Auditory vs. audiovisual prominence ratings of speech involving spontaneously produced 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
dgilbert.ambranzaitis@lnu.se, johan.frid@humlab.lu.se, davidh@speech.kth.se

Abstract

Visual information can be integrated in prominence perception, but most available evidence stems from controlled experimental settings, often involving synthetic stimuli. The present study provides evidence from spontaneously produced head gestures that occurred in Swedish television news readings. Sixteen short clips (containing 218 words in total) were rated for word prominence by 85 adult volunteers in a between-subjects design (44 in an audio-visual vs. 41 in an audio-only condition) using a web-based rating task.

As an initial test of overall rating behavior, average prominence across all 218 words was compared between the two conditions, revealing no significant difference. In a second step, we compared normalized prominence ratings between the two conditions for all 218 words individually. These results displayed significant (or near significant, p<.08) differences for 28 out of 218 words, with higher ratings in either the audiovisual (13 words) or the audio-only-condition (15 words).

A detailed examination revealed that the presence of head movements (previously annotated) can boost prominence ratings in the audiovisual condition, while words with low prominence tend to be rated slightly higher in the audio-only condition. The study suggests that visual prominence signals are integrated in speech processing even in a relatively uncontrolled, naturalistic setting.

Index Terms: prominence perception, multimodality, beat gesture, head movement, pitch accent, visual prosody

1. Introduction

The production and the perception of spoken language are essentially audio-visual, or multimodal phenomena, comprising audible, acoustic information and visible, kinematic information, where the latter concerns both articulatory (e.g., lip) movements, e.g., [1], [2], [3], [4], [5], and [6], as well as gestures, produced, for instance, with the limbs, the fingers, the torso, the head, or the eyebrows, e.g., [4], [5], [6], [7], [8], [9], and [10]. 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., [4], [5], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], and [21].

Following a standard account of gesture classification going back to McNeill [7], we can distinguish between iconic, metaphorical, deictic, and beat gestures (or rather, dimensions of gestures, see [22] and [23]), 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., [24], [25], [26], [27], [28], and [29]. For instance, seeing a head movement may improve speech intelligibility [24], and seeing manual beats has been shown to affect semantic processing, e.g., [26], and comprehension in L2 learners, e.g., [28] and [29]. Furthermore, there is evidence suggesting that visually perceived gestures can contribute to perceived prominence, e.g., [30], [31], [32] [33], [34], and [35]. 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., [3], [4], [30], and [32], or where carefully controlled synthetic stimuli are used, e.g., [30], [31], and [35].

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-Aguíar [36], 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 relatively high level of ecological validity, compared to many previous studies. Furthermore, while in [36], 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. Based on the cited previous studies, our general expectation is to find evidence for an integration of the visual modality in the perception of prominence in our materials, and specifically higher perceived prominence of words uttered with a head beat when the visual modality is available compared to an audio-only condition.

2. Method

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 through a specific color (see below for details).

2.1. The (audio/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 sample comprises quotes of four different speakers (news anchors; two female) and were taken from a larger corpus [18] 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, see [37], also known as the ‘sentence accent’ or the ‘focal accent’.

Swedish has a binary tonal word accent distinction (Accent 1, Accent 2), implying that a content word is usually pitch-accented even when not uttered prominent at phrase level. This kind of low-level accenitation, be it Accent 1 or 2, can be conceived of as a ‘small accent’, which is realized as a fall in pitch, represented phonologically as a high-low (HL)-accent (with a difference in temporal alignment of the HL-pattern between Accent 1 and Accent 2). A ‘big’ accent, then, is realized, phonologically, by adding a high tone (H) to the existing HL-pattern, resulting in an HHL-pattern involving a pronounced rise in pitch (....LH), typically used to signal a high level of phrase-level prominence. Big accents can most often easily be identified by trained annotators and labelled with a high inter-rater reliability, see e.g. [18].

In this study, we use two versions of the speech sample described here: one including both the original audio and video display and an audio-only version.

2.2. Data collection

2.2.1. 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 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.

2.2.2. 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 Närta (‘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 color 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 (an empty display in the audio-only condition) 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/audio 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.

2.2.3. 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.

2.3. Analysis

The collected prominence ratings were analyzed in two steps. First, average word prominence ($\bar{x}_w$) was calculated per rater as a heuristic measure in order to explore overall rating behavior as a function of rating condition (audio-visual vs. audio-only). $\bar{x}_w$ was calculated according to (1) as the sum of ratings of all 218 words (where each word could be rated 0, 1, or 2), divided by the total number of words.

$$\bar{x}_w = \frac{\text{sum of all ratings}}{218}$$  \hspace{1cm} (1)

The rationale behind this measure was that if words that are produced with a head beat are rated more prominent in the audio-visual condition, then even average word prominence should be slightly higher in the audio-visual than in the audio-only condition.

