Laboratory Phonology
Journal of the Association for Laboratory Phonology

Ambrazaitis, G., & House, D. 2022. Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 13(1): 15, pp. 1–35. DOI: https://doi.org/10.16995/labphon.6430

Open Library of Humanities

Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters

Gilbert Ambrazaitis, Department of Swedish, Linnæus University, Växjö, Sweden, gilbert.ambrazaitis@lnu.se

David House, Division of Speech, Music and Hearing, KTH (Royal Institute of Technology), Stockholm, Sweden, davidh@kth.se

This study investigates the multimodal implementation of prosodic-phonological categories, asking whether the accentual fall and the following rise in the Swedish word accents (Accent 1, Accent 2) are varied as a function of accompanying head and eyebrow gestures. Our purpose is to evaluate the hypothesis that prominence production displays a cumulative relation between acoustic and kinematic dimensions of spoken language, especially focusing on the clustering of gestures (head, eyebrows), at the same time asking if lexical-prosodic features would interfere with this cumulative relation. Our materials comprise 12 minutes of speech from Swedish television news presentations. The results reveal a significant trend for larger f_o rises when a head movement accompanies the accented word, and even larger when an additional eyebrow movement is present. This trend is observed for accentual rises that encode phrase-level prominence, but not for accentual falls that are primarily related to lexical prosody. Moreover, the trend is manifested differently in different lexical-prosodic categories (Accent 1 versus Accent 2 with one versus two lexical stresses). The study provides novel support for a cumulative-cue hypothesis and the assumption that prominence production is essentially multimodal, well in line with the idea of speech and gesture as an integrated system.

1. Introduction

Speech and gesture form an integrated system and are widely considered to be co-generated in spoken language production (e.g., McNeill, 1985, 2005; Kendon, 2004; Willems & Hagoort, 2007). This claim is supported, for instance, by evidence suggesting that speech and gesture have a common developmental origin in early hand-mouth linkages (Iverson & Thelen, 1999; see also, e.g., Esteve-Gibert & Prieto, 2014). Furthermore, it has been shown that "[w]hen speech stops, gesture stops" (Graziano & Gullberg, 2018, title; see also McNeill, 1985), meaning that disfluencies in speech production tend to affect both speech and gesture. Also, a multitude of studies have shown that gesture and speech are coordinated in time and space.

For instance, gestures are temporally aligned with units of speech (e.g., McNeill, 1992; McClave, 2000; Kendon, 2004; Yasinnik, Renwick, & Shattuck-Hufnagel, 2004; Jannedy & Mendoza-Denton, 2005; Flecha-García, 2010; Swerts & Krahmer, 2010; Leonard & Cummins, 2011; Loehr, 2012; Alexanderson, House, & Beskow, 2013; Esteve-Gibert & Prieto, 2014; Krivokapić, 2014; Ambrazaitis & House, 2017a; Graziano & Gullberg, 2018; Shattuck-Hufnagel & Ren, 2018, Ambrazaitis, Zellers, & House, 2020b; *inter alia*). This alignment has, furthermore, shown to be affected by linguistic prosodic structure (e.g., Esteve-Gibert & Prieto, 2013; Krivokapić, 2014; Esteve-Gibert, Borràs-Comes, Asor, Swerts, & Prieto, 2017), making an even stronger case for speech and gesture being planned in conjunction than the mere co-occurrence. Moreover, speech and gesture not only converge in the temporal domain, but also in the 'spatial' domain in a broad sense, displaying correlations between the presence, magnitude, or complexity of movements with the extension of articulatory or acoustic parameters of speech (e.g., Krahmer & Swerts, 2007; Parrell, Goldstein, Lee, & Byrd, 2014; Pouw, Harrison, & Dixon, 2020a; Pouw, Harrison, Esteve-Gibert, & Dixon, 2020b; Pouw, de Jonge-Hoekstra, Harrison, Paxton, & Dixon, 2021). However, less is known on convergence in space than on convergence in time.

The present study critically extends these lines of research on the spatio-temporal convergence of speech and gesture, by means of studying the multimodal production of prominence, considering a couple of understudied areas: First, we focus on the spatial domain, asking how the acoustic realization of pitch accents co-varies with the presence of accompanying gestures. Second, we investigate non-manual gestures (head and eyebrow gestures, and thereby also the clustering of different gestures). Third, we study spontaneous, non-elicited gestures in an ecologically valid material. And finally, we wonder if a cumulative relation between acoustic and kinematic prominence signals, as we refer to it, might be sensitive for the prosodic-phonological structure of the accented word. In particular, we ask whether an accentual pitch movement is enlarged when accompanied by head or eyebrow gestures, irrespective of whether the pitch movement reflects a lexical or a post-lexical tone, or whether it is associated with a stressed syllable or not.

1.1. Gesture and prominence

A growing line of research focuses on the role of gestures in prominence production and perception (e.g., McNeill, 1992; McClave, 2000; House, Beskow, & Granström, 2001; Beskow, Granström, & House, 2006; Krahmer & Swerts, 2007; Flecha-García, 2010; Roustan & Dohen, 2010; Rusiewicz, 2010; Swerts & Krahmer, 2010; Al Moubayed, Beskow, Granström, & House, 2011; Leonard & Cummins, 2011; Alexanderson et al., 2013; Parrell et al., 2014; Prieto, Puglesi, Borràs-Comes, Arroyo, & Blat, 2015; Ambrazaitis & House, 2017a; Krivokapić, Tiede, & Tyrone, 2017; Ambrazaitis, Frid, & House, 2020a; Jiménez-Bravo & Marrero-Aguiar, 2020). In a standard account of gesture classification going back to McNeill (1992), a distinction is made between iconic, metaphoric, deictic, and beat gestures. Beat gestures differ from the other categories in that they, rather than functioning to convey semantic content, have been assumed to be used to construct rhythmical structures or highlight words or expressions, in other words: to signal prominence. Although the term beat gesture often refers to movements by the hands or a finger (e.g., Casasanto, 2013), we and others have earlier used it in connection with head and eyebrow movements (e.g., Krahmer & Swerts, 2007; Ambrazaitis & House, 2017a) and suggest extending its usage to all types of prominence-related movements. Henceforth, we thus use the term beat gesture or the short-hand terms head beat and eyebrow beat for the two types of beat gestures of interest in this study.

Just as gesture and speech in general appear to be best characterized as tightly integrated, much of the available evidence concerning beats and prominence in particular points in the same direction: It has been shown, for instance, that seeing beat gestures can facilitate speech processing (Biau & Soto-Faraco, 2013; Wang & Chu, 2013), suggesting a firm integration of beat gestures with the human language capacity. Also, beat gestures are typically co-produced and aligned with pitch accents or stressed syllables in speech production (e.g., Yasinnik et al., 2004; Jannedy & Mendoza-Denton, 2005; McNeill, 2005; Flecha-García, 2010; Swerts & Krahmer, 2010; Leonard & Cummins, 2011; Loehr, 2012; Alexanderson et al., 2013; Esteve-Gibert & Prieto, 2013; Shattuck-Hufnagel, Ren, Mathew, Yuen, & Demuth, 2016; Esteve-Gibert et al., 2017; Ambrazaitis & House, 2017a; Kelterer, Ambrazaitis, & House, 2018; Shattuck-Hufnagel & Ren, 2018; Ambrazaitis et al., 2020b; see also Pouw, Trujillo, & Dixon, 2020c). For instance, Flecha-García (2010) found that eyebrow raises preceded pitch accents by on average 60 ms in a corpus of English face-to-face dialogue. In studies by Leonard and Cummins (2011) and Loehr (2012), a correlation was found between the apices of gestural strokes and pitch accents. Synchronization between three different phases of head nods and stressed syllables carrying a so-called big accent in Swedish (see Section 1.2) was found by Alexanderson et al. (2013).

Another type of temporal convergence concerns the concurrent lengthening, rather than timing, of gesture and units of speech. For instance, Rusiewicz, Shaiman, Iverson, and Szuminsky

(2014) showed that manual pointing gestures are lengthened as a function of prominence. Manual pointing gestures were also studied by Krivokapić et al. (2017), who combined electromagnetic articulography (EMA) and a motion capture system to simultaneously record articulatory movements and hand movements. Their results showed that both lip movements (for /b/) and a part of the manual gesture were lengthened as a function of the elicited prominence level.

However, while many studies have focused on the temporal aspects of co-production, less is known about the spatial kinematic or the acoustic and articulatory characteristics of co-produced pitch accents and gestures. Moreover, to our knowledge, the available evidence for a conversion of acoustic or articulatory and spatial gestural features is limited to manual gestures (Krahmer & Swerts, 2007; Parrell et al., 2014, Pouw et al., 2020a; Pouw et al., 2020b; Pouw et al., 2021), and the available results are, we argue, not entirely conclusive. For instance, Roustan and Dohen (2010) examined acoustic and articulatory measures and observed no differences between accents with a (deictic or beat) gesture compared to accents without a gesture. By contrast, Krahmer and Swerts (2007) found that words produced with a beat gesture differ from words without gesture in terms of duration and F2 of the stressed vowel. Crucially, both studies tested for, but did not observe any differences in f_o. Pouw et al. (2021), however, did observe effects of arm movements on f_o, which were more pronounced for larger movements of the arms than for smaller wrist movements, suggesting a biophysical coupling between arm movements and respiratory-related activity.

Evidence for a spatial convergence involving less forceful movements has also been found, although in an articulatory, rather than acoustic study, by Parrell et al. (2014). They used EMA to record movements of the lips and the right index finger while participants repetitively tapped the finger and simultaneously repeated a single spoken syllable. Their results showed that prominence will affect both the duration (as in Krivokapić et al., 2017, reviewed above) and the magnitude of articulatory movements, irrespective of whether the prominence is intentionally expressed through a gesture (a tapped finger) or through speech. Evidence of this kind suggests a tight link between gesture and speech in production, indicating in particular that "implementation of prosody is not domain-specific but relies on general aspects of the motor system" (Parrell et al., 2014, abstract).

Most of the studies reviewed in the previous paragraphs suggest a cumulative relation between speech and gesture in the production of prominence: the higher the prominence level to be produced, the (temporally) longer or (spatially) larger the speech- or gesture-related movements, and the stronger the acoustic correlates. However, the available evidence so far stems from studies involving manual gestures only, elicited in controlled experiments and predetermined. For instance, participants were instructed to produce a pointing gesture with a given word (e.g., Roustan & Dohen, 2010; Krivokapić et al., 2017). In some studies, the movements were rather unlike authentic co-speech gestures (e.g., Parrell et al., 2014, Pouw et al., 2021). Although the results of the cited studies are informative, the hypothesis of a cumulative relation should be further evaluated for spontaneous gestures, including non-manual gestures.

