Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses

Patrick Rohrer, Elisabeth Delais-Roussarie, Pilar Prieto

Available online at www.sciencedirect.com

ScienceDirect

Lingua 293 (2023) 103583

Lingua

www.elsevier.com/locate/lingua

Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses

Patrick Louis Rohrer [a,b,*], Elisabeth Delais-Roussarie [b], Pilar Prieto [c,a]

[a] Grup d'Estudis de Prosòdia (GrEP), Department of Translation and Language Sciences, Universitat Pompeu Fabra, Carrer Roc Boronat, 138 office 53.204, 08018 Barcelona, Catalonia, Spain

[b] Nantes Université, Laboratoire de Linguistique de Nantes (LLING-UMR 6310), Chemin de la Censive du Tertre, B.P. 81227, 44312 Nantes, France

[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

Received 12 January 2023; revised 20 July 2023; accepted in revised form 21 July 2023; available online 4 August 2023

Abstract

Research has shown a close temporal relationship between prominence-lending tonal movements in speech and prominence in manual gesture. However, prosodic structure consists of not only prosodic heads (i.e., pitch accentuation) but also prosodic edges. To our knowledge, no previous studies have assessed the value of prosodic edges (nuclear vs. phrase-initial prenuclear pitch accents) as anchoring sites for different types of gestures (i.e., referential vs. non-referential) while independently controlling for the relative degree of prominence associated with the pitch accent. The English M3D-TED corpus, which contains over 23 minutes of multimodal speech, was analyzed in terms of prosody and gesture. Results showed that while the majority of manual gesture strokes overlapped a pitch accented syllable (85.99%), apex alignment occurred at a relatively low rate (50.4%) and alignment rates did not significantly differ between referential and non-referential gestures. At the phrasal level, crucially our results also showed that strokes align with phrase-initial prenuclear pitch accents over nuclear accents, and this relationship is not driven by relative prominence. These findings show that both prosodic heads and prosodic edges (i.e., phrase initial and final positions) are key sites for both referential and non-referential gesture production.

Keywords: Gesture-speech synchronization; Prominence; Manual gesture; Pitch accentuation; Phrasal prosodic structure

* Corresponding author at: Carrer Roc Boronat, 138 office 53.204, 08018 Barcelona, Catalonia, Spain.
E-mail addresses: patrick.rohrer@upf.edu (P.L. Rohrer), Elisabeth.Delais-Roussarie@univ-nantes.fr (E. Delais-Roussarie), pilar.prieto@upf.edu (P. Prieto).

https://doi.org/10.1016/j.lingua.2023.103583
0016-7037/

1. INTRODUCTION

In communication, speakers naturally make use of a variety of multimodal resources. Two such resources are the use of speech prosody and the use of co-speech gestures. Speakers move their hands and body in a communicative way and evidence from the gesture field has revealed that (a) these manual gestures are semantically and pragmatically coherent with speech, and (b) a close temporal relationship exists between prominence-lending tonal movements (i.e., pitch accentuation) and prominence in gesture (for more information on the synchrony rules, see McNeill, 1992). Indeed, initial qualitative observations suggest that the stroke of a manual gesture (that is, the interval in time in which the peak of effort in the gesture occurs, Kendon, 1980; McNeill, 1992) generally co-occurs with, or slightly precedes, a stressed syllable. Such findings have given way to much more quantitative analyses across languages. Numerous studies have since investigated the relationship between prominence in speech and prominence in gesture. These studies have varied widely in terms of the type of speech studied (i.e, natural, [semi-]spontaneous speech vs. speech produced in highly controlled tasks in a laboratory setting) and the target types of gesture studied, as well as the landmarks chosen in speech and gesture to assess synchrony.

1.1. Gesture types, landmarks, and their association with prominence

In terms of gesture types, among the most widely used gesture typologies is that proposed by McNeill (1992), which divides manual co-speech gesture into iconic, metaphoric, deictic, and beat gestures. Iconic gestures imagistically represent concrete objects or ideas in speech, while metaphoric gestures imagistically represent abstract ideas in speech. Deictic gestures refer to spatial relations with concrete or abstract entities (i.e., pointing). Finally, beat gestures have been described as gestures which do not represent semantic content in speech, but rather are gestures with simple biphasic movements of the hands that associate with speech prominence and rhythm and have special discourse-pragmatic functions (McNeill, 1992:15). Crucially, most studies on the temporal association between speech and gesture prominence have either focused on only one type of gesture or have not considered the effects of gesture type at all. The approach adopted in the current study divides gestures into two broad categories based on their referentiality: gestures which are referential to semantic content in speech (corresponding to McNeill's (1992) iconic, metaphoric, deictic types) and those which are non-referential and therefore do not show any semantic content in speech (encompassing McNeill's (1992) beat gestures) (see, e.g., Rohrer et al., 2021; Shattuck-Hufnagel and Ren, 2018). Given the previous claims by McNeill (1992) on beat gestures being specifically associated with prosodic prominence, there is clearly a need to further assess the temporal alignment patterns of referential vs. non-referential gestures.

In terms of gesture landmarks, studies generally assessed the alignment behavior of two elements within the phasing structure of a gesture that represent its prominence, namely the *stroke* and the *apex* of a gesture. The stroke refers to the most prominent movement that bears gestural meaning, usually identified by factors such as speed, direction, or hand shape (particularly for referential gestures). A few studies have assessed the overlap between gestural strokes and prosodic prominence. For example, Karpiński et al. (2009) studied task-oriented dialogues carried out by Polish speakers and found that 75% of gesture strokes (without taking gesture type into account) overlap with a strong metrical prominence according to the RaP (Rhythm and Prominence) method of annotation (Dilley and Brown, 2005). Another study by Shattuck-Hufnagel and Ren (2018) investigated the temporal overlap of gesture strokes and ToBI (Tones and Breaks Indices)-defined pitch accented syllables in a 30-minute English academic lecture. The authors found that 83.13% of non-referential strokes overlapped with syllables that were annotated as having a pitch accent, with similar rates for referential gestures (82.85%).

The most commonly studied gestural landmark for testing temporal association is the *apex* of a gesture, which refers to an instant in time representing the "kinetic goal of the stroke" (Loehr, 2004; see also Loehr, 2007:190). This phenomenon has alternatively been termed "hits" in some studies (e.g., Yasinnik et al., 2004, see also Rohrer et al., 2021 for more details) and generally refers to points in time in which the hands reach maximum extension, suddenly stop, or change direction. A number of laboratory-based studies have found a robust relationship between gesture apexes and pitch accentuation. For example, Leonard and Cummins (2011) employed motion-tracking devices to investigate non-referential gesture production by a native English speaker. They used a reading task in which the speaker was instructed to produce beat gestures on 3 prominent syllables, which had been chosen beforehand. The authors found that gesture apexes tended to be the least variable in terms of their timing with prosodic landmarks (compared to 4 other gestural kinematic landmarks: onset of movement, peak velocity of extension/retraction, and offset of movement). They also found that the closest prosodic landmark to gesture apexes was the peak of the pitch within the pitch accented syllable. A number of laboratory-based studies have focused on the production of deictic (i.e., pointing) gestures, and how this is modulated by the phonetic realization of the target words. For example, Esteve-Gibert and Prieto (2013) investigated the production of deictic gestures in a picture-naming task carried out in Catalan, where

15 participants uttered target words with varying metrical structures in an embedded sentence. They found that the apex of the pointing gesture occurred during the pitch accented syllable, and that the apex showed a stronger correlation with the pitch peaks than other gestural landmarks (i.e., the stroke onset and offset) did, regardless of the metrical structure of the target word (see also, Rochet-Capellan et al., 2008).

Studies assessing the apex as the gestural landmark in natural speech data (that is, naturally occurring, [semi-] spontaneous speech) have generally not distinguished gestures by their referentiality. Loehr (2004) analyzed a total of 2 minutes and 44 seconds of conversational speech from 4 speakers. The author subsequently found that gesture apexes occurred within 275 ms of a pitch accent (annotated as the highest or lowest point of pitch within the vowel of the stressed syllable) 74.8% of the time, with the average distance being 17 ms before the pitch accent (SD = 341 ms). Jannedy and Mendoza-Denton (2005) analyzed 59 seconds of multimodal speech from an audience member at a public congressional town hall meeting and found that 95.7% of apexes co-occur with pitch accentuation. Similarly, Yasinnik et al. (2004) analyzed approximately 5 minutes of speech from a single speaker giving an academic lecture and found that 90% of hits (apexes) occurred within the boundaries of a pitch accented word. Another study by Esposito et al. (2007) analyzed two 4-minute Italian dialogues by two native speakers (one male and one female) and reported rates of alignment between hits and pitch accented syllables to be 78% and 84% for each speaker, respectively. Finally, in a narrative retelling task carried out by native Turkish speakers, Turk (2020) identified different tonal events (*f0* minima and maxima) associated with pitch accents, phrase accents, and boundary tones and assessed their temporal relationship with the gesture apex. The author found that apexes associated mainly with pitch accents. Finally, Pouw and Dixon (2019) used motion trackers to measure the gesture productions of 4 speakers carrying out a narrative retelling task. Peak *f0* values were extracted from the associated words and then the distance between the pitch peak and 3 kinematic measures from gesture were assessed. These measures were peak acceleration, peak velocity, and peak deceleration (the latter corresponding closest to the gesture apex). The distributions showed no significant difference by gesture type. Across all gestures, the two landmarks in gesture most closely associated with peak pitch were peak velocity (leading peak pitch by an average of 39 ms; SD = 454 ms) and peak deceleration/apex (occurring on average 44 ms after peak pitch; SD = 424 ms). Thus, the results of this cohort of studies suggest a close temporal association between pitch accentuation and gestural apexes, in particular. However, a closer evaluation of this literature raises some key points worth considering.