In a second step, we compared the ratings obtained in the two conditions for each individual word (i.e., 218 comparisons). To this end, ratings were standardized by transforming them into z-scores according to (2).

$$z_w = \frac{x_w - \bar{x}_w}{s}$$  \hspace{1cm} (2)

In interpreting the results of these word-wise comparisons, we considered the available annotations of head movements and big accents (see 2.1). The purpose of this analysis was to explore in greater detail possible effects of the visual perception of head gestures.

The effect of rating condition on all measures (step 1 and 2) was assessed by means of linear regression models (equivalent to independent samples t-tests) using the lm() function from the stats package in R [38].

3. Results

Figure 1 displays the results of the first step of analysis: a comparison of average word ratings ($\bar{x}_w$) in the two conditions audio-visual vs. audio-only. These results suggest no effect of the rating condition on $\bar{x}_w$ ($t = .11, df = 83, p = .91$).

In a second step, the normalized ratings of each of the 218 individual words involved in the test were analyzed, yielding significant differences ($p < .05$) between the two rating...
conditions for 18 words, and for an additional 10 words if also near-significant differences (p<.08) are considered. Differences were found to occur with both polarities, that is either displaying a higher average prominence rating in the audio-visual (in about one half of the observed effects), or in the audio-only condition. An overview of these results is provided by Tables 1 and 2.

Table 1: Number of significant (p<.05) or near significant (p<.08) differences between the two rating conditions (audio-visual [AV] vs. audio-only [A]), based on comparisons for 218 individual words; see text for explanations of noBA, BA, and BA+HB.

<table>
<thead>
<tr>
<th>Direction of effect</th>
<th>noBA</th>
<th>BA</th>
<th>BA+HB</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No effect</td>
<td>132</td>
<td>20</td>
<td>38</td>
<td>190</td>
</tr>
<tr>
<td>A/V &gt; A</td>
<td>5</td>
<td>0</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>A &gt; A/V</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>148</td>
<td>22</td>
<td>48</td>
<td>218</td>
</tr>
</tbody>
</table>

While Table 1 displays a simple count of effects and their polarities (or ‘directions’ in Tab. 1), Table 2 provides details. In both tables, we distinguish between three different types of words, considering whether the word has been uttered with a big pitch accent (BA, see Sec. 2.1 for further explanation) but without a head beat (HB), or with a BA and a HB combined (BA+HB), or without any big accent or head beat (noBA).

Table 1 shows that for a majority of words (190 out of 218), no significant difference is found between the rating conditions, in line with the results of the analysis of average word ratings above. It also shows that of the 28 (near-) significant differences, a large proportion is found for words that lack a big pitch accent (11+5=16 word, or 57.1%), but also a fairly large proportion (8+2=10 words, or 35.7%) for words uttered with a combination of big accent and head beat. For most of these BA+HB words (8 words), the observed effect is in the expected direction, that is, a tendency for higher prominence ratings is found when the head gesture can be seen by the rater (i.e. in the audio-visual condition).

Table 2 shows that many of the words lacking a BA or HB (for which a significant or near significant effect was found), were function words, while BA and BA+HB words were generally content words. It also suggests that function words tend to be associated with weaker effects. Furthermore, it shows that effects were observed for all four speakers involved in the study, and that they occurred in all clips except one (clip 14).

4. Discussion

The results suggest only few and weak effects of the rating condition, and thus provide only limited evidence for the contribution of visually perceived head movements to the perceived prominence of a word. However, the results add to previous studies in some important respects. Furthermore, we argue that the lack of stronger effects is plausible, to be discussed in what follows.

To start with, the first part of the analysis suggested that 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, or vice versa. However, our second analysis – involving comparisons for all individual words – suggests that the lack of an overall effect probably has two sources: First, indeed, there is no effect for a majority of words. Second, however, the words that do display a difference between rating conditions happened to fall into two about equally-sized groups with opposite polarity of the effect, either displaying higher prominence values in the audio-visual or in
the audio-only condition. Obviously, in a measure of overall rating behavior, these two opposite effects cancel out each other.

