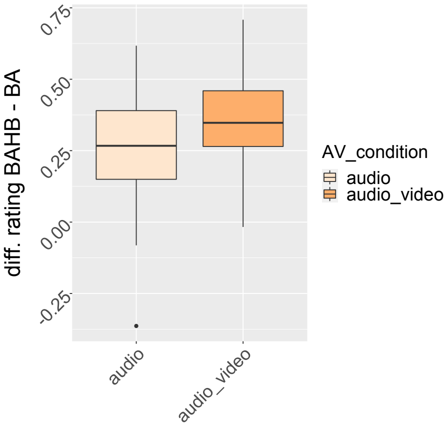


Figure 3. Boxplots for the difference measure ($\bar{x}_{\Delta HB}$) comparing the audio-only and the audio-visual condition.

Conclusions

According to a recent proposal by Holler and Levinson (2019), multimodal information can be expected to support spoken language processing in natural communicative settings, rather than to impede processing. We therefore should explicitly look for, and expect to find, evidence for audio-visual integration in the perception of ecologically valid speech samples. However, few previous studies on the impact of visually perceived gestures on perceived prominence have tested spontaneous speech or spontaneously produced gestures. Furthermore, little is known about the individual contributions of different gestural articulators (Jiménez-Bravo & Marrero-Aguiar, 2020).

The present study has provided novel evidence based on spontaneous gestures, albeit produced with non-spontaneous speech (news readings). However, the choice of this speech genre enabled us to isolated head beats from manual gestures. The results suggest that the visual perception of head beats can add to perceived prominence, although the observed effect of the availability of the visual modality was relatively small – much smaller than the effect of acoustic prominence cues.

The relatively small predictive power of the rating condition may have several explanations. First, not only the speech material was uncontrolled, but also the rating set-up. We did not, in this study, control for factors such as screen size or viewport size in the audio-visual condition (but see Ambrazaitis et al., 2019), nor did we have any possibility to control to what degree the participants really looked at the video. Moreover, and most importantly, in an uncontrolled setting, a multitude of multimodal signals are processed simultaneously (Holler & Levinson, 2019). Given these circumstances, it is noteworthy that the impact of the visual modality on the perception of prominence is robust enough to be measurable in ecologically valid data using relatively uncontrolled rating setting.

Acknowledgements

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

References

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

Al Moubayed S, Beskow J, Granström B & House D (2011). Audio-visual prosody: Perception, detection, and synthesis of prominence. In: A Esposito, A M Esposito, R Martone, V C Müller & G Scarpetta, eds, *Toward autonomous, adaptive, and context-aware multimodal interfaces. Theoretical and practical issues. Lecture Notes in Computer Science, 6456*. Berlin, Heidelberg: Springer, 55-71.

Ambrazaitis G, Frid J & House D (2019). Multimodal prominence ratings: Effects of screen size and audio device. In *Book of abstracts from the 6th European and 9th Nordic Symposium on Multimodal Communication (MMSYM 2019)*. Leuven, Belgium, 2-3.

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.

Ambrazaitis G & House D (submitted). Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters.

Ambrazaitis G, Zellers M & House D (2020). Compounds in interaction: Patterns of synchronization between manual gestures and lexically stressed syllables in spontaneous Swedish. In: *Proc of Gesture and Speech in Interaction (GESPIN2020).* Stockholm, Sweden.

Barton K (2020). *MuMIn: Multi-Model Inference.* R package version 1.43.17.

Bates D M, Maechler M & Bolker B (2012). *lme4: Linear mixed-effects models using S4 classes.* R package version 1.1-15.

Dohen M & Loevenbruck H (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech,* 52: 177-206.

Esteve-Gibert N, Borràs-Comes J, Asor E, Swerts M & Prieto P. (2017). The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America*, 141(6): 4727-4739.

Esteve-Gibert N & Prieto P (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research,* 56(3): 850–864.

Holler J & Levinson S C (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences,* 23(8): 639-652.

House D, Beskow J & Granström B (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In: *Proc of Eurospeech 2001*. Aalborg, Denmark, 387-390.

Jiménez-Bravo M & Marrero-Aguiar V (2020). Multimodal perception of prominence in spontaneous speech: A methodological proposal using mixed models and AIC. *Speech Communication,* 124: 28-45.

Kendon A (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

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

Leonard T & Cummins F (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes,* 26(10): 1457–1471.

Loehr D (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology,* 3(1): 71–89.

McNeill D (1992*). Hand and mind: what gestures reveal about thought.* Chicago: University of Chicago Press.

McNeill D (2005). *Gesture and thought.* Chicago: University of Chicago Press.

Myrberg S & Riad T (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics,* 38(2): 115–147.

Prieto P, Puglesi C, Borràs-Comes J, Arroyo E & Blat J (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics,* 49(1): 41–54.

R Core Team (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

Scarborough R, Keating P, Mattys S L, Cho T, Alwan A & Auer E T (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech,* 52: 135–175.

Shattuck-Hufnagel S, Prieto P (2019). Dimensionalizing co-speech gestures. In*: Proc. of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia, 1490-1494.

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(2): 197–206.

Wang L & Chu M (2013). The role of beat gesture and pitch accent in semantic processing: an ERP study. *Neuropsychologia,* 51(13): 2847–2855.

Yasinnik Y, Renwick M & Shattuck-Hufnagel S (2004). The timing of speech-accompanied gestures with respect to prosody. In: *Proc. of From Sound to Sense*. Cambridge, MA, USA, 97–102.