Therefore, the present study focuses on spontaneous head and eyebrow gestures, albeit produced with scripted speech and in a very specific genre (news readings). For news readings, there is evidence suggesting that prominence-lending head and eyebrow gestures are more likely to occur with perceptually strong accents than with weak ones, indeed suggesting a cumulative relation: Swerts and Krahmer (2010) had a listener panel rate the prominence of all words in a 1000-word corpus of Dutch news presentations. Listeners were asked to mark words that appear to be highlighted, and from that simple binary decision for each word (yes/no) they derived a three-way classification of words into strongly, weakly, and not accented. The rating was performed on the audio-signal, without access to the video channel. Then, an independent annotation of gestures (head and eyebrow movements) was performed using the video channel only. The results showed that words classified as strongly accented often (in 67% of the cases) co-occurred with both head and eyebrow annotations, while weak accents were most typically produced without a gesture (47%) or with only a head (16%) or an eyebrow movement (13%).

However, Swerts and Krahmer (2010) did not perform any acoustic measurements in order to explore whether the number of accompanying movements would correlate with the production of pitch accents. Acoustic data would be informative, because another factor that might explain their perceptual ratings are top-down processes (e.g., Fant, Kruckenberg, & Liljencrants, 2000; Wagner, 2005; Kleber & Niebuhr, 2010; Baumann & Winter, 2018). In particular, a part-of-speech effect may lie behind the obtained higher prominence-ratings for words with gestures, if we assume that a majority of beat gestures in Swerts and Krahmer (2010) co-occurred with content words (which are inherently more prominent than function words, see Fant et al., 2000; Baumann & Winter, 2018). However, words with gestures might still have been produced with stronger acoustic prominence correlates, such as enlarged accentual f_o ranges or longer stressed syllable durations, than words without gestures, since we know that such acoustic measures strongly correlate with perceived prominence (e.g., Fant et al., 2000; Baumann & Winter, 2018). Thus, the data by Swerts and Krahmer (2010) might well entail further evidence for a cumulative relation.

Moreover, this cumulative relation in the production of prominence might not only relate to the gradual, phonetic domain: A common generation process of the acoustic and the kinematic dimensions of speech might also, generally, predict that gesture production may reflect traces of phonological structure (e.g., Esteve-Gibert & Prieto, 2013; Krivokapić, 2014; Kelterer et al., 2018). For instance, results by Esteve-Gibert and colleagues suggest that the coordination of gestures with prominence-related landmarks in the acoustics is governed by prosodic structure (Esteve-Gibert & Prieto, 2013, for manual gestures; Esteve-Gibert et al., 2017, for head gestures). Another relevant study by Kelterer et al. (2018) suggests that a prominence-lending head beat might require a lexically stressed syllable (be it a primary or secondary stress) to be associated with, thus generally behaving comparable to a pitch accent. We might thus formulate the stronger prediction that gesture production is in some way affected by lexical or phonological prominence structures. The Swedish language comprises a set of lexical prosodic conditions that can serve as a case to test this prediction.

1.2. Swedish lexical prosody

Swedish exhibits lexical stress, but also a binary tonal distinction traditionally referred to as the 'word accents' or 'lexical pitch accents' Accent 1 (A1) and Accent 2 (A2). Accentual tones are associated with the stressed syllable, as shown in (1). In the Stockholm variety, treated in this study, A1 is typically understood as an (H)L*-tone, while A2 is modelled as an H*L-tone.

The word accents can be characterized as to some degree lexical but are generally assigned to words by means of phonological or morphological rules. This is exemplified in (1) illustrating a case where the plural suffix /ar/ is lexically specified for A2; note, however, that phonologically, the tone associates with the stressed syllable located in the stem.

The pitch pattern associated with a word accent does, of course, induce a certain level of phonetic (i.e., perceptual) prominence. However, this prominence is traditionally not regarded as a phrase- or sentence-level phenomenon (e.g., Bruce, 1977; Gussenhoven, 2004; but see also Ambrazaitis, 2009; Myrberg, 2010). Hence, the two word accent categories (A1 and A2) are generally not assumed to induce different levels of prominence either (Bruce, 2007).

Prominence at the phrase level is, in the Stockholm variety, achieved by means of an additional tonal rise (usually modelled as an H-tone), realized in sequence with the word accent pattern (a pitch fall) (Bruce, 1977). This additional pitch movement is most commonly referred to as a 'focal accent,'[1] but we will here adopt the term 'big accent' (as opposed to a 'small accent' when the additional pitch rise is missing) as proposed by Myrberg and Riad (2015). A big accent, then, is characterized by a (falling-)rising (H)L*H pattern in A1, and a two-peaked H*LH pattern in A2. The initial (H) for A1 is set in parentheses, since the accentual HL-fall in A1 is frequently realized as very small or is even absent in a big accent (e.g., Engstrand, 1997). The four resulting basic accentual patterns of Swedish are schematically displayed in **Figure 1**.

[1] The term 'focal accent' is problematic because it suggests that the HLH-accent (i.e., with the additional final H-tone) is (exclusively) used to mark focus. However, studies by Ambrazaitis (2009) and Myrberg (2010) have clearly shown that this is a misconception, for two reasons. First, there are situations where an HLH-accent fulfils other functions than signaling focus (e.g., Myrberg, 2010; this is also evident in our current news data, where there are far more HLH-accents than focus constituents), and second, focus can under certain conditions be realized without the H-tone (Ambrazaitis, 2009). In fact, Bruce's (1977) original term was the more adequate (i.e., less function-related) term 'sentence accent,' which was indeed reactivated by Ambrazaitis (2009). Myrberg and Riad (2015) went a step further and coined the terms 'small' and 'big accents' in order to strictly focus on form when describing the accents. This seems adequate because even the 'small' accent can function as a 'sentence accent.'

```
(1)    Small accents in Stockholm Swedish, tonal associations
       a.  Accent 1 (A1):       bilen         /ˈbiː   lɛn/    ‘the car’
                                                |
                                            H   L*
       b.  Accent 2 (A2):       bilar         /ˈbiː   lar/     ‘cars’
                                                |
                                               H*L

(2)    Big accents in Stockholm Swedish, tonal associations
       a.  Accent 1 (A1):       bilen         /ˈbiː   lɛn/                ‘the car’
                                                |
                                          (H)  L*H
       b.  Accent 2 (A2):       bilar         /ˈbiː   lar/                 ‘cars’
                                                |
                                               H*L    H
       c.  Accent 2 (A2),    bilförsäljare   /ˈbiːl   fœr   ˌsɛlː  jarɛ/  ‘car dealer’
           compound:                            |             |
                                               H*L            H
```

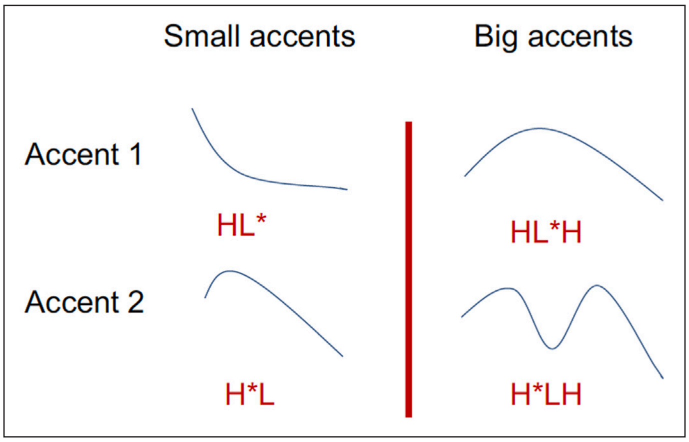

Figure 1: Schematic representations of pitch patterns associated with big and small pitch accents in Stockholm Swedish. Tonal representations in this illustration are based on Bruce (1977), assuming an H-tone in Accent 1, even if the HL-fall is often elided in a big accent, as also illustrated here (i.e., no fall prior to the rise L*H of the big Accent 1). The contours represent a phrase-final position (with an L% boundary).

Moreover, words in Swedish can have two lexical stresses—a primary and a secondary one. Words with two stresses (usually compounds) are generally assigned A2, and the H-tone of the big accent associates with the secondary stress in these words. Hence, unlike in simplex (i.e., single-stress) A2 words, where only the accentual fall associates with the stressed syllable, in double-stress words the H-tone of the big accent is also associated with a stressed syllable (the secondary stress), as illustrated in (2).

Finally, it is most widely assumed that the distinction between A1 and A2 is privative, where A2 is the marked case. This is related to the assumption that only A2 is specified for lexical tone. The tones that surface in A1 are assumed to be assigned post-lexically. However, the account of Riad (e.g., Riad, 2006; Myrberg & Riad, 2015) includes a more detailed analysis where tone is assumed to be lexical only in single-stress A2 words. Compounds (i.e., words with two stresses) are assumed to be assigned A2 post-lexically. We can thus distinguish between three lexical-prosodic word categories in Stockholm Swedish, displayed in **Table 1**.

Word accent	Stressed syllables	Lexical tone	Typical morphological structure	Lexical-prosodic category ID used in this study
Accent 1	One (primary)	no	simplex	simp_A1
Accent 2	One (primary)	yes	simplex	simp_A2
Accent 2	Two (primary + secondary)	no	compound	comp_A2

Table 1: Prosodic-phonological features of words in Swedish.

1.3. Goals and scope

Most of the existing evidence on speech-gesture coproduction suggests that it would be reasonable to predict a cumulative relation between gestural, articulatory, and acoustic prominence cues (cf. our review in Section 1.1), in the sense that spatial and durational features of gestures are positively correlated with acoustic, spatial, and durational features of speech. Furthermore, the results by Swerts and Krahmer (2010) suggest that prominence correlates not only with the spatial extension, but also with the number of visual prominence signals, as a moderate increase in prominence is more likely to coincide with either a head or an eyebrow beat (but not both), while both types of visual beats in combination are only expected with higher prominence. We therefore formulate the following *Cumulative-cue hypothesis* (note the wording “more or stronger”): Acoustic and kinematic correlates of prominence are cumulative, in the sense that they correlate positively—the more or stronger the acoustic cues applied in the production of a target prominence level, the more or stronger the kinematic cues.

Direct evidence in favor of this hypothesis is, to the best of our knowledge, mostly based on elicited, manual gestures (cf. Section 1.1). In addition, the available evidence is mostly relevant for one part of the hypothesis with respect to gestures (the ‘stronger,’ rather than the ‘more’). The present study aims to shed new light on the hypothesis by means of testing it for two, in this connection, previously understudied types of gestures and their combination (head and eyebrow movements) as produced spontaneously in naturalistic speech (albeit in a special genre, news readings). Crucially, we thereby explicitly aim at the 'more'-part of the hypothesis with respect to gestures, testing whether the number of accompanying gestures (0, 1, or 2 gestures) correlates with the strength of the acoustic signal.

Thus, although the present study critically extends previous evidence, its scope is still limited, covering the 'stronger'-part of the hypothesis with respect to acoustics, but the 'more'-part with respect to gestures. It thus focuses on acoustic, rather than on kinematic detail, not least because it is based on pre-existing annotations of big accents (henceforth, BA), head beats (HB), and eyebrow beats (EB) (Ambrazaitis & House, 2017a, although the dataset has been considerably increased for the present study). These were categorical: Only the presence versus the absence of events (BA, HB, EB) was judged upon, but no spatiotemporal kinematic detail was labelled. This appeared adequate for the exploratory approach in Ambrazaitis and House (2017a), where patterns of co-occurrence of accents and gestures were of interest. For the present study, we added a detailed analysis of the realization of the BAs. A similar manual enrichment of the gesture annotations was not deemed feasible within the present study, and also not essential for testing the 'more'-part of the hypothesis.