First, studies that reported high rates of alignment between apexes and pitch accented syllables were mostly laboratory-based studies with controlled conditions, in which participants were often instructed to produce gestures on particular words (e.g., Esteve-Gibert and Prieto, 2013; Leonard and Cummins, 2011). Second, studies using (semi-)spontaneous speech have shown methodological differences in the procedure used to assess temporal alignment. Some studies chose a specific time window to assess alignment. For example, Loehr's (2004) alignment criteria was based on a 275 ms time window around the occurrence of a pitch accent. This number was calculated based on his own data, considering the average distance between any gestural event (i.e., begin time and end time of gesture phases, apexes) and tonal event (i.e., pitch accents, phrase accents, and boundary tones), and found that "the majority of the tones . . ., regardless of type, tended to occur within a distance of 272 msec from the nearest gestural annotation." (Loehr, 2004:103). As a result, the author considered any apex occurring within 275 ms of a pitch accent to be "aligned," which also happens to loosely correspond to the average word length in his data. An alternative approach employed by Turk (2020) involved a two-step procedure, in which apexes were first paired with the nearest *f0* tonal event that shared the semantic meaning with the gesture, and then the pairings were tested for synchrony. Specifically, the two phenomena were considered synchronized if they occurred within the time window of the average syllable duration of the entire corpus (160 ms). The use of such time windows means that gesture apexes may have been considered as aligned even when the apex occurred outside the boundaries of a pitch accented syllable. Furthermore, studies which have relied solely on the pitch peak as a prosodic landmark (e.g., Leonard and Cummins, 2011; Pouw and Dixon, 2019) take *f0* as a proxy to prominence through (rising or high) pitch accentuation. However, prominence is encoded not only in *f0*, but also in combination with other phonetic cues, such as intensity and duration. Additionally, *f0* rises may occur in non-pitch-accented syllables, such as in edge tones or displaced pitch accent peaks. Thus, relying solely on *f0* on a phonetic level may not give a holistic picture of how gestures associate with speech prominence.

Only two studies have specifically assessed the co-occurrence of the apex within the boundaries of pitch accented syllables. The first study by Esposito et al. (2007) included hits (apexes) from other articulators, including the head, shoulders, and eyebrows. When focusing only on the manual gestures, the male speaker produced 138 manual hits, of which 87 aligned with a pitch accented syllable (63%), while the female speaker produced 3 manual hits, of which 2 aligned with a pitch accented syllable (66.7%). In the second study, Yasinnik et al. (2004) found that in polysyllabic words which contained a hit, 90% also contained a pitch accent (in other words, the authors took a larger time window, the word instead of the syllable, to assess alignment). In monosyllabic words, the authors found much lower rates of alignment (65%).

Finally, some studies have reported conflicting results. McClave (1994) investigated the timing of 50 rhythmic "beat" gestures produced in conversational speech. The author found that these rhythmic gestures did not all align with pitch accented syllables, but rather that one of the gestures within the rhythmic group would align with the tone nucleus, and the others would rhythmically span out, falling on both accented and unaccented syllables. A few laboratory-based studies have also found conflicting results with deictic gesture production in a picture-naming task. In Dutch, De Ruiter (1998, as cited in Esteve-Gibert and Prieto, 2013) found that lexical stress did not influence the production of deictic gestures (though in a second experiment, the author found that contrastive focus acted as an attractor for gesture apexes), and in English, Rusiewicz (2010) found that the apexes of pointing gestures tended to align with word onset regardless of the metrical structure of the target words or their contrastive status. In any case, regardless of methodological differences and some contrasting findings, it is held that there is generally a close relationship between prominence in gesture and prominence in speech.

Crucially, it is important to note that not all prosodically prominent positions in speech attract gesture. Even though several authors have claimed that nuclear pitch accents within the phrase attract more gestures (e.g., Kendon, 1980; McClave, 1994, 1998), very few studies have empirically assessed this issue. An important question is thus whether the degree of perceived accentual prominence is the main factor in the attraction of gesture or whether it is their position within a prosodic phrase (that is, phrasal prosodic structure, or the position of the pitch accent within the phrase). In the next subsection, we review the studies that have considered phrasal prosodic structure (incorporating nuclearity and prosodic phrasing) in the synchrony between gesture and speech prosody.

1.2. The role of phrasal prosodic structure in gesture production

In the Autosegmental-Metrical approach to prosodic theory (Pierrehumbert, 1980), nuclearity is defined in terms of phrasal position (see also Ladd, 2008:133). The term "nuclear" was inherited from the British school of intonation (see Nolan, 2022 for a review) and designates the last instance of a phenomenon in a phrase. As such, the nuclear pitch accent is the final pitch accent that occurs, either at the intermediate phrase (ip-) level (thus, being ip-nuclear) or at the intonational phrase (IP-) level (thus, being IP-nuclear). Any pitch accent that occurs in the phrase before the nuclear pitch accent is designated prenuclear (and unaccented syllables that are uttered after the nuclear pitch accent are designated post-nuclear). Though not explicitly integrated in this conceptual definition, many authors agree that in English, the nuclear pitch accent is generally (though not always) the most prominent pitch accent in the phrase (e.g., Calhoun, 2010b; Ladd, 2008). The only study to our knowledge to directly investigate the effect of pitch accent nuclearity in gesture attraction was by McClave (1998), who investigated the position of referential gestures in English spontaneous dyadic conversation. The author found that over half of the gestures co-occurred with the nuclear pitch accent. A few studies have investigated the effects of higher-level phrasal prosodic structure by specifically looking at the temporal relationship between the onsets and offsets of prosodic phrases and the temporal realization of gesture. For example, Loehr (2012) investigated the timing of gesture phrases (i.e., strokes and any associated preparations or holds, as per Kendon, 1980; henceforth *g-phrase*) and found that g-phrase onsets occur in close temporal proximity to ip onsets. Furthermore, both the temporal location of pitch peaks in rising pitch accents, as well as pointing gesture apexes are sensitive to an upcoming prosodic boundary (Esteve-Gibert and Prieto, 2013), and gesture lengthens both under prominence and at prosodic boundaries (Krivokapić et al., 2017). Looking higher in the prosodic hierarchy, Shattuck-Hufnagel and Ren (2018) found that Gesture Units with similar kinematic forms tended to align with prosodic groupings above the level of the IP. While the existence of such prosodic phrases is not explicitly described in the AM model as they do not seem to be regularly marked by intonational phenomena, Kendon (1980) describes these as Locution Clusters (potentially coinciding with what Selkirk (1981) described as the "utterance" level). Thus, more research is needed to investigate the effects of higher-level prosodic structure on gesture production.

At this juncture, we expect that pitch accents in edge positions (in particular, phrase-initial and phrase-final positions) may display strengthening effects and that these positions will attract both prosodic and gestural prominence. From the prosodic angle, there is clear evidence that prosodic structure (and, specifically, prosodic edges) modulates the phonetic realization of segments (see Cho, 2016, for an overview). While pre-boundary lengthening seems to privilege *domain-final strengthening* (where lengthening is larger at IP- and ip-final positions than medial and initial positions), other phenomena seem to display *initial strengthening effects*. Bolinger (1985:85) described how some pitch accents that are located at prosodic edges have phrasal marking effects. The author specifically offers the example of reciting a list (e.g., "One, two, three, four, five"), whereby the first and last element in the list receive a pitch accent. The author described how the last item in the list, by default, would receive a nuclear pitch accent, yet an "attention-getting" accent may occur towards the beginning of the phrase. In terms of intonation, evidence has been found that, in French, phrase-initial *f0* rises at the smallest phrase level (the AP, or accentual phrase following Jun and Fougeron, 2000, 2002) tend to occur more frequently at IP-initial positions than in IP-medial ones (e.g., Astésano et al., 2007; see also Fougeron and

Keating, 1997; Portes et al., 2012). The edge strengthening hypothesis has been further reinforced when looking at cases of "stress shift" (e.g., the shift of the pitch accent in a polysyllabic word such as *MassaCHUsetts* - capital letters indicating the pitch accented syllable - from the third syllable to the first syllable when occurring in contexts such as *the MASsachusetts MIracle*). Investigating such cases in a radio news corpus, Shattuck-Hufnagel et al. (1994) found a significantly higher rate of such early accentuation within words that carry the first (or only) accent of a phrase, relative to words with phrase-medial or final accents. The authors thus claim that "speakers seek to actively indicate that a new intermediate intonational phrase has begun by placing a pitch accent on the first accentable syllable" (Shattuck-Hufnagel et al., 1994:382), which also coincides with strategies to avoid pitch accent clash in English. Thus, there is good evidence that phrase onsets attract pitch accents, but to our knowledge very few studies have assessed the role of phrasal prosodic structure in the association of gestures. More specifically, to our knowledge, no previous studies have assessed the role of pitch accent nuclearity in the attraction of gestures by controlling independently for relative degree of prosodic prominence.