Next, let us recognize that parts of the results are straightforwardly in line with our expectations. Eight (of 48) words uttered with a head beat (BA+HB) tended to be rated more prominent in the audio-visual than in the audio-only condition, suggesting that seeing a head beat has added to perceived prominence. Notably, this ‘visual boost effect’ was never observed for BA-words, which lack a head beat, providing further indirect support for the conclusion that seeing a head beat in the audio-visual condition indeed mattered. Furthermore, the magnitude of the effect as measured in the audio condition was relatively large for most of these BA+HB words (see slopes in Tab. 2). Yet, the impact of head beats was not very consistent, as the ‘visual boost effect’ was observed in only eight of 48 words. However, this lack of effect might have a couple of plausible explanations. First, words uttered with a head beat are always, in our materials, co-produced with a big accent, and there is evidence suggesting that such multimodal accents tend to be produced acoustically more salient than accented words lacking a head beat, e.g., [31] and [39]. Thus, BA+HB are likely to be rated very prominent already in the audio-only condition, leaving little room for even higher ratings in the audio-visual condition. This might also be related to the news speech genre, which aims to be very clear in terms of audio information alone, that is, news readers tend to produce acoustically very prominent accents. In addition, our use of a 3-point scale might have impeded a clear distinction between the conditions for BA+HB words. This explanation could be tested using a more fine-grained scale, which could be implemented using a continuum of shades, instead of only two colors, in our rating buttons. A second possible explanation is that head beats were produced naturally and can be expected to differ in the extent and visibility of the movement. Thus, in some words, the movements might simply have been too weak to have a serious impact on the prominence ratings. We have, however, no kinematic measurements available to support this claim. Finally, the rating situation was largely uncontrolled. We did not, in this study, control for factors such as screen size or viewport size in the audio-visual condition, but see [40]. Also, raters might not have been as attentive as in a lab setting. Moreover, and most importantly, in an uncontrolled setting, a multitude of multimodal signals are processed simultaneously [6]. Thus, given all these circumstances, it is noteworthy that for some words we indeed could observe a ‘visual boost effect’.

Let us now turn to those results that may appear unpredicted at first sight. First, in a considerable number of cases (11), a word that was neither produced with a big accent nor a head gesture, hence most likely a word with an overall low prominence level, was rated more prominent in the audio-only than in the audio-visual condition. This relatively frequent phenomenon suggests that rather subtle acoustic cues that might signal some low degree of prominence had a greater impact on prominence ratings when no video display was available. This in turn might mean that perception is more sensitive for acoustic differences when only audio is available, which, we would argue, provides further indirect evidence for the integration of the visual modality in prominence perception. This increased sensitivity, or ‘audio boost effect’, in the audio-only condition might then, likewise, apply to higher prominence levels as well and explain the two instances of BA-words that received higher ratings in the audio-only condition. These less expected results might thus well be argued to be in line with our general expectation, that is, prominence perception being affected by the presence or absence of the visual modality.

Another unexpected result are the five instances of words lacking a BA and an HB that nevertheless tended to be rated more prominent in the audio-visual condition. As shown in Figure 2, three of them happened to occur in the same sentence. Notably, two of them (ta and om) directly preceded or succeeded the same prominent word (hand) uttered with BA+HB displaying a highly significant ‘video boost effect’. It would seem likely to assume that the two adjacent noBA-words (ta and om) were boosted as a side effect of the HB on hand. This example reveals a level of complexity in prominence perception that has hardly been discussed in previous research. Finally, we will leave the few so-far uncommented aspects of the results unexplained, as they would appear marginal.

Figure 2: Annotated text for an excerpt of one of the clips used in this study. The words öppna and hand were produced with big accents (BA), and hand additionally with a head beat (HB); the label VID denotes words which were rated more prominent in the A/V condition (significant at p<.01(**), p<.05(*) or p<.08(†), respectively); red circles mark words with a VID effect despite lacking a HB.

5. Conclusions

According to a recent proposal by Holler and Levinson [6], multimodal information can be expected to support spoken language processing in natural communicative settings, rather than to impede processing. We therefore might expect to find evidence for audio-visual integration even in the perception of ecologically valid speech samples. 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 isolate head beats from manual gestures, which can be difficult using spontaneous speech, see [36]. Our results suggest, once again, that the visual perception of head beats can add to perceived prominence, and more generally, that the presence vs. absence of visual information has an impact on prominence perception. While visual information can boost prominence ratings when available, acoustic information can boost prominence ratings when the visual modality is absent. This study showed that the impact of the visual modality on the perception of prominence is robust enough to be measurable in speech data involving spontaneously produced (head) gestures, using a relatively simple and uncontrolled rating setting.

6. Acknowledgements

This work was supported by two grants from the Swedish Research Council (VR-2017-02140 and VR-2013-2003) and by the research network GEHM (GEstures and Head Movements in language), funded by the Independent Research Fund Denmark (9055-00004B).
7. References