Moreover, a more specific goal of this study is to evaluate the cumulative-cue hypothesis from a phonological angle, asking whether a potential cumulative relation between the visual and the auditory mode would in some respect be sensitive for lexical prosody. To this end, we focus on Swedish words uttered with a so-called big accent (see Section 1.2) and study the range of the accentual fall (the (H)L* or H*L in A1 or A2, respectively), as well as of the big-accent rise (H), as a function of accompanying head and eyebrow movements. Our previous analyses (Ambrazaitis & House, 2017a) of the data chosen for the present study have shown that the occurrence of head and eyebrow movements in Swedish news presentations is cumulative *per se*: Words may either carry a big accent (without additional movement: BA), or a big accent and a head beat (BA + HB), or a big accent, a head beat, and an eyebrow movement (BA + HB + EB); eyebrow movements hardly ever occur independently of head movements in our data. Thus, the data at hand provide us with a three-level factor describing the multimodal make-up of an intonationally prominent word in Swedish: BA, BA + HB, BA + HB + EB.

The phonological angle in our analysis is two-fold. First, we treat the accentual fall and the big-accent rise separately, thereby acknowledging their fundamentally different phonological functions (see Section 1.2). Second, we distinguish between three different prosodic word types, according to **Table 1**. This design enables us to compare, for instance, big-accent rises that are associated with a stressed syllable with those that are not, or accentual falls that are assumed to represent a lexical versus a post-lexical tone. The research question we ask in this study is thus as follows: Assuming that the presence of a head or combined head/eyebrow gesture would correlate positively with the f_o range of an accentual f_o movement, would such a relation be observed both for the accentual fall and the accentual rise, and irrespectively of the phonological characteristics of the movement?

2. Method

2.1. Materials

This study is based on audio and video data of 60 brief news readings from Swedish Television (Sveriges Television, 2013), comprising 1936 words in total, or about 12 minutes of speech. Each news reading typically contains one–three sentences. The data set includes speech from five news anchors: two female (Sofia Lindahl, Katarina Sandström) and three male (Pelle Edin, Filip Struwe, Alexander Norén). We refer to the news anchors by their first names. The selection of news anchors was random (only meeting the requirement of including both male and female speakers). The recordings were retrieved on DVD from the National Library of Sweden (Kungliga Biblioteket). The broadcasts were cut into 'stories,' and 60 of them were included in this study. They can be identified using the list provided in Appendix A, where date, time of the news broadcast, and a key phrase (in Swedish) is provided that summarizes the content of the story. The amount of material included per news anchor varied slightly, as 20 stories stem from Alexander,[2] whereas only ten each were included from the other four anchors.

2.2. Annotations

The material was transcribed, segmented at the word level, and annotated for big accents (BA), head beats (HB), and eyebrow beats (EB) using ELAN (Sloetjes & Wittenburg, 2008) and Praat (Boersma & Weenink, 2018). These annotations are detailed in Sections 2.2.1 and 2.2.2 (see also Appendix B.1). In addition, f_o landmarks were annotated using Praat and the Praat script ProsodyPro (Xu, 2013) (see Section 2.2.5).

The materials were annotated in two phases, involving different groups of annotators (six different individuals in total), and thus form two subsets with respect to some of the annotation principles and inter-rater reliability testing (Section 2.2.3 and Appendix B.2). The older part of the dataset (30 stories, used in Ambrazaitis & House, 2017a) was labelled by three annotators (subset 1; Annotators 1–3, including the two authors of this paper). For the present study, 30 additional stories were annotated by two new annotators (subset 2; Annotators 4–5). No distinction is made between these subsets in the presentation of the results. A sixth annotator was recruited for an additional round of control annotations (Annotator 6; see Sections 2.2.2 and 2.2.3 for details).

[2] The imbalance in the materials (20 stories from Alexander, 10 each from the other speakers) is explained by the fact that we, for a first exploratory analysis (Ambrazaitis & House, 2017a), decided to focus on Alexander as our primary speaker. This first analysis therefore included 20 files from Alexander, and only 11 additional files (from three additional speakers) for the purpose of validation. For the present paper, we aimed to increase the amount of annotations considerably, but our resources were limited. (We increased the data volume by 100%, from 30 to 60 files [one of the original 31 files was excluded for technical reasons], but a certain overweight of Alexander remained in the data set).

2.2.1. Annotation of big accents

A big accent was annotated when a rising f_o movement corresponding to the second H-tone (cf. Section 1.2) was recognizable in the f_o contour; note that this f_o movement was expected in the stressed syllable for A1 words, but later in the word, surfacing as a second peak, in A2 words. Big accents were annotated with access to the audio channel, an f_o display, and the word segmentations, but without using the video display. Obvious annotation errors were corrected (see Appendix B.3 for details).

2.2.2. Annotation of head and eyebrow movements

The annotation scheme for head and eyebrow beats (HB, EB) was simple in that only the presence versus absence of the event of interest was judged upon. That is, no time-aligned annotations were undertaken and hence, no decisions had to be made upon temporal onsets and offsets of the HB and EB movements (see Section 1.3). Likewise, no distinctions were made in the annotations between types or directions of movements, meaning that any kind of head or eyebrow movement was considered (e.g., upwards, downwards, to the left or right), as long as it adhered to the following criteria: 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 (cf. Esteve-Gibert et al., 2017, who argue that the prosodic word is the scope of a head movement). That is, slower movements were ignored, which could occur, for instance, in connection with the re-setting of the head position, which often spanned several words. Most typically, the observed movements were monodirectional (i.e., a simple movement, e.g., upwards, downwards, or diagonally upwards towards the left or right etc.), but could also be more complex (e.g., up and down again).

Instructions to annotators were deliberately kept simple, as in the description in the previous paragraph, because we intended to rely on the annotator's human capacity of recognizing beat gestures. The instructions turned out to be sufficiently detailed to generate a good inter-rater agreement (see Section 2.2.3 below).

As the instructions made direct reference to the spoken words ("roughly within the temporal domain of the word"), we annotated gestures with access to the audio (and to the graphic representation of the written transcript). Annotating gestures is frequently performed using the video display only, with the audio muted, in order to avoid possible confounds introduced by the spoken message (e.g., Swerts & Krahmer, 2010; Graziano & Gullberg, 2018; *inter alia*). The rationale behind our procedure was that beat gestures, if they work as prominence markers, must relate to words, and hence, annotating with reference to the spoken words would be best suited to find the relevant movements, and to sort out potentially irrelevant movements such as the occasional slow re-settings mentioned above. However, this procedure might entail the risk that the auditory perception of acoustic prominence cues influences the visual perception of beats.

In order to test for a possible confound of this kind, a random selection of eight files (of 60), containing 247 words, were additionally annotated for head movements (HB), using only the video display (i.e., using neither the audio nor the graphic representation of the text). A new annotator was recruited for this purpose (Annotator 6), who had not listened to or watched the materials or otherwise been involved in the study previously. The annotator was instructed in a similar way to the previous annotators, but without reference to the words: He was asked to look for "relatively quick, clearly visible" head movements and to determine the onset and offset of the movement. It was explained that there may be slower resetting movements which should be ignored. Before going through the video frame by frame (to determine onsets and offsets), the annotator was instructed to watch each video a few times at normal speed in order to gain an impression of the relevant movements that are to be annotated. Again, all possible directions of head movements were considered. The annotator was instructed to identify the onset and the offset of a HB as the point in time where the head leaves or enters a resting position. It was explained that the 'resting' position only means that the head is not moving; the head may 'rest' in a neutral, but also raised, lowered, or tilted position. The movement defined in this way could either be monodirectional or more complex. These additional annotations were then compared with the original annotations made with access to the audio, revealing a strong correspondence (see Section 2.2.3).

A problem that arose with our original procedure, however, is that beat gestures can occur in the vicinity of the border between two adjacent words (or even span larger parts of two words), while annotators were requested to associate each beat with one word at a time. That is, they were forced to decide upon which word to label for HB/EB in cases where a movement spanned two adjacent words. This decision was made to ensure that each annotation of a beat gesture would correspond to exactly one observation in the analysis. As a result, some variation was observed across annotators with respect to which one of two adjacent words was associated with a beat gesture. However, this variation could be easily handled in connection with the association of gesture labels with BA-labels, as described in Appendix B.5.

2.2.3. Annotators and inter-rater reliability

The first 30 files (subset 1) were annotated (for BA, HB, EB) by three annotators (Annotators 1–3), independently of each other. Inter-rater reliability was tested using Fleiss' κ (Fleiss, 1971), and turned out fair to good (BA: $\kappa = 0.77$; HB: $\kappa = 0.69$; EB: $\kappa = 0.72$). For the analysis, our three-fold annotations were converted to a single, consensus (i.e., majority) rating for each word. The additional 30 stories (subset 2) were labelled by two new annotators (Annotators 4–5), where one annotator was responsible for either HB or EB and BA in a given file (see Appendix B.2 for details). However, an overlap of 13 files was annotated twice completely, that is, with annotations for BA, HB, and EB by both annotators, in order to test for inter-rater reliability using

Cohen's κ (Cohen, 1960). This, again, provided satisfying results (BA: $\kappa = 0.88$; HB: $\kappa = 0.77$; EB: $\kappa = 0.77$).

Annotator 6 was recruited for the additional HB annotations performed with access to the video only. Note that this annotation was restricted to head movements. After the annotation procedure described in Section 2.2.2 was completed, the time-aligned HB annotations were, in a second step, associated with the words they overlapped with (for details concerning this procedure, see Appendix B.4). A Cohen's κ was then calculated based on these word-associated annotations comparing them to the original (word-based) annotations which were made with access to the audio track, revealing a good correspondence ($\kappa = 0.80$). That is, the correspondence was not perfect, but it reached the same order of magnitude as when comparing different annotators within the original (audio- and video-based) annotations, meaning that the deviations observed lie within the expected range of annotator variability, and can most likely be explained as such. This comparison of annotation strategies (with versus without audio) did thus not reveal any strong evidence for an audio-confound.

All calculations of κ-values were made in R (R Core Team, 2012) using the irr package (Gamer, Lemon, Fellows, & Singh, 2019).

2.2.4. Classification of movements and accents as multimodal prominence constellations (MMP)

As a next step, words with a BA were classified as belonging to either of the three conditions of interest, namely, BA, BA + HB, or BA + HB + EB. This classification was not entirely trivial and is explained in detail in Appendix B.5. **Table 2** displays the outcome of this classification in terms of sample sizes for the three multimodal prominence constellations (henceforth, MMP) included in this analysis. The total number of words included is thus 543. Note that words associated with other combinations (such as BA + EB, lacking HB) were sparse in this material (see Ambrazaitis & House, 2017a) and were hence excluded from this study. The table shows that the three MMP constellations are neither equally frequent, nor equally distributed across speakers, which should be kept in mind when interpreting the results.