1.3. Motivation and research questions

The present study aims to contribute with more evidence to the previous work on the role of phrasal prosodic structure (e.g., prosodic heads or speech prominence, as well as prosodic edges in terms of accentual positions within the phrase) in the gesture-speech alignment interface. Specifically, it has two objectives, namely (a) to assess the temporal overlap between manual gesture strokes and apexes (comparing both referential and non-referential gestures) with pitch accented syllables; and (b) to assess the role of pitch accent nuclearity (ip-prenuclear vs. ip-nuclear) on gesture production, specifically to determine if the location of manual gesture is driven by relative degrees of pitch accentual prominence or by a phrase-initial edge effect. To our knowledge, no study has thoroughly investigated the effects of pitch accent nuclearity (prenuclear vs. nuclear) on gesture production within the ip-level in English discourse while controlling for relative degree of prosodic prominence, and position within the phrase. The current study aims to respond to three research questions: (1) Do gesture strokes and apexes align with pitch accented syllables in English TED Talks, and is this relationship modulated by gesture referentiality? (2) Do gestures associate with pitch accents in nuclear positions more than with those in prenuclear positions? (3) Is this relationship driven by relationships of relative prominence or phrasal position (i.e., the left-most pitch accent, phrase-medial, or nuclear positions)?

In terms of the first research question, we expect strokes to be largely aligned with pitch accented syllables (around 80%, as per Shattuck-Hufnagel and Ren, 2018). By contrast, though apexes will largely align with pitch accented syllables as well, this relationship may be more variable than that of strokes (e.g., Pouw and Dixon, 2019). In terms of the second question, following recent work showing domain-initial prosodic strengthening effects and that gestures tend to begin at the onset of ips (Loehr, 2004, 2012), we expect that prenuclear pitch accents will also be key in the temporal association between prosody and gesture, in the sense that gestures will often be produced with prenuclear pitch accents. In terms of the final research question, given that nuclear pitch accents are generally more prominent than prenuclear pitch accents (e.g., Ladd, 2008), and that phrase-initial positions have been described as "attention-getting", marking the onset of a new prosodic phrase (e.g., Bolinger, 1985; Shattuck-Hufnagel et al., 1994), we expect that this relationship will not be directly driven by relative prominence but rather modulated by positional effects in phrasal prosodic structure. Specifically, we expect to find evidence of domain-initial effects, with gestures often co-occurring with pitch accents at phrase-initial positions.

2. METHODS

2.1. Materials: The English M3D-TED corpus

The English M3D-TED corpus was used in the current analysis, which was independently annotated for prosody and gesture. The entire audiovisual corpus is available online (https://osf.io/ankdx/) in the format of ELAN files (Wittenburg et al., 2006), as well as the M3D labeling manual which explicitly describes the annotation procedure and each tier that is available in the corpus (Rohrer et al., 2021). The corpus contains over 23 minutes of multimodal annotated speech and gesture from 5 different native English speakers giving a TED Talk (mean duration per speaker: 4 m 47 s). The corpus contains a total of 1156 gesture strokes, 1307 apexes, and 2033 pitch accented syllables. After removing stretches of silence or disfluent speech, a total of 1139 strokes and 1257 apexes remained in the database for analysis.

TED Talks are a form of academic speech that has been described as a "hybrid genre" (Caliendo, 2012:101, as cited in Mattiello, 2019). Similar in format to a conference talk (a presentation with a limited time slot, given by an expert), TED Talks differ in that the members of the audience are often not specialists in the field. This results in a rather informal register being adopted by TED speakers, which is more similar to spontaneous conversation. Of particular interest is

the use of narration within the genre (Mattiello, 2017; see Mattiello, 2019 for an overview). Such contexts make TED Talks an ideal genre for the study of gesture, as TED speakers are generally quite expressive and a good number of gestures typically appear in TED Talks (see, e.g., Harrison, 2021). Specifically in the English M3D-TED corpus, the mean rate of words per manual gesture, considering both referential and non-referential, is 3.93 (i.e., a gesture is produced approximately every 4 words on average). Regarding the naturality of the data, though TED Talks are oftentimes rehearsed and/or trained, the official TED guide to public speaking (Anderson, 2016) does not give details on how speakers should employ specific prosodic or gestural features in their speech. In fact, the guide proposes that speakers speak naturally and conversationally. Specific points regarding the use of prosody include recommendations such as the use of varied prosody (speech rhythm, intonational patterns, etc.) to match the intended meaning. Similarly, in terms of gesture, the guide proposes that speakers move intentionally and make use of their hands and arms to amplify their message in speech. In all cases, however, the guide highlights that this should come naturally and that there are no "rules" for speakers to follow. Thus, TED Talks can be classified as natural, academic-style discourse.

2.2. Gestural annotation

Gestural annotation was carried out by the first author and a research assistant within the context of developing the MultiModal MultiDimensional (M3D) labeling system, following the annotation guidelines fully described in the labeling manual (Rohrer et al., 2021). Specifically, it makes use of the gesture phasing tiers and the gesture referentiality tierset described below.

2.2.1. Gesture phasing

Specifically, only manual co-speech gestures were annotated (i.e., meaningful manual movements that act as an utterance, or part of an utterance, as per Kendon, 2004). All gesture annotation was carried out using frame-by-frame analysis in ELAN, with initial gesture phasing annotation undertaken without audio. Stroke identification was largely based on the kinematic properties of the movement (salient movements based on speed, hand configuration, etc.). The apex was identified in frame-by-frame analysis as corresponding to the frame in which the image of the hand(s) goes from blurry to suddenly clear, or the frame immediately preceding a change in the direction of movement. Fig. 1 (upper panel) shows an example of the gestural annotation of a Gesture Unit containing 3 non-referential gesture strokes.

Inter-annotator reliability was assessed for 2 labelers on approximately 25% of the data for each of the two key aspects of gesture phasing, namely gesture phases (i.e., the segmentation of gestures into gesture phases including preparation, stroke, hold, recovery, etc.), as well as apex annotation. The built-in inter-annotator reliability tool in ELAN was used to assess reliability for gesture phasing, which uses an algorithm to assess both temporal overlap as well as label assigned together (Holle and Rein, 2015). The algorithm returned kappa values above 0.76 for the identification of each type of phase, indicating substantial reliability. Apex location was assessed in terms of distance (in frames) between the two raters and found that 50% of apex annotations were within one frame of each other (33 ms), and 73.8% were within two frames (66 ms). This qualitative assessment of apex coding seems to indicate high rates of agreement, particularly considering that Loehr (2004) considered up to 6 frames of distance as acceptable for agreement.

2.2.2. Gesture referentiality

Once gesture phase structure was coded in ELAN without the audio, gesture referentiality (i.e., referential vs. non-referential) was then assessed with the audio. Referential gestures have a clear referent in speech through the direct or indirect representation of speech content via degrees of iconicity or metaphoricity (e.g., if a speaker is describing a ball, a referential gesture can use the hands to directly represent the ball, and may indirectly convey information about its size, which may not be mentioned in speech). Alternatively, referential gestures may show spatial relationships via deixis (e.g. pointing to a ball in the environment). Non-referential gestures, however, do not have a clear referent in speech (e.g., McNeill's (1992) "beat" gesture). Inter-annotator reliability for gesture referentiality was assessed using Gwet's Agreement Coefficient 1 (AC1, Gwet, 2008) with MASI distances as the distance metric (Artstein and Poesio, 2008; Passonneau, 2006a,b). The resulting coefficient (which can be interpreted similarly to traditional Kappa) indicated excellent agreement (AC1 = 0.895, CI (0.856, 0.933), p <.001).