MMP	Katarina (f)	Sofia (f)	Alexander (m)	Filip (m)	Pelle (m)	Total
BA	60	36	60	75	49	280
BA + HB	29	36	71	12	34	182
BA + HB + EB	16	6	34	2	23	81
Total	105	78	165	89	106	543

Table 2: Sample size per multimodal prominence constellation (MMP) and speaker; f = female, m = male.

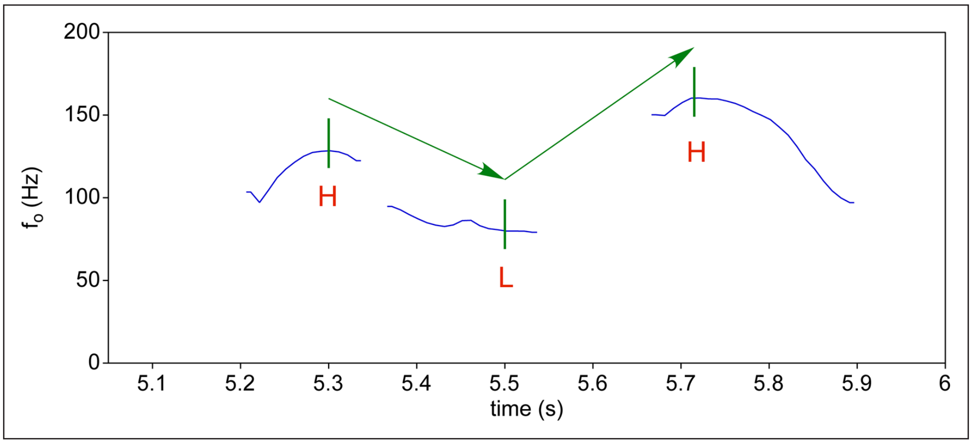

Figure 2: Schematic illustration of the f_o landmarks labelled. Example from the present material. Only f_o is shown in this illustration, while in the actual labelling process in Praat, annotators had access to waveform, spectrogram, word segmentations, and orthographic transcriptions (see Figure B1 in Appendix B). The arrows represent the dependent variables (the fall and the rise) that were calculated from the labelled landmarks.

2.2.5. Annotation of f_o landmarks

Returning to the labelling of f_o landmarks, for each word labelled as BA, three f_o landmarks were determined: an initial f_o maximum (representing the H-tone of the HL-word accent pattern), the following f_o minimum (L), and the following f_o maximum representing the big-accent H-tone (cf. Section 1.2). **Figure 2** displays an example f_o contour from the present material with these landmarks labelled. Details concerning the annotation of f_o landmarks, as well as concerning data extraction and additional tagging, are provided in Appendix B.6 and B.7.

2.3. Measurements and data analysis

The data file resulting from data extraction and additional tagging (see Appendix B.7) was used to calculate two dependent variables and perform statistical analyses in R (R Core Team, 2012). Data file and R-code are included in the supplementary materials (Appendices C and D). The two variables of interest were the accentual fall (henceforth, fall) and the big-accent rise (rise). These were calculated in semitones, based on the f_o values extracted at the labeled f_o landmarks. Due to missing values for the initial H-landmark (see Appendix B.6) in 125 cases (23% of 543 words), the subsequent analysis is based on 418 data points for the fall, and 543 data points for the rise.

The data were first explored using boxplots, means, and standard deviations. Our goal was to investigate the role of the three-level MMP factor (levels: BA, BA + HB, BA + HB + EB) as a predictor for the size (in semitones) of the fall and the rise variables, as well as its interaction with lexical prosody. As our data set contains data from several speakers, while each speaker contributes with many data points (and hence independence of data points cannot be assumed), we have opted to assess the role of MMP using linear mixed-effects models. Our rationale for evaluating the effect of

MMP as a predictor is to (i) build a full model, including MMP and its hypothetic interactions, as well as all other potentially relevant and available fixed- or random-effects factors, in order to (ii) compare it to simpler models, lacking MMP as a predictor or its interaction with other predictors. Models are then evaluated by means of R^2 values as an effect size measure (reflecting the degree of variability in the data described by the model). In addition, models are compared using likelihood ratio tests (χ^2) in order to obtain *p*-values, testing whether the contribution of MMP as a predictor is significant (i.e., whether a model containing MMP [or an interaction involving MMP] performs differently from a simpler model). This procedure was pursued for the fall and the rise. All modeling was done in R using the lmer function from the lme4 package (Bates, Maechler, & Bolker, 2012). R^2-values were obtained using the function r.squaredGLMM from the MuMIn package (Barton, 2020), χ^2-tests using the anova function from the stats package (R Core Team, 2012). Further details on decisions in model construction are presented in Section 3.2.

3. Results

3.1. Data exploration

In describing the trends seen in the results in this section, we refer to the results of the statistical modeling, which is presented in detail in Section 3.2. **Figure 3** displays boxplots for the accentual fall and the big-accent rise as a function of accompanying head and eyebrow movements, pooled across all five speakers. To gain an overall impression of the predictive value of the multimodal prominence constellation factor (MMP), the data are pooled across lexical-prosodic conditions. **Table 3** provides corresponding means and standard deviations. In **Figure 4** and **Table 4**, however, the results are shown separately for the three lexical-prosodic categories (henceforth, LexPros).

The boxplots (**Figures 3** and **4**) and the standard deviations (**Tables 3** and **4**) show a large variability of the size of both fall and rise, both ranging from <1 *st* to about 15 *st*. The standard deviations are large (≈ 3 *st* for most conditions) relative to the means. Nevertheless, some trends are noticeable: **Figure 3b** and **Table 3** suggest a tendency for larger rises when visible beats are coproduced with the big accent, as the size of the rise tends to be slightly larger for BA + HB (with head beat) than for BA (without head beat), and even slightly larger again when an eyebrow movement is added (BA + HB + EB). This trend is significant for the rise (i.e., the contribution of MMP as a predictor is significant), but there is no such trend seen for the fall (see **Figure 3**, **Table 3**, and also **Table 6** below).

In **Figure 4b**, we can observe a significant interaction of the predictors MMP and LexPros (cf. also **Table 6**), as the correlation between MMP and the size of the rise differs largely between the prosodic word types. In particular, a straightforward tendency for larger rises as a function of added visual beats is best observable for A1 words (**Figure 4b**; **Table 4**): The mean rise size is larger for BA + HB than for BA, and even larger for BA + HB + EB. However, this tendency is only partial, and complementary for the simplex A2 versus compound A2 words: For the simplex A2 words, the

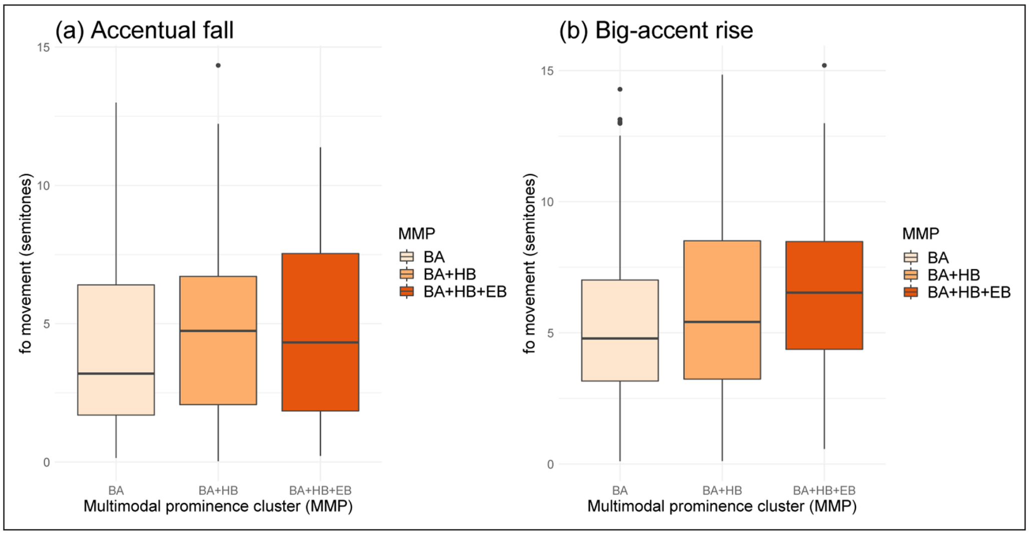

Figure 3: Boxplots[3] for the accentual fall **(a)** and the following big-accent rise **(b)** measured in semitones [*st*] as a function of the multimodal prominence constellation (MMP); sample sizes, fall (a): $n_{BA} = 224$, $n_{BA+HB} = 135$, $n_{BA+HB+EB} = 59$, rise (b): $n_{BA} = 280$, $n_{BA+HB} = 182$, $n_{BA+HB+EB} = 81$.

	Accentual fall [*st*]	**Big-accent rise [*st*]**
BA	4.07 (2.90)	5.24 (2.91)
	$n = 224$	$n = 280$
BA + HB	4.74 (2.99)	6.04 (3.43)
	$n = 135$	$n = 182$
BA + HB + EB	4.76 (3.13)	6.42 (3.02)
	$n = 59$	$n = 81$

Table 3: Means of f_o range in semitones [*st*] (standard deviations in parentheses) and sample sizes (*n*) for the fall and the rise per multimodal prominence constellation (BA, BA + HB, BA + HB + EB). Note that the sample of falls contains 125 fewer data points than the sample of rises, since a fall is often not measurable in A1 words.

big-accent rise tends to be slightly enlarged only when both a head and an eyebrow beat are added (**Figure 4b**). For the compound words, however, larger rises are obtained already when a head beat is added (BA + HB), while no further extension of the rise is observed with an additional eyebrow movement. For the fall, **Figure 4a** suggests a trend for a predictive function of MMP for simplex A2 words (which is not significant), but not for the other two word types. However, there is no significant interaction between MMP and LexPros for the fall (**Figure 4a**; **Table 6**).

[3] The upper and lower ends of the box mark the first and the third quartile (the box thus contains 50% of the data points around the median), the horizontal line within the box marks the median, the whiskers mark the highest and the lowest data point still within 1.5 times the inter-quartile range, and the dots represent extreme values.

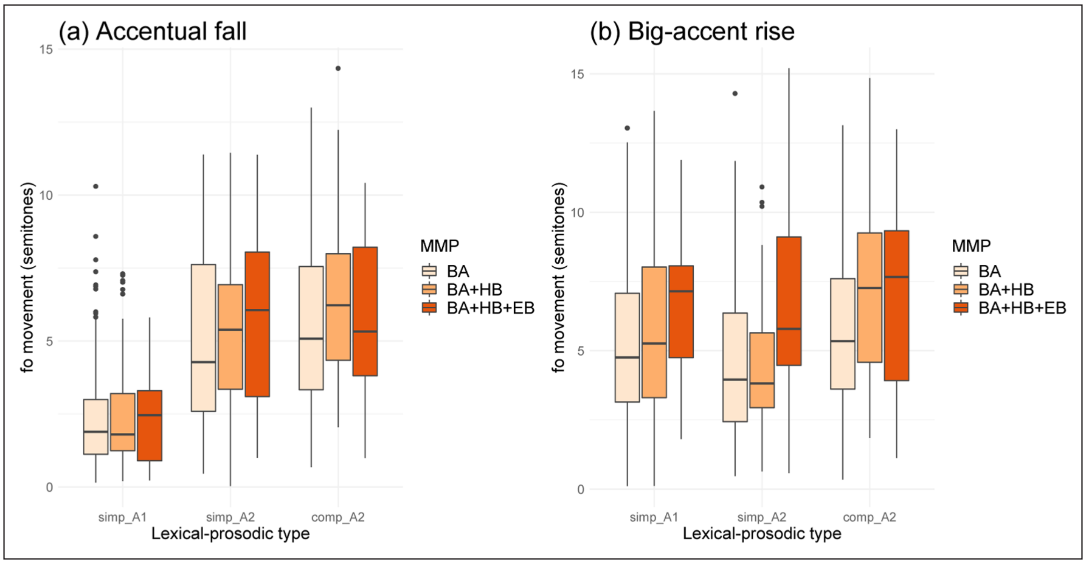

Figure 4: Boxplots for the accentual fall **(a)** and the following big-accent rise **(b)** measured in semitones as a function of the multimodal prominence constellation (MMP) and the lexical-prosodic condition (x-axis: simplex stress [A1], simplex stress + lexical tone [A2], compound stress [A2]); for sample sizes see Table 4.