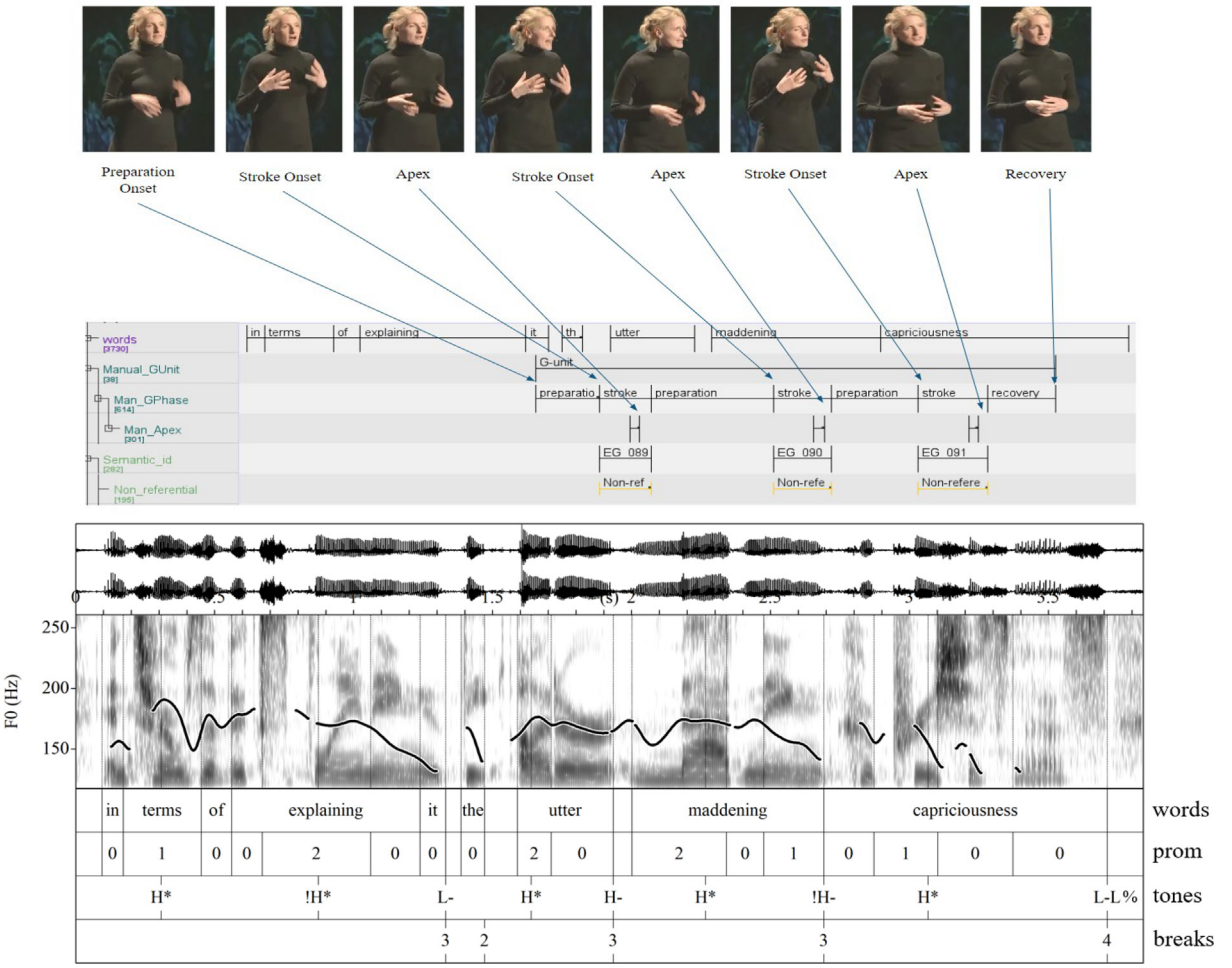

Fig. 1. Gestural and prosodic annotation of the utterance . . . in terms of explaining it. The utter, maddening capriciousness taken from the English M3D-TED corpus, speaker EG at 09:53 (Gilbert, 2009). (Upper panel): Gestural annotation in ELAN of a Gesture Unit containing 3 non-referential strokes. (Lower panel): Prosodic annotation of an intonational phrase, composed of 4 intermediate phrases.

2.3. Prosodic annotation

Prosodic annotations were carried out by the first author of this article. An orthographic transcription of speech was initially carried out in Praat (Boersma and Weenink, 2022). The transcription was then automatically aligned and segmented into words, syllables, and phones with the Montreal Forced Aligner (McAuliffe et al., 2017).

2.3.1. Phonological analysis with MAE-ToBI

Prosodic labeling was carried out following the Mainstream American English (MAE) ToBI (Tones and Breaks Indices) system (Silverman et al., 1992; Veilleux et al., 2006). Two main domains were labeled, namely phrasing and pitch accentuation. Regarding phrasing, a “breaks” tier was used to assess phrasing across 4 levels, where, importantly, a 3-break indicates an ip (intermediate phrase) boundary, and a 4-break indicates an IP (intonational phrase) boundary. IPs generally corresponded to entire clauses and had greater pre-boundary lengthening, often followed by a large pause. Intermediate phrases were identified as smaller groupings of words within the IP, which generally showed some degree of pre-boundary lengthening or a much smaller pause. Regarding pitch accentuation, a “tones” tier was used to assign the tonal target to prominent (pitch accented) syllables, as well as phrasal accents (at ip boundaries) and boundary tones (at IP boundaries).

2.3.2. Annotation of the degree of accentual prominence

An additional tier was added in Praat to the ToBI tiers to assess the degree of prosodic prominence of each syllable within an IP. Prominence annotation was adapted from the "prominence layer" described in the DIMA (Deutsche Intonation, Modellierung und Annotation) system for German (Kügler et al., 2015). The degree of prominence was annotated for each syllable on a 4-point scale. Syllables with no prominence were encoded as 0. A prominence value of 1 was assigned for weak prominences which do not necessarily coincide with an *f0* movement. Such prominences often corresponded to rhythmically motivated prominences (Calhoun, 2010a), post-focal prominences produced in a reduced pitch register, syllables that contained phrasal accents, or syllables that contain lexical stress. A prominence value of 2 was assigned to strong prominences that coincided with an *f0* tonal movement. Such prominences are said to occur with a typical pitch accent (regardless of its position within the phrase). A prominence value of 3 was assigned to extra strong prominences. These prominences show an additional emphasis that goes beyond a typical 2 prominence, oftentimes showing phonetic differences (e.g., a stronger *f0* excursion, greater intensity, etc.) but are phonologically the same as a typical 2 prominence.

To carry out the prominence annotations, the first author listened to the entire IP to identify the most prominent syllables, assigning them 2 or 3 values of prominence. Weaker prominences were then assessed relative to the stronger prominences, accounting for rhythmic constraints (i.e., rhythmically derived full pitch accents or lexical stress). Finally, remaining syllables that were not deemed prominent were assigned a value of 0. Fig. 1 (lower panel) shows an example of the prosodic annotations of the sentence ". . . in terms of explaining it, the utter maddening capriciousness." in Praat, where the first tier corresponds to the orthographic transcription (words), the second tier corresponds to the annotations of relative prominence of each syllable (prom), and the final two tiers refer to the ToBI annotation (tones, breaks). Inter-annotator reliability for 2 labelers was assessed for degree of prosodic prominence on approximately 25% of the data using Gwet's Agreement Coefficient 1 (AC1, Gwet, 2008). The resulting coefficient indicated substantial agreement (AC1 = 0.733, CI (0.71, 0.756), p <.001).

Once the prosodic annotations were completed in Praat, the annotations were imported into ELAN. The gestural and prosodic annotation data was then exported together in a time-aligned database for further processing in R (R Core Team, 2021). Finally, two important data transformations were done in R. First, pitch accented syllables were labeled in R as being either prenuclear or nuclear relative to the ip (following the definition that the nuclear pitch accent is the final pitch accent in an ip). Additionally, to operationalize the *relative degree of prominence at the level of the ip* in R (that is, to see which syllables were the most prominent in the phase), each pitch accented syllable was assessed. Specifically, if the pitch accented syllable received the highest prominence value and no other syllable was annotated at the same level of prominence in the ip, it was labeled as the "strongest prominence in the phrase." If two or more syllables shared the highest prominence value in the ip, it was labeled as sharing the "equally strongest prominence." Finally, if the prominence value was lower than another syllable in the ip, it was labeled as a "weaker prominence in the phrase." This data transformation resulted in a three-level categorical variable that was then used for the subsequent analysis.

2.4. Gesture-speech alignment criteria

In order to assess the association of gestures with speech, the temporal overlap between prosodic and gestural landmarks was assessed. The prosodic landmark of interest was the temporal span of pitch accented syllables. This prosodic landmark was specifically chosen for several reasons. First, the phonological synchrony rule describes gesture as associating with the"phonological peak syllable of speech" (McNeill, 1992:26). Additionally, this choice is motivated by some key points in Autosegmental-Metrical theory (which includes the metrical phonology and autosegmental phonology frameworks) as well as the Intonational Phonology framework. Namely, syllables themselves are organized into higher level prosodic units (metrical phonology), and act as tone-bearing units to which tones associate (autosegmental phonology). The intonational phonology framework in turn holds that pitch accents associate with metrically prominent syllables in speech. Thus, the syllable is the principal domain for accentual prominence, and this holds even in cases where tonal alignment is variable, as *f0* peaks may be displaced, occurring outside of the bounds of the pitch accented syllable (e.g., see Pierrehumbert, 1980; Prieto, 2011 for a review on the relevant work within the AM framework). In terms of gesture, two key landmarks were assessed: the stroke phase, and the apex. First, if any part of the stroke annotation temporally occurred within any part of the annotation of a pitch accented syllable, then the stroke was considered to have aligned with a pitch accented syllable (Shattuck-Hufnagel and Ren, 2018). Apexes were considered aligned with pitch accented syllables if the apex annotation fell within the boundaries of a pitch accented syllable.

2.5. Statistical analyses

A series of Generalized Linear Mixed Effects Models (GLMMs) were run using the *lme4 package* (Bates et al., 2015) in R. The random effects structure of each model was determined using the *buildmer* function (Voeten, 2022), which compares all potential combinations of random effects (e.g., random slopes by speaker, and random intercepts for relevant fixed factors as well as their interactions) and returns the best-fitting model. Models which raised convergence issues or overfit the data were re-run as Generalized Linear Models (GLMs). Omnibus test results were then carried out to assess significant main effects, which were then assessed through a series of Bonferroni-corrected pairwise tests carried out with the *emmeans* package (Lenth, 2022).

For the assessment of stroke alignment, a GLM with a poisson regression was built, which used the number of gesture strokes as the dependent variable and included a fixed factor of Gesture Referentiality (2 levels: Referential and Non-referential), a fixed factor of Alignment (2 levels: Aligned and Not aligned), and their two-way interaction. The model was offset by the total number of gestures by type. Likewise, for the assessment of apex alignment, a GLM with a poisson regression was built, which used the number of gesture apexes as the dependent variable and included a fixed factor of Gesture referentiality, a fixed factor of Alignment (2 levels: Aligned and Not aligned), and their two-way interaction. The model was offset by the total number of apexes by type.

For the assessment of the role of pitch accent position (prenuclear vs. nuclear pitch accents) in the attraction of gesture, a GLMM with a poisson regression was built, with the number of gesture strokes which aligned with a pitch accent as the dependent variable and a fixed factor of Position (2 levels: Prenuclear and Nuclear), of Gesture Referentiality (2 levels: Referential and Non-referential), as well as their two way interaction, and with by-participant varying intercepts and varying slopes for gesture referentiality. For the assessment of the relative degree of prominence (i.e., whether a gesture associated with the strongest prominence in a prosodic phrase, a syllable which shared the strongest prominence with another syllable, or with a weaker prominence), a GLM with a poisson regression was built, with the number of prenuclear-associated gesture strokes as a dependent variable and a fixed factor of Relative Prominence (3 levels: Strongest, Equally Strongest, or Weaker), a fixed factor of Gesture Referentiality, and their two-way interaction. The model was offset by the total number of gestures by referentiality. Finally, for the assessment of ip-initial edge effects, a GLM with a poisson regression was built. The dependent variable was the number of gesture strokes. The model included a fixed factor of Phrasal Position (3 levels: Phrase-initial, Phrase-medial, and Nuclear). The model was offset by the total number of gestures by position (prenuclear vs. nuclear).

3. RESULTS

3.1. Temporal alignment between pitch accented syllables and both manual gesture strokes and apexes

In response to the first research question, one goal of the current study was to assess the temporal alignment between pitch accented syllables and 2 gesture landmarks: the stroke and the apex. Table 1 below shows the by-speaker comparisons for both levels of temporal alignment.

Though there are some minor differences by speaker, the average rate of alignment across speakers between strokes and pitch accented syllables was shown to be 84.32% (SD: 5.71%). Apexes showed substantially lower rates of alignment, with the apex occurring within the pitch accented syllable at an average rate across speakers of 49.12% (SD: 7.72%). Gesture referentiality was assessed to determine if potentially one type of gesture (referential or non-referential) showed greater rates of alignment over the other. We found that for strokes, referential gestures aligned with pitch accents 88.34% of the time, and non-referential gestures aligned with pitch accented syllables 82.62% of the time.

Table 1
The alignment rates between gestures and pitch accented syllables in the English M3D-TED corpus, separated by speaker (Column 1), and between gesture strokes and apexes (Columns 2 and 3).

Speaker	Stroke Alignment (%)	Apex Alignment (%)
AS	84.08%	44.71%
EG	93.62%	57.53%
ES	84.33%	56.61%
MS	78.62%	46.99%
SJ	80.95%	39.76%
Average Rate Of Alignment	84.32%	49.12%

The GLM revealed a significant main effect of Alignment ($\chi^2(1) = 615.54$, $p < .001$), indicating that there were more gestures that aligned with a pitch accent than those that did not ($z = -19.395$, $p < .001$), as well as a significant interaction between Gesture Referentiality and Alignment ($\chi^2(1) = 13.01$, $p < .001$). The post-hoc pairwise analyses showed that when gesture strokes aligned with a pitch accented syllable, they were equally likely to be referential or non-referential in nature. However, when a gesture did not align with a pitch accented syllable, they were significantly more likely to be non-referential in nature than referential ($z = 3.190$, $p = .006$).

In terms of the apex, non-referential gesture apexes fell within the bounds of a pitch accented syllable 50.91% of the time, while referential gesture apexes fell within the bounds of a pitch accented syllable 47.52% of the time. The GLM revealed no significant effect of Gesture Referentiality ($\chi^2(1) = 0$, $p = 1$), Alignment ($\chi^2(1) = 0.096$, $p = .756$) nor a significant interaction between Gesture Referentiality and Alignment ($\chi^2(1) = 1.447$, $p = .229$). Taken together, these results indicate no tendency for gesture apexes to be either aligned or misaligned with pitch accented syllables, regardless of gesture referentiality.

3.2. Temporal association between manual gesture strokes and prenuclear and nuclear pitch accentuation

In terms of prosodic phrasing, there were two reasons for the intermediate phrase being chosen as the principal unit of analysis to understand the effect of nuclearity. First, in terms of prosodic phrasing, by choosing a smaller phrase, there is less bias in terms of the number of pitch accents in each category. The number of ips which contained more than one prenuclear accent was relatively small ($N = 101$) compared to the number containing exactly 1 prenuclear pitch accent ($N = 391$). The intonational phrase may be too large as it would naturally have many more potential prenuclear anchoring points (yet only one potential IP-nuclear anchoring point). Second, in terms of gesture, the majority of strokes occurred completely within the boundaries of the ip, which ensures that these phrases are not too short in duration to accommodate gesture production (that is, if many strokes overlapped ip boundaries, it would be difficult to assess positional effects). Thus, by choosing the ip, we can better control for the number of potential anchoring points of individual gestures, providing more insight into the relevance of the prenuclear/nuclear distinction for gesture attraction.

In order to find adequate contexts that allow us to assess whether prenuclear or nuclear pitch accents acted as stronger attractors for gesture production, the data was filtered. Of the 1139 gesture strokes that were annotated, 216 were removed from the analysis as they crossed an ip-boundary (thus, 923 strokes occurred completely within the bounds of

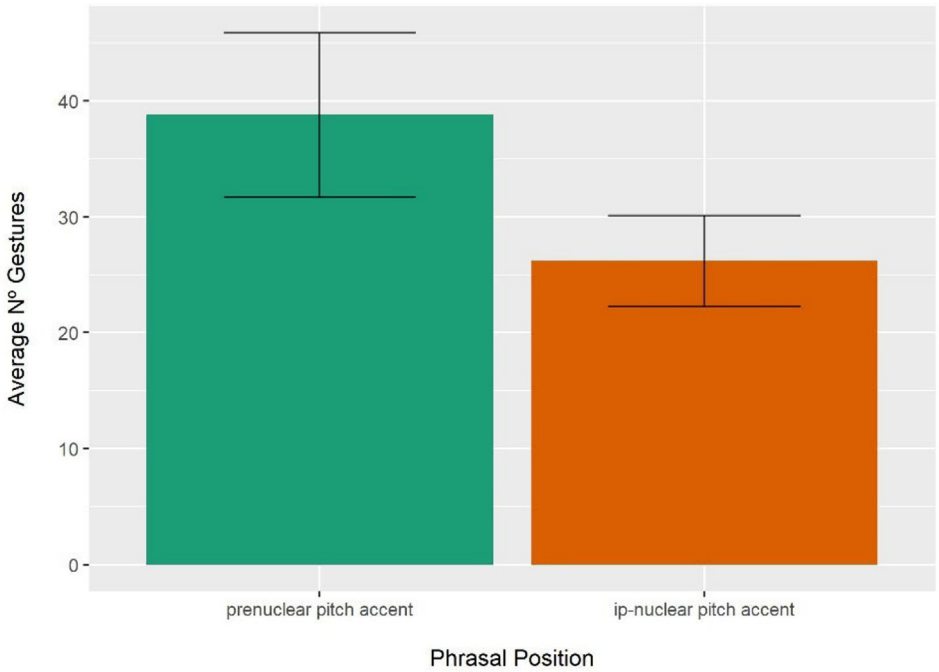

Fig. 2. The average number of gestures per speaker as a function of the phrasal position with which they align (prenuclear vs. ip-nuclear) when the ip contains at least two potential anchoring points.

an ip), and an additional 267 strokes were omitted as they either did not overlap a pitch accent (N = 171) or they overlapped multiple pitch accents (N = 96). Finally, gestures which occurred in ips that contained only one (nuclear) pitch accent were removed (N = 282). A total of 325 gestures remained for analysis, as they occurred within the boundaries of one ip which contained at least two potential anchoring positions (one or multiple prenuclear pitch accents and one nuclear pitch accent), and each stroke overlapped with only one of the potential anchoring points. Looking at such contexts, it is thus possible to assess gesture production patterns when at least two potential landmarks are present within the phrase. Of the 325 gestures in such contexts, 194 (59.69%) align with a prenuclear accent. Fig. 2 shows the average number of gestures aligning with each phrasal position. The results of the GLMM showed a significant main effect of Position ($\chi^2(1) = 12.09$, $p < .001$), where there were significantly more gestures aligning with prenuclear pitch accents than nuclear ones ($z = 3.6$, $p < .001$). A significant main effect of Gesture Referentiality was found ($\chi^2(1) = 7.97$, $p = .005$), indicating that there are significantly more non-referential gestures than referential ones ($z = 2.9$, $p = .004$). Importantly, no significant interaction was found between Position and Gesture Referentiality ($\chi^2(1) = 1.16$, $p = .282$).