The boxplots in **Figure 4** display some further trends: The fall clearly tends to be smaller in A1 words than in A2 words (**Figure 4a**). For the rise, however, **Figure 4b** does not suggest such a tendency. It suggests nonetheless a slight tendency for larger rises in compounds than in simplex words: Compare the medians for comp_A2 and simpl_A1 in **Figure 4b** (see also per word type totals in **Table 4**). Both trends (for the fall and the rise, respectively) are backed-up by the modeling, as the contribution of LexPros turns out significant for both the fall and the rise (**Table 6**).

As a final step in this data exploration, let us consider speaker variability. **Figure 5** displays general trends with respect to the occurrence of visual beats, subsuming the lexical-prosodic conditions, comparable to **Figure 3** above. (Additional speaker-specific plots for the rise are shown in Figure E1 in Appendix E, split up by lexical-prosodic category as in **Figure 4**.) Comparing **Figure 5** to **Figure 3** suggests that the overall trend—larger rises when head or eyebrow movements are added—is generally evident for the individual speakers, with partial exceptions for speakers Katarina and Filip. In the case of Filip, the deviant result for BA + HB + EB may be related to the low number of data points (see **Table 2**).

3.2. Statistical modelling

The role of MMP and LexPros as predictors for our fall and rise variables were assessed using linear mixed-effects regression models (see Section 2.3). Starting with the structure of our full models, we included MMP (3 levels: BA, BA + HB, BA + HB + EB), LexPros (3 levels: simp_A1, simp_A2, comp_A2), as well as the speakers' Sex (2 levels) as fixed-effect factors. Sex was included even

	Simplex, A1		Simplex, A2		Compound, A2	
	Accentual fall [*st*]	**Big-accent rise [*st*]**	**Accentual fall [*st*]**	**Big-accent rise [*st*]**	**Accentual fall [*st*]**	**Big-accent rise [*st*]**
BA	2.46 (2.06)	5.13 (2.70)	5.03 (3.08)	4.69 (3.24)	5.40 (2.73)	5.83 (3.01)
	n = 95	*n* = 150	*n* = 52	*n* = 52	*n* = 77	*n* = 78
BA + HB	2.66 (2.10)	5.85 (3.40)	5.26 (2.71)	4.61 (2.64)	6.37 (2.74)	7.39 (3.56)
	n = 49	*n* = 95	*n* = 35	*n* = 36	*n* = 51	*n* = 51
BA + HB + EB	2.42 (1.71)	6.30 (2.58)	5.78 (3.18)	6.29 (3.43)	5.62 (3.05)	6.79 (3.43)
	n = 17	*n* = 39	*n* = 22	*n* = 22	*n* = 20	*n* = 20
Total per word type	2.52 (2.03)	5.53 (2.96)	5.26 (2.97)	4.98 (3.14)	5.77 (2.80)	6.49 (3.32)
	n = 161	*n* = 284	*n* = 109	*n* = 110	*n* = 148	*n* = 149

Table 4: Means of f_o range in semitones [*st*] (standard deviations in parentheses) and sample sizes (*n*) for the fall and rise per multimodal prominence constellation (BA, BA + HB, BA + HB + EB) and lexical-prosodic structure (simplex versus compound; A1 versus A2).

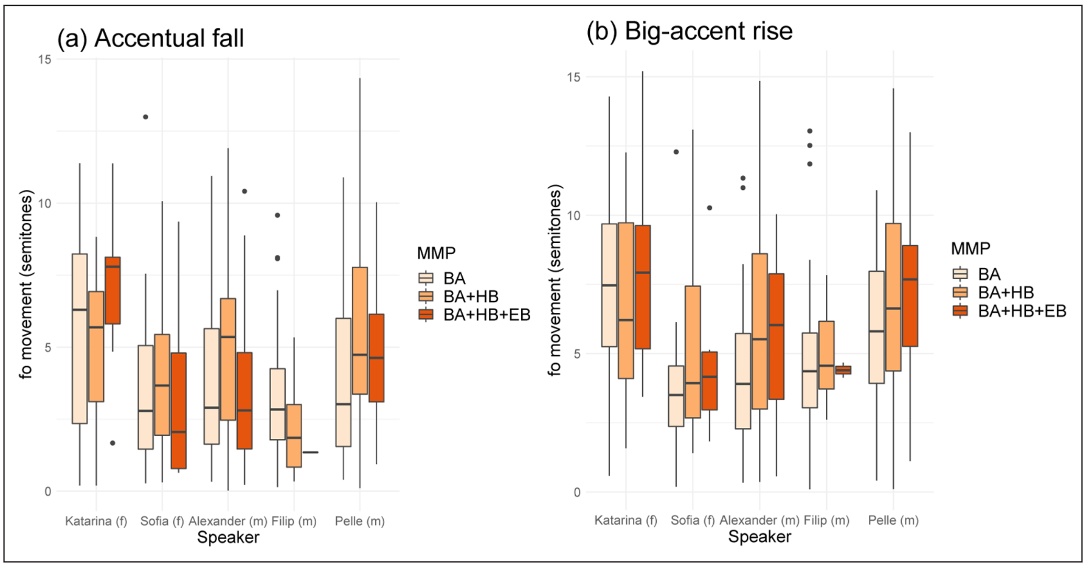

Figure 5: Boxplots for the accentual fall **(a)** and the following big-accent rise **(b)** measured in semitones as a function of the multimodal prominence constellation (MMP). The figure displays the individual results for the five speakers included in this study (two female = f, three male = m); for sample sizes see Table 2.

though our f_o measures are expressed in semitones, which should largely eliminate sex-related f_o differences (cf. **Figure 5** suggesting individual speaker, rather than sex-related variations). Speaker sex might nonetheless still account for some degree of data variability (see Figure E2 in Appendix E) and was hence included. Crucially, we were also interested in possible interactions between LexPros and MMP, hence the interaction term ('*') in **Table 5**.

To the three fixed-effect factors (MMP * LexPros + Sex) we added two random-effects factors: obviously, Speaker, but also Topic (see Appendix B.7), as a rather large variation between some topics can be observed both for the accentual fall and the rise (see Figure E3 in Appendix E). However, we allowed the model to assume varying intercepts, but no varying slopes, neither for different Speakers nor Topics. Random slopes for Speaker could have been considered given the results shown in **Figure 5**, as the contribution of MMP appears to differ in strength (and partly takes different directions for two speakers). However, we preliminarily tested a model with random slopes for Speaker, which ran into a convergence issue. We hence opted for random intercepts only, thereby accepting a potentially worse model fit, but on the other hand a simpler, easier interpretable, model. The resulting full models (for the rise and fall) are displayed in R syntax in **Table 5**.

Based on the full models, we constructed three sets of reduced models: one set lacking the interaction between MMP and LexPros ("Reduced 1" in **Table 5**), one set lacking the MMP factor altogether, but keeping Lexpros ("Reduced 2"), and one set lacking the LexPros factor, while keeping MMP ("Reduced 3").

Visual inspection of residual plots (Figure E4 in Appendix E) for the rise did not reveal any obvious deviations from homoscedasticity or normality. For the fall, the plots suggest that homoscedasticity might be slightly violated, meaning that the model for the fall should be interpreted with some caution.

Table 5 displays R^2-values of all six models. For each model, two different R^2-values are presented. The (marginal) R^2_m value measures the variation described by the fixed factors, while the (conditional) R^2_c measures the variation described by the entire model, including the random-effects factors (Nakagawa & Schielzeth, 2013). **Table 5** shows that the full models provide the

Dependent variable	**Model**		R^2_m	R^2_c
Fall	Full	MMP * LexPros + Sex + (1|Speaker) + (1|Topic)	0.284	0.350
	Reduced 1	MMP + LexPros + Sex + (1|Speaker) + (1|Topic)	0.278	0.342
	Reduced 2	LexPros + Sex + (1|Speaker) + (1|Topic)	0.274	0.338
	Reduced 3	MMP + Sex + (1|Speaker) + (1|Topic)	0.022	0.076
Rise	Full	MMP * LexPros + Sex + (1|Speaker) + (1|Topic)	0.080	0.276
	Reduced 1	MMP + LexPros + Sex + (1|Speaker) + (1|Topic)	0.064	0.253
	Reduced 2	LexPros + Sex + (1|Speaker) + (1|Topic)	0.044	0.230
	Reduced 3	MMP + Sex + (1|Speaker) + (1|Topic)	0.023	0.200

Table 5: 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 factors; '*' denotes the interaction between factors, as opposed to '+.'

best fit, as they reach the highest R^2-values. The models account for 35% of the variation observed in the falls ($R^2_c = 0.350$) and for 27.6% in the case of the rises.

The table also displays a salient difference between the fall and the rise concerning the contribution of the fixed factors. For the falls, R^2_m suggests that the fixed factors contribute relatively strongly to overall model performance (28.4% of variation described) and that this contribution can be largely explained by the LexPros factor (cf. the low R^2_m for "Reduced 3," lacking LexPros, compared to all other models for the fall). In fact, only the contribution of LexPros, but not of MMP, proves significant in the likelihood ratio tests (cf. **Table 6**). For the rise, on the other hand, the fixed factors contribute overall relatively little to model performance (at most 8% of variation described by the full model). However, not only LexPros, but also MMP and their interaction appear to contribute to the model, as the removal of any of these factors drastically reduces the R^2_m value.

This conclusion is supported by the results from the likelihood ratio tests presented in **Table 6**. Three comparisons were performed for each of the two dependent variables (fall, rise). The first comparison (Full versus Reduced 1) evaluates the interaction between MMP and LexPros. The table reveals a significant interaction for the rise, but not for the fall. The second and third comparisons (Reduced 1 versus 2; Reduced 1 versus 3) evaluate the predictive values of MMP and LexPros, respectively. This is done by comparing a reduced model (lacking the factor of interest) with the more complex model (including the factor, but no interaction). The results show that LexPros makes a significant contribution to both rise and fall, while the contribution of MMP is only significant for the rise.