These results suggest that prenuclear pitch accents have an important role as anchoring sites for the temporal integration of manual gesture. Specifically, when gestures have multiple potential anchoring points, they tend to associate with prenuclear pitch accents over nuclear pitch accents, regardless of their referentiality. The following subsections will assess whether this relationship is driven or modulated by the degree of relative prominence (that is, by assessing whether stronger prenuclear pitch accentual prominences within the ip are attracting more gestures, Section 3.3) or by a phrase-initial strengthening effect (a preference for the initial rather than medial or final positions regardless of their relative prominence, Section 3.4).

3.3. Gestural attraction towards prenuclear pitch accents: An effect of relative prominence?

In order to assess whether the attraction of gestures to prenuclear pitch accents is modulated by their relative prominence, we undertook several analyses. First, an initial analysis of the relative prominence ratings across the database showed that nuclear pitch accents were on average perceived to be more prominent than prenuclear pitch accents, with the former receiving an average prominence score of 1.74, and the latter receiving an average score of 1.53.

Fig. 3 shows the number of gestures that aligned with a pitch accent per type of pitch accent (prenuclear vs. nuclear) and their relative degree of prominence in the phrase. Of the 325 gestures used for analysis, a total of 194 aligned with a prenuclear pitch accent (the left bar of the Fig. 3), of which 39 (20.1%) occurred in cases where the associated prenuclear pitch accent was the most prominent pitch accent in the phrase, 86 (44.33%) occurred with a prenuclear pitch accent that was assessed as having the same degree of prominence as another pitch accent in the phrase, and 69 (35.57%) occurred in cases where the pitch accent had a relatively weaker prominence to another pitch accent in the phrase.

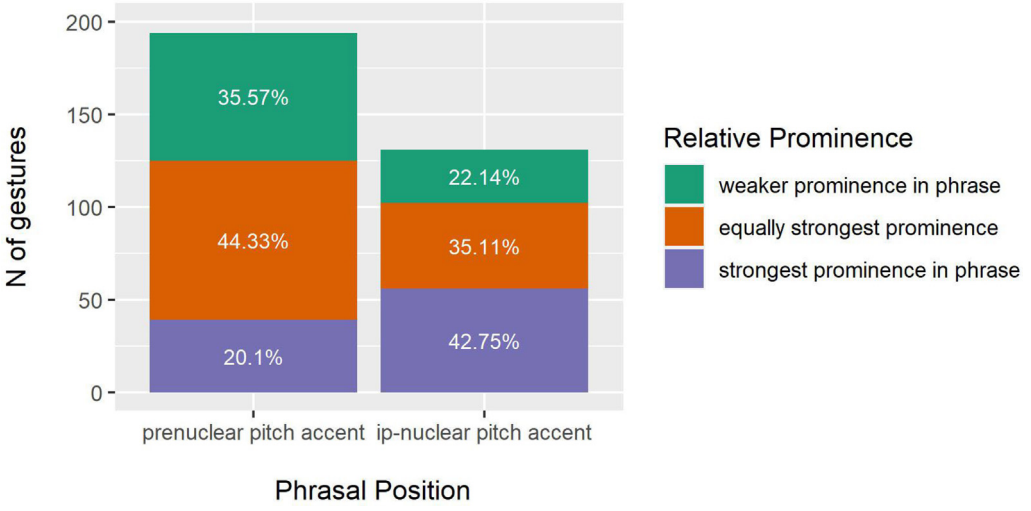

Fig. 3. The number of gestures co-occurring per phrasal position (prenuclear vs. ip-nuclear) and the relative degree of prominence within the phrase (weaker, equally, or strongest prominence in the phrase, shown as percentage of occurrence).

The GLM showed a significant effect of Relative Prominence ($\chi^2(2) = 9.72$, $p = .008$), but no significant effect of Gesture Referentiality ($\chi^2(1) = 2.84$, $p = .092$), nor their interaction ($\chi^2(2) = 1.4$, $p = .497$). Pairwise comparisons of the significant effect of Relative Prominence showed that when gestures align with prenuclear pitch accents, those accents are significantly more likely to be equally prominent to another pitch accent in the phrase than the strongest one in the phrase. Thus, the attraction to prenuclear pitch accents does not seem to be driven by prominence, and there is no effect of gesture referentiality in this relationship.

3.4. Gestural attraction towards prenuclear pitch accents: A phrase-initial strengthening effect?

In the present section we assess the presence of a phrase-initial strengthening effect through gesture production. To do so, it is necessary to assess with which position the first occurrence of a gesture associates (that is, to see if gesture tends to be produced with the left-most pitch accent). Of the 325 gestures that were used in the previous analysis, 88 were removed as they were not the first occurrence of a gesture within the phrase. Thus, the remaining 237 gestures reflect those that either occurred at initial positions (i.e., the left-most pitch accent in an ip), occurred in a medial position (where no earlier pitch accent aligned with a gesture), or occurred with the nuclear pitch accent (and no earlier pitch accent aligned with a gesture). Fig. 4 shows the average number of gestures produced by speakers according to their relation to the phrasal edges. Specifically, gestures were most often produced with the initial pitch accent of the prosodic phrase (66.67%) compared to medial (8.02%) or nuclear positions (25.32%). Results of the GLM showed a significant main effect of Phrasal Position ($\chi^2(2) = 131.87$, $p < .001$), with gestures occurring significantly more often at initial positions than in nuclear positions ($z = 6.385$, $p < .001$) or in medial positions ($z = 8.723$, $p < .001$). Additionally, gestures were significantly more likely to occur in nuclear positions than medial positions ($z = -3.96$, $p < .001$).

4. DISCUSSION AND CONCLUSIONS

The aim of the current study was to investigate the temporal association between manual gesture production and speech by taking into account the role of phrasal prosodic structure in a multimodal corpus of English academic speech (5 TED Talks containing over 28 minutes of multimodal speech). The objectives of the study were twofold, namely (a) to assess the temporal overlap between manual gesture strokes and apexes (both referential and non-referential) with pitch accented syllables; and (b) to assess the role of pitch accent nuclearity (ip-prenuclear vs. ip-nuclear) in gesture production, specifically to determine if the location of manual gesture is driven by relative degrees of pitch accentual prominence or by a phrase-initial edge effect. The current study is the first to thoroughly investigate the temporal alignment patterns of strokes/apexes within the boundaries of pitch accented syllables by taking into account the phrase-level constraints in a large English speech corpus.

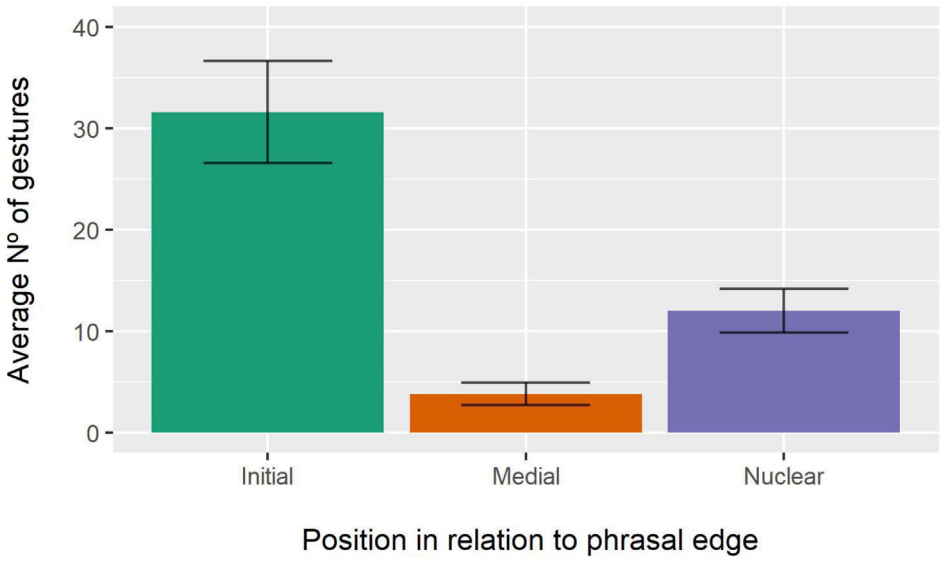

Fig. 4. The average number of gestures produced by speakers per phrasal position of the pitch accent (error bars show standard error).

Regarding the first objective, we found that gesture strokes tended to overlap with pitch accented syllables at an average rate of 84.32%, with no significant differences between how referential and non-referential gestures align with pitch accented syllables. These results reinforce the idea that all gesture types, whether referential or not, align with prosodic prominence to similar degrees. This finding is in line with what has been reported previously in the literature (e.g., Shattuck-Hufnagel and Ren, 2018 for English; Karpiński et al., 2009 for Polish). Furthermore, it closely resembles previous results that have compared gesture types, where referential and non-referential gestures have been reported to align at rates of 82.85% and 83.13%, respectively (Shattuck-Hufnagel and Ren, 2018). Interestingly, the results suggest that when gestures do not align, they are significantly more likely to be non-referential in nature, which runs contrary to the idea that non-referential (beat) gestures are inherently more prosodic in nature.

The results for apex alignment showed an average alignment rate of 50.91%. This may seem low compared to the results found in previous studies. These differences can mainly be attributed to how alignment was operationalized (see Section 1.1). Most studies have taken a set time-window around the pitch accent to assess alignment (e.g., Loehr, 2004, 2012), or used continuous measures of distance between landmarks in both gesture and prosody (e.g., Pouw and Dixon, 2019). However, a closer look at their results indicates that although the average distance is quite close, there is a rather high amount of variability. For example, Loehr (2004) reports a mean distance of 17 ms, with a standard deviation of 341 ms, where the standard deviation is slightly larger than the average word duration in his data (approx. 300 ms). A closer look at the misaligned apexes in our data showed that these points still co-occurred very close to pitch accented syllables, with over 66% of misaligned apexes occurring on the syllable immediately preceding or following a pitch accented syllable (with a slight preference for occurring on the syllable following the pitch accented one, and over 91% occurring within a distance of two syllables). Thus, it is reasonable to conclude that many apexes do not fall within the bounds of a pitch accented syllable but in neighboring syllables, and thus only loosely associate with prosodically prominent syllables in speech.

The result that the stroke as a whole (and not precisely the apex of the stroke) stably aligns with pitch accentuation is in line with results of other studies suggesting that movement phases of gesture in general may be prominence-lending and thus be more prone to align with prosodic prominence. For example, McClave (1998) found preliminary evidence that, occasionally, some speakers may tend to speak and gesture so that pitch and manual movements mirror each other (i.e., the right hand rises as pitch rises, and goes down as pitch falls). Similarly, Ambrazaitis et al. (2020) suggested that not only strokes, but any movement phase of a gesture may be prominence-lending. The authors assessed the temporal association of gesture movement phases with Swedish compound words in a spontaneous speech corpus. Swedish compound words contain two lexical stresses, the first of which usually being associated with primary stress. Prosodically, the primary lexical stress is marked by a “late fall” (H*+L) followed by a subsequent peak (H) in the secondary stress, which acts to mark sentence-level prominence (and is, consequently, associated with high levels of prominence). The authors found that not only the stroke, but any potential movement phase could align with prominent syllables. Specifically, the authors found a preference for stressed syllables to co-occur with preparations and gesture strokes but showed that holds and even retractions could co-occur with stressed syllables. Similarly, a study by Fung and Mok (2018) described a speaker who regularly showed a lag between the apex of their deictic gestures and prominent syllables. The authors suggest that there might be individual by-speaker variation in terms of strategies to achieve gesture-speech synchrony, and that the speaker may have been aligning the movement phase of the stroke with the stressed syllable, as opposed to the apex. Thus, it remains unclear exactly the degree to which the apex can be said to be a meaningful and robust gestural anchor compared to other kinematic landmarks within the stroke or even gestural movements outside of the stroke.

Regarding the second objective on the role of the nuclearity of the target pitch accents (i.e., prenuclear or nuclear pitch accents), the current study unexpectedly found that gestures have a tendency to shift towards phrase-initial prenuclear positions, and that this was not driven by relative prominence at the phrasal level. To our knowledge, the only previous study that has assessed the relationship between nuclear and prenuclear pitch accentuation is that by McClave (1998). Even though the study found that gestures tended to associate with nuclear pitch accents, it only assessed referential gestures and did not offer any quantitative analysis. Similar results to the current study have been found in Swedish compound words. Specifically, Ambrazaitis et al. (2020) found that when only one of the two syllables in Swedish compound words overlapped with the movement phase of a gesture, it was almost always with the first (primary) lexical stress. Importantly, they found that movement phases associated with the first lexical stress, even in cases in which the secondary stress (marked by the additional pitch rise, see above) was considered more prominent, acoustically. The authors suggest that the movement phases of gesture may be marking the primary stress of compounds (regardless of their relative prominence) and by doing so, the gesture may be functioning to identify compound words as a single unit (as opposed to being two separate words), potentially aiding disambiguation and speech processing for the listener (e.g., Guellaï et al., 2014).

Our results put into question a strict view of the one-to-one relationship between gesture prominence and prosodic prominence, as in our data, relative prominence cannot predict the pitch accents that attract gesture alignment. Rather, it seems that gesture may have a preference for prenuclear pitch accents which occur in initial positions of the prosodic phrase. Indeed, when separating prenuclear pitch accents according to their position in the phrase (being in initial, medial, or nuclear positions), we found that most gestures associated with initial positions of the ip. This finding is in line with Loehr (2004, 2012), who found that g-phrases begin in close temporal proximity to intermediate phrase onsets. In sum, the current study has found that ip-initial prenuclear pitch accents are key players in the gesture-speech temporal interface, acting as a prosodic anchor for both referential and non-referential gesture production, regardless of their relative degree of prominence in the phrase.

The results of the current analysis highlight the role of prenuclear pitch accents as potential prosodic anchors for gesture production. However, it is important to keep in mind that the analysis excluded 282 gestures as they occurred in intermediate phrases that contained only one nuclear pitch accent as an anchoring point. Thus, ip-nuclear pitch accents still play a key role for gesture-speech association, particularly when they are the only pitch accent in a phrase. The fact that nuclear positions still acted as an important prosodic anchor for gesture lends further support to the idea that edge positions are key in communication. As previously mentioned, Bolinger (1985) hypothesized that speakers of English prefer to indicate both phrase beginnings and endings with prosodic prominences. The current study suggests that a similar edge-strengthening effect is found in gesture, where speakers tend to produce gestures that associate with pitch accents in phrasal edge positions at the ip-level. While the current study did not look at the IP-level, an exploratory analysis on the 325 gestures was carried out at the request of an anonymous reviewer. When comparing ip- and IP-nuclear accents, there was greater association with IP-nuclear pitch accents than ip-nuclear pitch accents occurring in the middle of the IP; however, the difference between the two was not significant. Importantly, gestures still largely associated with prenuclear accents more than either ip- or IP-nuclear ones, highlighting the key role of prenuclear pitch accentuation as prosodic anchors in gesture-speech temporal alignment.

The reason for such patterns of multimodal speech production remains to be further investigated. It may be that, as according to Bolinger, these early prominences function as “attention-getting” multimodal prominences, and that producing multimodal cues early in speech may aid listeners in better predicting upcoming speech, which in turn may better facilitate conversation (e.g., in terms of turn-taking, see Holler and Levinson, 2019). More perceptual studies would be needed to assess the functions of such production patterns. In any case, our results support the view that higher-level prosodic structure (understood as categories above the syllable and the feet, and specifically in our case, the marking of ip constituent edges) is actively modulating the gesture-speech temporal interface, as has been suggested in previous studies (e.g., Esteve-Gibert and Prieto, 2013; Krivokapić et al., 2017; Loehr, 2012; Shattuck-Hufnagel and Ren, 2018).

The present study has some limitations. First, it involves the multimodal analysis of TED Talks, which can be seen as a specific genre of discourse under very particular contexts (rehearsed speech, given under a time limit and in front of a large audience). Though we argue that such speech is semi-spontaneous (in that it is not scripted) and natural (no explicit instructions are given to the speakers on how to speak or gesture, see Section 2.1), future studies should work to include other types of discourse, such of spontaneous conversation between multiple speakers. Only by investigating gesture-speech alignment in a variety of discourse settings and following similar methodologies can we better understand multimodal human communication. Second, methodologically, the current study also considered strict overlap from independent annotations in gesture and speech. While such an approach avoids perceptual bias, it is also limiting in that a distance of only a few milliseconds could be the determining factor of whether a gesture stroke aligns with a pitch accented syllable or not. For example, a gesture that was classified as not aligned because the stroke annotation ended a few milliseconds before the beginning of the accented-syllable annotation may strike the viewer as perceptually aligned. The use of independent annotations is quite common in the field as studies have shown that listeners are more likely to perceive gesture and speech as co-occurring (especially when the gesture occurs just before a pitch accented syllable), even if the two do not temporally co-occur (e.g., Leonard and Cummins, 2011, see also Rohrer et al., 2019). However, future studies may consider taking advantage of multiple approaches (e.g., a strict assessment of alignment, a perceptual assessment, and a more fine-grained continuous analysis of time alignment) in order to avoid edge cases such as that described above and achieve a more holistic picture. Moreover, the current study did not further investigate gestures that did not align with a pitch accent, which is indeed an important question for future work on gesture-speech synchrony. Third, while the current study has controlled for the semantic contribution of gesture (i.e., whether the gesture is referring to propositional content in speech), other pragmatic factors could be at play, such as the structuring of information in discourse. Indeed, the ip constituents may roughly align with constituents in information structure, such as topic, focus, or discourse referents. As proposed in Ambrazaitis et al. (2020), it may be possible that gesture is marking such constituents as a whole, working conjointly with prosody to offer a multimodal marking of relevant information in

discourse. Future studies should address the interplay between gesture marking, prosodic marking, and information structure marking in discourse.