To sum up, the results suggest that the size of both the accentual fall and the big-accent rise are predicted in a significant manner by the lexical prosodic category. In addition, they suggest a slight but significant contribution of the MMP-factor as a predictor for the size of the big-accent rise, but not for the accentual fall. Likewise, there is an interaction between MMP and lexical prosody on the rise, meaning that MMP predicts the rise differently in the different word types.

Dependent variable	**Model comparison**	**Effect tested**	χ^2	*df*	*p*
Fall	Full vs. Reduced 1	Interaction (MMP*Lexpros)	3.63	4	0.458
	Reduced 1 vs. Reduced 2	MMP	2.82	2	0.245
	Reduced 1 vs. Reduced 3	LexPros	135.54	2	<.001***
Rise	Full vs. Reduced 1	Interaction (MMP*Lexpros)	11.26	4	.024*
	Reduced 1 vs. Reduced 2	MMP	14.08	2	<.001***
	Reduced 1 vs. Reduced 3	LexPros	28.83	2	<.001***

Table 6: Results of likelihood ratio tests comparing full and reduced models.

4. Discussion

The results of this study provide partial, yet significant evidence in favor of the hypothesis of a cumulative relation between acoustic and kinematic correlates of prominence. However, this evidence heavily depends on the prosodic-phonological characteristics of the accented word and on the part of the accentual pattern measured (fall versus rise in the two-peaked pitch accents of Swedish). In particular, a significant trend for a cumulative relation between the realization of the pitch accent and accompanying head or eyebrow beats was only observed for the big-accent rise, but not for the accentual fall. Furthermore, the cumulative relation takes different forms in different types of words: The rise of a big accent tends to be greater when accompanied by head and/or eyebrow gestures than when produced without any beat. However, prosodic word types differ in respect to how straightforward the number of beat gestures (head, or head and eyebrows) predicts a further increase, as discussed in more detail in Section 4.1. In Section 4.2, we discuss the cumulative relation from a more general perspective (disregarding the role of lexical prosody), and then conclude with an outlook in Section 4.3.

4.1. Effects of lexical prosody

This study enables us to evaluate the role of several prosodic-phonological features in the multimodal production of prominence. In particular, we are able to discuss:

1. whether the hypothesized cumulative relation between beat gestures and pitch accent realization would affect both the accentual fall and the big-accent rise,
2. whether it affects the fall in both A1 (where it is realized before the stressed syllable) and in A2 (realized in the stressed syllable),
3. whether it affects such f_o movements that originate from lexical tones (i.e., the fall in simplex A2 versus in compound A2 words), and
4. whether it would affect the prominence-encoding H tone (the big accent rise) irrespective of it being associated with the primary stress (A1), the secondary stress (compound A2), or not with a stressed syllable at all (simplex A2).

The answer to the first question suggested by our results is ‘no,’ since a significant contribution of the predictor MMP (i.e., the addition of head and eyebrow beats) was found only for the rise, but not for the fall. For the same reason, our results cannot confidently answer question 2. If we focus our attention on the A2 words only (simplex and compounds in **Figure 4a**), the range of the fall seems to be slightly predictable by MMP in a comparable manner as the rise in A1 words, but the trend did not prove significant. This lack of significance might be related to the strongly deviating results for A1 words (to be further discussed in Section 4.1.2) and hence the fact that the significant contribution of the LexPros factor accounted for the bigger part of the variability in the data (**Tables 5** and **6**).[4] However, it may also be true that the accentual fall is not at all, or only weakly, involved in the cumulative relation between beat gestures and pitch accent realization. This in turn may indicate that the accentual fall is more related to lexical stress and less to the encoding of post-lexical prominence.[5] Concerning question 3, our results do not reveal any difference between the results for the fall in simplex versus compound A2 words, and hence they do not provide any evidence suggesting that the lexical status of tone would have any impact on the multimodal implementation of prominence. Question 4, concerning the rise, deserves a more detailed discussion which is presented in Section 4.1.1.

4.1.1. The big-accent rise

The results for the rise revealed a significant contribution of the MMP factor (i.e., accompanying beat gestures) as well as an interaction with lexical prosody, whereby all lexical-prosodic categories seem to respond differently to the presence of accompanying gestures. In A1 words, the cumulative relation is observed most straightforwardly: The rise tends to be larger when a head beat is present than when not (BA + HB versus BA), and yet larger when also an eyebrow beat is present (BA + HB + EB). In simplex A2 words, however, no modification of the rise is observable when only a head beat is added, but only if also an eyebrow beat is present. For compound A2 words, finally, this relation is reversed: The rise tends to be enlarged already when a head beat is added (as in A1 words), but not any further with an additional eyebrow movement.

An explanation for the results for the compounds may perhaps be developed along the following lines. Compound words are semantically more complex and might, at least in the genre of news presentations, simply be subject to stronger overall emphasis than simplex words. This assumption is supported by the generally larger big-accent rise in compounds as observed in the data (cf. **Table 4**). Thus, when a head beat is added in compounds, a certain threshold for f_o might already have been reached, and therefore, for a further increase of prominence, eyebrow movements may be added, but without any further expansion of the f_o fall and rise. An independent measure of prominence (i.e., perceptual prominence ratings) would be helpful in order to evaluate this explanation (e.g., Ambrazaitis et al., 2020a; Ambrazaitis, Frid, & House, 2022).

[4] That is, the inclusion of fall measurements for A1 words might well have caused an analytical artefact: Figure 4a suggests that the low values obtained for the fall in A1 may have caused the significant contribution of LexPros as a predictor, and possibly have prevented a significant contribution of MMP, and its interaction with LexPros. However, when replicating the models and model comparisons for the fall with a reduced data set, including only the A2 words, still no significant contribution of MMP can be established [$\chi^2(2) = 2.45, p = .29$], but only, again, a significant contribution of LexPros [$\chi^2(1) = 4.19, p = .04$].

[5] As the tones associate with stress, they only surface in words that are stressed. In a sense, thus, they can also be characterized as tonal correlates of lexical stress. Words in Swedish may surface unstressed—and hence also unaccented—which is often the case for function words (e.g., prepositions, conjunctions, often also pronouns or parts of lexicalized phrases; see, e.g., Myrberg, 2010; Myrberg & Riad, 2015).

This tentative attempt of explaining the patterns observed in compounds, however, does not help us in understanding the reversed pattern observed in simplex A2 words, where larger values for the rise are only observable in connection with an eyebrow beat, but not when only a head beat is added (**Figure 4b**). In this case, prosodic-phonological characteristics might well provide an explanation: It would seem reasonable to assume that a spatial/acoustical convergence of gesture and speech generally presupposes a certain temporal convergence. However, in simplex A2 words, the big-accent rise is not realized in a lexically stressed syllable, but typically in the (unstressed) post-stress syllable. The head movement, though, might well be aligned with the stressed syllable (where the accentual fall is realized in A2 words) (e.g., Alexanderson et al., 2013; Esteve-Gibert et al., 2017; *inter alia*). This lack of alignment between head movement and pitch rise might, thus, lie behind the lack of correlation between the presence of head beat and the size of pitch rise in simplex A2 words. However, this explanation does not directly account for the fact that we still observe a trend for larger pitch rises in these words when also eyebrow beats are added. We could speculate, though, that temporal convergence is a prerequisite for spatial/acoustical convergence only in the case of head beats, but not in the case of eyebrow beats. It has been observed that eyebrow beats tend to slightly precede the head beat they co-occur with (informal observation in this study, see Appendix B.5, but also House, Ambrazaitis, Alexanderson, Ewald, & Kelterer, 2017, and Flecha-García, 2010, who showed that eyebrow movements precede pitch accents). That is, it is possible that eyebrow beats generally do not display a direct temporal association with the pitch accent, but rather with the head beat, and therefore, their spatial/acoustical convergence with pitch accents is not inhibited by a lack of temporal convergence.

Note that we presently cannot test these assumptions, as our annotation scheme does not reveal temporal detail. However, we have concluded in previous studies that head beats seem to require a lexically stressed syllable to associate with (be it a primary or a secondary stress; cf. Ambrazaitis & House, 2017a; Kelterer et al., 2018). In addition, previous research has shown that head movements temporally align with stressed syllables (e.g., Alexanderson et al., 2013; Esteve-Gibert et al., 2017; *inter alia*). Under the assumption that these association and alignment patterns are found in the present data, too, we would expect a cumulative relation between the occurrence of head beats and f_o excursion for the rise in A1 words (where the rise occurs in the primary stress) and possibly in A2 compounds (where the rise occurs in the secondary stress), but not for the rise in A2 simplex words, which is what our results display.

4.1.2. A note on the accentual fall in A1

The results revealed a significant contribution of lexical prosody (i.e., the LexPros factor) on the realization of the fall, with generally lower values for the fall in A1 than for A2 words. This result replicates what is well known about the word accents' phonetic realization: The fall in the case of A1 precedes the stressed syllable (and typically even the word, as many Swedish words have initial stress) and is hence probably much less perceptually relevant than the upcoming big-accent rise. It is thus more likely subject to reduction (Engstrand, 1997).

Yet, in our data, the fall was measurable in many cases, typically reaching a range of several semitones, although considerably smaller than in A2. However, the status of the fall in A1 is unsettled (Bruce, 1977; Engstrand, 1997), and hence we also need to consider the possibility that the fall we have measured as the accentual fall in A1 is actually not related to the accented word at all. Bruce (1977) argued convincingly for the existence and relevance of an accentual fall in A1 in small accents (i.e., when the sentence accent is missing, in Bruce's terms), but it might well be more adequate to assume its elision in big accents (e.g., Engstrand, 1997; Myrberg & Riad, 2015). That is, what we have measured as fall in A1 might represent an invalid measure in the present context; we included it because it is assumed to exist in Bruce's account.

4.2. Prominence production: Essentially multimodal and cumulative?

The conceptualization of prominence production as an essentially multimodal process is by now well established (e.g., Yasinnik et al., 2004; Beskow et al., 2006; Swerts & Krahmer, 2010; Parrell et al., 2014; Ambrazaitis & House, 2017a; Esteve-Gibert et al., 2017; Krivokapić et al., 2017; *inter alia*). A cumulative relation between speech-related and gesture-related signals is, we would argue, well in line with this multimodal conception, but not equally well established yet, as previous evidence in favor of the cumulative-cue hypothesis is limited (see Section 1.1), and the evidence from the present study is rather weak, too. Despite this lack of strong evidence, we would argue that it makes sense to assume that prominence production is essentially multimodal and cumulative in nature, meaning that, *ceteris paribus*, the more/stronger the prominence-lending gesture, the stronger the speech cues.

It would make sense to assume this cumulative relation as an essential feature of prominence production, partly because some existing evidence clearly supports this idea (e.g., Parrell et al., 2014; Krivokapić et al., 2017). It also makes sense, however, because a cumulative relation has been repeatedly attested even when only considering the acoustic domain: For instance, pitch-accented words usually not only stand out by virtue of a pronounced inflection in f_o, but also in terms of longer segmental durations and certain spectral characteristics (e.g., Cooper, Eady, & Mueller, 1985; Sluijter & van Heuven, 1996; Fant et al., 2000; Heldner, 2003; *inter alia*). These acoustic properties of prominence were also the basis for Gussenhoven's proposal of the effort code (Gussenhoven, 2004), as an attempt to explain why prominence is generally expressed through a strengthening of the acoustic signal. The effort code can, we suggest, likewise account for gestural signals, and would, if extended to a multimodal concept, be well in line with the cumulative-cue hypothesis.

However, the evidence for a multimodal, cumulative relation is still not strong and not entirely conclusive, and we propose that there might be several explanations for this. First, different mechanisms might lie behind a cumulative relation, which might all come into play to various degrees. One mechanism might be located at the neural level: A bulk of studies suggests that speech and gesture are planned in conjunction (e.g., McNeill, 1985, 1992, 2005; Kendon, 2004; Esteve-Gibert & Prieto, 2013; Krivokapić, 2014; Graziano & Gullberg, 2018; *inter alia*). Admittedly, a co-generation of gestural and speech prominence cues at the planning stage would not seem to necessarily require a cumulative relation, but, we suggest, it would be the simplest possible mechanism, well in line with the effort code. However, at the planning stage, various constraints (e.g., phonological) might possibly interfere, which was one motivation for testing a possible role of lexical prosody in the present study.

Another mechanism behind the cumulative relation most likely lies in the (common) motor system, since the expression of prominence in one modality would seem to be able to trigger activation of the other modality (e.g., Parrell et al., 2014; see also Kelso, Tuller, & Harris, 1983). Finally, a different type of connection between gesture and speech production is suggested by findings on larger movements of the upper limbs, which have been shown to affect speech acoustics biomechanically, through the force that the movement transfers "onto the musculoskeletal system, thereby modulating respiration-related muscle activity, leading to changes in the intensity of vocalization" (Pouw et al., 2021, p. 90; see also Pouw et al., 2020a, 2020b). However, to the best of our knowledge we lack evidence for this kind of causal relationship when it comes to less forceful movements, such as those produced with the head or the eyebrows. To round off the argument, the weak evidence for the cumulative relation might be explained with reference to the various (parallel) mechanisms behind the relation, which might possibly come into play to different degrees. In particular, the last-mentioned mechanism (biomechanical force) will only have an effect when certain types of movements are involved; also, it is possible that the cross-modal connection in the motor system is a relatively weak force; and finally, the co-generation process at the neural level is, as already mentioned, subject to interaction with various non-prominence related constraints.

This interaction of constraints thus possibly involves an important source of variability in itself, which can easily lead to a weakening or a complete cancellation of the cumulative relation at the surface. In other words, even if prominence production may be determined by a multimodal effort code in the first place, both gesture and speech are also heavily shaped by other acoustic and kinematic form-function relations, which are encoded in parallel (cf. Liu & Xu, 2005; Xu, 2005, for the acoustic channel). Both speech and gestures are multifunctional. Even if we have referred to our head and eyebrow movements as beat gestures, a gesture can function as a beat while fulfilling other, say, deictic or iconic functions at the same time (e.g., Prieto, Cravotta, Kushch, Rohrer, & Vilà-Giménez, 2018; Shattuck-Hufnagel & Prieto, 2019). This might in general be most obvious for manual gestures, but it is probably valid even for head and eyebrow gestures.

Seemingly at odds with the cumulative-cue hypothesis are, at first sight, results by Prieto et al. (2015) who showed that, in the perception of prominence, beat gestures and pitch accents can compensate for each other, suggesting a cue-trading, rather than a cumulative-cue relation (cf. Ambrazaitis & House, 2017b). However, this cue-trading effect in perception can be accounted for, we propose, even if a cumulative relation might underlie the production of prominence, by referring to the multifunctionality of the kinematic and the acoustic speech channels just discussed. This multifunctionality can be expected to create a high degree of variation when it comes to the encoding of prominence at the surface. A particular prominence level might thus, when gesturing is not available for the encoding of prominence, be encoded primarily through acoustic parameters, while in another setting, it may be achieved through a gestural beat combined with weaker acoustic cues. Speech perceivers are, of course, accustomed to this variation in the production of prominence, which could very well lie behind the cue-trading between auditory and visual prominence cues observed by Prieto et al. (2015).

4.3. Conclusions and outlook

This study has provided novel evidence in favor of a cumulative-cue hypothesis on the multimodal encoding of prominence, suggesting in particular that the acoustic realization of pitch accents (here: f_o range of accentual rise) co-varies with the number of accompanying gestures (head beat only, or head beat plus eyebrow beat), and that this cumulative relation might be to some degree sensitive for lexical prosody. A cumulative relation, we have argued, is well in line with the assumption that "implementation of prosody is not domain-specific but relies on general aspects of the motor system" (Parrell et al., 2014, abstract), but has possibly multiple sources (Section 4.2). In a wider perspective, it further supports the idea of speech and gesture as an integrated system (e.g., McNeill, 1992; McNeill 2005; Iverson & Thelen, 1999; Kendon, 2004; Willems & Hagoort, 2007; Graziano & Gullberg, 2018).

Previous evidence for a cumulative relation in multimodal prominence encoding stems mostly from manual gestures elicited in less naturalistic experimental tasks and has to some degree been inconclusive. This evidence has typically been based on the comparison between a simple gesture versus a no gesture condition (i.e., a binary factor) or the spatial magnitude of gestures (the 'stronger' in our hypothesis), but studies have neglected the possible clustering of gestures (the 'more' in our hypothesis). The present study thus extends our understanding of the cumulative relation in several ways, as it was based on spontaneous gestures produced in naturalistic speech, focusing on head and eyebrow movements and how their clustering correlates with pitch accent realization.

Our results also suggest that the cumulative relation might be sensitive for the prosodic-phonological features of the accented word, as, for instance, a cumulative relation was only observed for the big-accent rise, but not for the accentual fall. Furthermore, the rise in simplex A2 words was not found to be enlarged in connection with an accompanying head beat. This is probably due to the rise not being aligned with the stressed syllable, while the head beat, presumably, is. However, to evaluate these explanations, we need to include precise temporal data on the alignment of movements and stressed syllables in a future study (see, e.g., Pouw et al., 2020c).

Another question to be addressed in the future is to what extent our head and eyebrow clusters indeed represent a 'more' of gestures, rather than a 'stronger.' Our previous study (Ambrazaitis & House, 2017a) has shown that, in our news speech material, eyebrow movements generally occur in connection with head movements; in only a very few cases did annotators see an EB where there was no simultaneous (or timely following) HB, which was also the reason for testing an MMP factor (BA, BA + HB, BA + HB + EB) in the present study. Possibly, then, an eyebrow beat should not be classified as an additional gesture, but rather as means to strengthen a head beat (but see also Swerts & Krahmer, 2010, where HB and EB occurred independently of each other). Future research should therefore address the relation between types of gestures that are perhaps more independent of each other, such as head beats and manual beats, and study how such gestures and their combinations correlate with pitch accent acoustics.

This leads us to two further directions for future research: For one, further acoustic dimensions (beyond f_0) should be included (such as segmental durations and spectral characteristics), as well as continuous kinematic data collected using, for instance, motion capture (e.g., Krivokapić et al., 2017; Pouw et al., 2021), electromagnetic articulography (e.g., Parrell et al., 2014; Krivokapić et al., 2017; Frid, Svensson Lundmark, Ambrazaitis, Schötz, & House, 2019), or video recognition techniques (e.g., Pouw et al., 2020c).

Furthermore, a variety of speech tasks and genres (including controlled experimental data and spontaneous conversation) should be considered in order both to enable the study of various types of gestures and to gradually establish a valid and comprehensive picture of gesture-speech integration in prominence production. Studying news readings has its advantages, when the focus is on prominence, as it would seem probable that the gestures produced by news presenters are less likely to be affected by many other communicative functions, compared to gestures produced in, say, spontaneous speech. News presenters are expected to move only minimally and to try to avoid the expression of emotions (at least on Swedish public service television). Obviously, they do not need to produce any dialog-regulating signals, which are known to be expressed both through speech and gesture. However, our data show that eyebrow movements and first and foremost head movements nevertheless are rather frequent. This is *per se* another piece of evidence for the emerging insight that gesture is an essential ingredient of spoken language.

It would seem reasonable to assume that the use of head and eyebrow movements is more or less restricted to prominence signaling in this genre. Co-speech head movements have been shown to have a variety of semantic and discourse functions (McClave, 2000), most of which would not seem to appear in news presentations. Likewise, eyebrow movements can signal discourse structure in dialogue (Flecha-García, 2010), but again, this function is less relevant in news presentations. That said, news presentations (at least from Swedish public service television) practically lack manual gestures, not least because presenters use their hands to hold a paper version of the script (which they, nevertheless, normally read from the prompter; the paper copy is used mostly as a backup).[6] Thus, although news presentations are perhaps advantageous for studying prominence encoding in some respect, we also need to include data from other sources.

In connection with the choice of materials, we would like to remind of the multifunctionality of gesture (and speech), which, we have argued above, may mask the hypothetically underlying cumulative relation of multimodal prominence cues at the surface. The material used in this study was probably 'clean' and 'noisy' at the same time: It was 'clean' in the sense that the expression of emotions and discourse-regulating signals was avoided. It was 'noisy' in the sense that, compared to controlled laboratory speech, our data were fairly uncontrolled, in that they contained different topics of news stories and different prosodic contexts (e.g., related to position in the utterance). Apart from controlling for individual differences between speakers, and to some degree effects of the topic (random effects in our linear mixed models), we did not include any further control, by means of, for instance, limiting the selection of data points to certain prosodic positions or the like, in order to avoid a critical diminishment of the sample. However, despite this noise, we observed a weak, yet significant, tendency for a positive correlation between the realization of pitch accents and the number of accompanying visual beats. It could be expected to find stronger correlations in future research when adding further levels of control, for instance by excluding certain prosodic contexts. This could seem even more necessary when turning to spontaneous speech, where a larger interference of various communicative functions, which need to be encoded multimodally and in parallel, can be expected.

Finally, we believe that for a reliable evaluation of the cumulative-cue hypothesis, future studies should also include perceptual prominence ratings (see Ambrazaitis et al., 2020a; Ambrazaitis et al., 2022) as an independent reference measure of prominence, to which correlations between acoustic and kinematic measures need to relate.

[6] Swedish news anchors are not explicitly trained for the specific task of presenting news on television. They are not explicitly taught how (not) to move their hands, head, or eyebrows. That is, the head and eyebrow beats observed in this study are not planned or rehearsed (Jane Andersson, SVT, personal communication).

Reproducibility

Data table and analysis code (R) for this study are included in the supplementary materials (Appendix C and D).

Additional files

The additional files for this article can be found as follows:

- **Appendix A.** A list of news stories included in the analysis. DOI: https://doi.org/10.16995/labphon.6430.s1
- **Appendix B.** A text presenting further details concerning methods and procedure not included in the article. DOI: https://doi.org/10.16995/labphon.6430.s2
- **Appendix C.** Data table. DOI: https://doi.org/10.16995/labphon.6430.s3
- **Appendix D.** R analysis code. DOI: https://doi.org/10.16995/labphon.6430.s4
- **Appendix E.** Some additional diagrams, such as detailed results per speaker and residual plots. DOI: https://doi.org/10.16995/labphon.6430.s5

Acknowledgements

Our research materials (television broadcasts) were retrieved 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 in that process, which also involved mediating between us and Swedish Television (SVT) to retrieve relevant permissions. A big thanks is also due to research assistants Malin Svensson Lundmark, Otto Ewald, Anneliese Kelterer, and Jonas Alexander David for preparing all data for annotation (Malin) and for annotations (Malin, Otto, Anneliese, Jonas), and to Dr. Sara Myrberg (Lund University) for discussions of Swedish lexical prosody. We are very grateful for the insightful and constructive comments by three anonymous reviewers. We also acknowledge feedback from delegates at the PaPE 2019 (3rd Phonetics and Phonology in Europe conference) in Lecce, Italy, as well as from members of the research network GEHM (GEstures and Head Movements in language), funded by the Independent Research Fund Denmark (9055-00004B). Finally, we deeply acknowledge Jane Andersson, news anchor at SVT, who took us to the studio for a study visit, providing us with essential information on the preparation and production of news broadcasts.

Funding information

This work was supported partly by the Marcus and Amalia Wallenberg Foundation (grant MAW 2012.01.03 to Gilbert Ambrazaitis) and partly by the Swedish Research Council (grant VR-2017-02140 to David House).

Competing interests

The authors have no competing interests to declare.

Author contributions

Both authors have made substantial contributions to the conception of the study, and both have actively participated in data acquisition (annotations), in collaboration with research assistants. Data analysis was performed by the first author. Both authors contributed significantly to the interpretation of the results. The manuscript was drafted by the first author and critically revised by the second author.

References

Alexanderson, S., House, D., & Beskow, J. (2013). Aspects of co-occurring syllables and head nods in spontaneous dialogue. In *Proceedings 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* (pp. 55–71). Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-18184-9_6

Ambrazaitis, G. (2009). *Nuclear intonation in Swedish: Evidence from experimental-phonetic studies and a comparison with German* (Doctoral dissertation). *Travaux de l'institut de linguistique de Lund, 49*. Centre for Languages and Literature, Lund University.

Ambrazaitis, G., Frid, J., & House, D. (2020a). Word prominence ratings in Swedish television news readings: Effects of pitch accents and head movements. In *Proceedings of the 10th International Conference on Speech Prosody* (pp. 314–318). Tokyo, Japan. DOI: https://doi.org/10.21437/SpeechProsody.2020-64

Ambrazaitis, G., Frid, J., & House, D. (2022). Auditory vs. audiovisual prominence ratings of speech involving spontaneously produced head movements. In *Proceedings of the 11th International Conference on Speech Prosody*. Lisbon, Portugal. DOI: https://doi.org/10.21437/SpeechProsody.2022-72

Ambrazaitis, G., & House, D. (2017a). 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. DOI: https://doi.org/10.1016/j.specom.2017.08.008

Ambrazaitis, G., & House, D. (2017b). Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization? In *Proceedings of the 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*. Stockholm, Sweden. DOI: https://doi.org/10.21437/AVSP.2017-17

Ambrazaitis, G., Zellers, M., & House, D. (2020b). Compounds in interaction: Patterns of synchronization between manual gestures and lexically stressed syllables in spontaneous Swedish. In *Proceedings 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.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38. DOI: https://doi.org/10.1016/j.wocn.2018.05.004

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Proceedings of Interspeech 2006* (pp. 1272–1275). Pittsburg, PA, USA. DOI: https://doi.org/10.21437/Interspeech.2006-375

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152. DOI: https://doi.org/10.1016/j.bandl.2012.10.008

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer*. Computer program. http://www.praat.org/

Bruce, G. (1977). *Swedish word accents in sentence perspective* (Doctoral dissertation). *Travaux de l'institut de linguistique de Lund, 12*. Lund University.

Bruce, G. (2007). Components of a prosodic typology of Swedish intonation. In T. Riad & C. Gussenhoven (Eds.), *Tones and tunes – volume 1: Typological studies in word and sentence prosody* (pp. 113–146). Berlin; New York: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110207569.113

Casasanto, D. (2013). Gesture and language processing. In H. Pashler (Ed.), *Encyclopedia of the mind* (pp. 372–374). Los Angeles; London; New Delhi; Singapore; Washington DC: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and *Psychological Measurement*, *20*, 37–46. DOI: https://doi.org/10.1177/001316446002000104

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142–2156. DOI: https://doi.org/10.1121/1.392372

Engstrand, O. (1997). Phonetic interpretation of the word accent contrast in Swedish: Evidence from spontaneous speech. *Phonetica*, *54*, 61–75. DOI: https://doi.org/10.1159/000262211

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. DOI: https://doi.org/10.1121/1.4986649

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. DOI: https://doi.org/10.1044/1092-4388(2012/12-0049)

Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, *57*, 301–316. DOI: https://doi.org/10.1016/j.specom.2013.06.006

Fant, G., Kruckenberg, A., & Liljencrants, J. (2000). Acoustic-phonetic analysis of prominence in Swedish. In A. Botinis (Ed.), *Intonation – Analysis, modelling and technology* (pp. 55–86). Dordrecht: Kluwer Academic Publishers. DOI: https://doi.org/10.1007/978-94-011-4317-2_3

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, *52*(6), 542–554. DOI: https://doi.org/10.1016/j.specom.2009.12.003

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. DOI: https://doi.org/10.1037/h0031619

Frid, J., Svensson Lundmark, M., Ambrazaitis, G., Schötz, S., & House, D. (2019). Investigating visual prosody using articulography. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference* (pp. 5–8). Copenhagen, Denmark.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84.1.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*, *9*, 879. DOI: https://doi.org/10.3389/fpsyg.2018.00879

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511616983

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, *31*(1), 39–62. DOI: https://doi.org/10.1016/S0095-4470(02)00071-2

House, D., Ambrazaitis, G., Alexanderson, S., Ewald, O., & Kelterer, A. (2017). Temporal organization of eyebrow beats, head beats and syllables in multimodal signaling of prominence. In *International Conference on Multimodal Communication: Developing New Theories and Methods*. Osnabrück, Germany.

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

Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, *6*(11–12), 19–40.

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. *Interdisciplinary Studies on Information Structure*, *3*, 199–244.

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. DOI: https://doi.org/10.1016/j.specom.2020.07.006

Kelso, J. A. S., Tuller, B., & Harris, K. (1983). A "dynamic pattern" perspective on the control and coordination of movement In P. MacNeilage (Ed.), *The production of speech* (pp. 138–173). New York: Springer. DOI: https://doi.org/10.1007/978-1-4613-8202-7_7

Kelterer, A., Ambrazaitis, G., & House, D. (2018). Head beats as pitch-accompanying visual correlates of primary and secondary lexical stress: Evidence from Stockholm Swedish compounds. In *Proceedings of the Sixth International Symposium on Tonal Aspects of Languages (TAL 2018)* (pp. 124–128). Berlin, Germany. DOI: https://doi.org/10.21437/TAL.2018-25

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511807572

Kleber, F., & Niebuhr, O. (2010). Semantic-context effects on lexical stress and syllable prominence. In *Proceedings of the 5th International Conference on Speech Prosody*. Chicago, USA.

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. DOI: https://doi.org/10.1016/j.jml.2007.06.005

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130397. DOI: https://doi.org/10.1098/rstb.2013.0397

Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 3. DOI: https://doi.org/10.5334/labphon.75

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471. DOI: https://doi.org/10.1080/01690965.2010.500218

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62*, 70–87. DOI: https://doi.org/10.1159/000090090

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, *3*(1), 71–89. DOI: https://doi.org/10.1515/lp-2012-0006

McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878. DOI: https://doi.org/10.1016/S0378-2166(99)00079-X

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, *92*(3), 350–371. DOI: https://doi.org/10.1037/0033-295X.92.3.350

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. DOI: https://doi.org/10.7208/chicago/9780226514642.001.0001

Myrberg, S. (2010). *The intonational phonology of Stockholm Swedish* (Doctoral dissertation). *ACTA Universitatis Stockholmiensis, 53: Stockholm Studies in Scandinavian Philology New Series*. Department of Scandinavian Languages, Stockholm University.

Myrberg, S., & Riad, T. (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, *38*(2), 115–147. DOI: https://doi.org/10.1017/S0332586515000177

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142. DOI: https://doi.org/10.1111/j.2041-210x.2012.00261.x

Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, *42*, 1–11. DOI: https://doi.org/10.1016/j.wocn.2013.11.002

Pouw, W., Harrison, S. J., & Dixon, J. A. (2020a). Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, *149*(2), 391–404. DOI: https://doi.org/10.1037/xge0000646

Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (2020b). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, *148*(3), 1231–1247. DOI: https://doi.org/10.1121/10.0001730

Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2021). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, *1491*(1), 89–105. DOI: https://doi.org/10.1111/nyas.14532

Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020c). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, *52*(2), 723–740. DOI: https://doi.org/10.3758/s13428-019-01271-9

Prieto, P., Cravotta, A., Kushch, O., Rohrer, P., & Vilà-Giménez, I. (2018). Deconstructing beat gestures: A labelling proposal. In *Proceedings of the 9th International Conference on Speech Prosody* (pp. 201–205). Poznań, Poland. DOI: https://doi.org/10.21437/SpeechProsody.2018-41

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. DOI: https://doi.org/10.1016/j.wocn.2014.10.005

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

Riad, T. (2006). Scandinavian accent typology. *Sprachtypologie und Universalienforschung (STUF)*, *59*, 36–55. DOI: https://doi.org/10.1524/stuf.2006.59.1.36

Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *Proceedings of the 5th International Conference on Speech Prosody*. Chicago, USA.

Rusiewicz, H. L. (2010). *The role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures* (Doctoral dissertation). University of Pittsburgh.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, *57*, 283–300. DOI: https://doi.org/10.1016/j.specom.2013.06.004

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

Shattuck-Hufnagel, S., Ren, A., Mathew, M., Yuen, I., & Demuth, K. (2016). Non-referential gestures in adult and child speech: Are they prosodic? In *Proceedings of the 8th International Conference on Speech Prosody* (pp. 836–839). Boston, USA. DOI: https://doi.org/10.21437/SpeechProsody.2016-171

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, *9*, 1514. DOI: https://doi.org/10.3389/fpsyg.2018.01514

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471–2485. DOI: https://doi.org/10.1121/1.417955

Sveriges Television (SVT) [Swedish Television]. (2013). *Rapport* [Television broadcast].

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. DOI: https://doi.org/10.1016/j.wocn.2009.10.002

Wagner, P. (2005). Great expectations – Introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proceedings of Interspeech 2005* (pp. 2381–2384). Lisbon, Portugal. DOI: https://doi.org/10.21437/Interspeech.2005-41

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, *51*(13), 2847–2855. DOI: https://doi.org/10.1016/j.neuropsychologia.2013.09.027

Willems, R. M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain and Language*, *101*(3), 278–289. DOI: https://doi.org/10.1016/j.bandl.2007.03.004

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech communication*, *46*(3–4), 220–251. DOI: https://doi.org/10.1016/j.specom.2005.02.014

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10). Aix-en-Provence, France.

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