All in all, the current study has shown that, regardless of gesture referentiality, gesture strokes are a robust landmark for assessing temporal gesture-speech temporal integration, while apexes are not so robust. Moreover, an important positioning effect was uncovered in the data, in which prenuclear pitch accents which occur in initial positions of ip prosodic phrases were key anchoring points for gesture association, regardless of their relative degree of prominence. This suggests that gesture visually highlights not only prosodic heads but also prosodic edges. More crosslinguistic work could potentially shed further light on how gestures visually represent prosodic structure.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for the current study are available in the Open Science Framework repository, https://osf.io/abh47/.

The annotated files generated in ELAN (Wittenburg et al., 2006), as well as a detailed description of the annotation procedures following the MultiModal MultiDimensional (M3D) labeling system, are also available in the Open Science Framework repository, https://osf.io/ankdx/.

Data is openly available on the OSF platform

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Patrick Louis Rohrer: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Visualization. **Elisabeth Delais-Roussarie:** Conceptualization, Supervision, Writing – review & editing. **Pilar Prieto:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support awarded by the Spanish Ministry of Science, Innovation and Universities (MCIU), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER) [grant numbers PGC2018-097007-B-100 “Multimodal Language Learning (MLL): Prosodic and Gestural Integration in Pragmatic and Phonological Development”; PID2021-123823NB-I00 “Multimodal Communication (MCOM): The integration of prosody and gesture in human communication and in language learning”], by the Generalitat de Catalunya [grant number 2017 SGR_971], and by the GEHM (Gesture and Head Movements in Language) Research Network, funded by the Independent Research Fund Denmark [grant number 9055-00004B]. The first author would like to acknowledge a joint Ph.D. grant, awarded by the Department of Translation and Language Sciences, Universitat Pompeu Fabra, and SGR Grant, Generalitat de Catalunya, [grant number 2017 SGR_971]. We would like to express our gratitude to Ulya Tütücünbasi for her help with the gestural annotation of the corpus. Many thanks are also extended to Dr. Stefanie Shattuck-Hufnagel and Ada Ren (Massachusetts Institute of Technology), as well as the members of the GrEP research lab for their feedback provided on early versions of this study. We would also like to thank Steve Dunne for his help with proofreading the final manuscript.

REFERENCES

Ambrazaitis, G., Zellers, M., House, D., 2020. Compounds in interaction: patterns of synchronization between manual gestures and lexically stressed syllables in spontaneous Swedish. Proceedings of Gesture and Speech in Interaction (GESPIN2020). KTH Royal Institute of Technology, Stockholm, Sweden https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1539106&dswid=-2101.

Anderson, C., 2016. TED Talks: The official TED guide to public speaking. Houghton Mifflin Harcourt, Boston, MA.

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Comput. Linguist. 34, 555–596. https://doi.org/10.1162/coli.07-034-R2.

Astésano, C., Bard, E.G., Turk, A., 2007. Structural influences on initial accent placement in French. Lang. Speech 50 (3), 423–446. https://doi.org/10.1177/00238309070500030501.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.
Boersma, P., Weenink, D., 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.14.
Bolinger, D., 1985. Two views of accent. J. Linguist. 21 (1), 79–123 https://www.jstor.org/stable/4175764.
Calhoun, S., 2010a. How does informativeness affect prosodic prominence? Lang. Cognit. Process. 25 (7–9), 1099–1140. https://doi.org/10.1080/01690965.2010.491682.
Calhoun, S., 2010b. The centrality of metrical structure in signaling information structure: A probabilistic perspective. Language 86 (1), 1–42 https://www.jstor.org/stable/40666298.
Cho, T., 2016. Prosodic boundary strengthening in the phonetics-prosody interface. Lang. Linguist. Compass 10 (3), 120–141. https://doi.org/10.1111/lnc3.12178.
Dilley, L., Brown, M., 2005. The RaP (Rhythm and Pitch) Labeling System, v. 1.0 [Unpublished Manuscript]. https://tedlab.mit.edu/tedlab_website/RaP%20System/RaP_Labeling_Guide_v1.0.pdf.
Esposito, A., Esposito, D., Refice, M., Savino, M., Shattuck-Hufnagel, S., 2007. A preliminary investigation of the relationship between gestures and prosody in Italian. In: Esposito, A., Bratanić, M., Keller, E., Marinaro, M. (Eds.), Fundamentals of verbal and nonverbal communication and the biometric issue, vol. 18. IOS Press, Amsterdam, pp. 65–74.
Esteve-Gibert, N., Prieto, P., 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. J. Speech Lang. Hear. Res. 56 (3), 850–864. https://doi.org/10.1044/1092-4388(2012/12-0049).
Fougeron, C., Keating, P.A., 1997. Articulatory strengthening at edges of prosodic domains. J. Acoust. Soc. Am. 101 (6), 3728–3740. https://doi.org/10.1121/1.418332.
Fung, H.S.H., Mok, P.P.K., 2018. Temporal coordination between focus prosody and pointing gestures in Cantonese. J. Phon. 71, 113–125. https://doi.org/10.1016/j.wocn.2018.07.006.
Gilbert, E., 2009. Your elusive creative genius. [video]. TED Conferences. https://www.ted.com/talks/elizabeth_gilbert_your_elusive_creative_genius.
Guellaï, B., Langus, A., Nespor, M., 2014. Prosody in the hands of the speaker. Front. Psychol. 5, 700. https://doi.org/10.3389/fpsyg.2014.00700.
Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61, 29–48. https://doi.org/10.1348/000711006X126600.
Harrison, S., 2021. Showing as sense-making in oral presentations: The speech-gesture-slide interplay in TED Talks by Professor Brian Cox. J. Engl. Acad. Purp. 53,. https://doi.org/10.1016/j.jeap.2021.101002 101002.
Holle, H., Rein, R., 2015. EasyDIAg: A tool for easy determination of interrater agreement. Behav. Res. Methods 47, 837–847. https://doi.org/10.3758/s13428-014-0506-7.
Holler, J., Levinson, S.C., 2019. Multimodal Language Processing in Human Communication. Trends Cogn. Sci. 23 (8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006.
Jannedy, S., Mendoza-Denton, N., 2005. Structuring information through gesture and intonation. In: Ishihara, S., Schmitz, M., Schwarz, A. (Eds.), Interdisciplinary Studies on Information Structure, vol. 3. Universitätsverlag Potsdam, Potsdam, pp. 199–244.
Jun, S.-A., Fougeron, C., 2000. A phonological model of French intonation. In: Botinis, A. (Ed.), Intonation, Text, Speech and Language Technology, vol. 15. Springer Netherlands, Dordrecht, pp. 209–242. https://doi.org/10.1007/978-94-011-4317-2_10.
Jun, S.-A., Fougeron, C., 2002. Realizations of accentual phrase in French intonation. Probus 14 (1), 147–172. https://doi.org/10.1515/prbs.2002.002.
Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z., 2009. Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. Speech Lang. Technol. 11, 113–122.
Kendon, A., 1980. Gesticulation and speech: Two aspects of the process of utterance. In: Key, M.R. (Ed.), The relationship of verbal and nonverbal communication. De Gruyter Mouton, Berlin, pp. 207–228.
Kendon, A., 2004. Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, UK.
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. Lab. Phonol. 8 (1). https://doi.org/10.5334/labphon.75.
Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., Peters, J., Ritter, S., Röhr, C.T., Schweitzer, A., Schweitzer, K., Wagner, P., 2015. DIMA – Annotation guidelines for German intonation. In: The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. The University of Glasgow, Glasgow, UK.
Ladd, D.R., 2008. Intonational Phonology. Cambridge University Press, Cambridge, UK.
Lenth, R.V., 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.4-.1. https://CRAN.R-project.org/package=emmeans.
Leonard, T., Cummins, F., 2011. The temporal relation between beat gestures and speech. Lang. Cognit. Process. 26 (10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218.
Loehr, D.P., 2004. Gesture and intonation (Doctoral dissertation). Georgetown University.
Loehr, D., 2007. Aspects of rhythm in gesture and speech. Gesture 7 (2), 179–214. https://doi.org/10.1075/gest.7.2.04loe.
Loehr, D.P., 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. Lab. Phonol. 3 (1), 71–89. https://doi.org/10.1515/lp-2012-0006.
Mattiello, E., 2017. The popularisation of science via TED Talks. Int. J. Lang. Stud. 11 (4), 77–106.

Mattiello, E., 2019. A corpus-based analysis of scientific TED Talks: Explaining cancer-related topics to non-experts. Discourse Context Media 28, 60–68. https://doi.org/10.1016/j.dcm.2018.09.004.
McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., Sonderegger, M., 2017. Montreal Forced Aligner [Computer program]. Version 1.0.
McClave, E., 1994. Gestural beats: The rhythm hypothesis. J. Psycholinguist. Res. 23 (1), 45–66. https://doi.org/10.1007/BF02143175.
McClave, E., 1998. Pitch and manual gestures. J. Psycholinguist. Res. 27 (1), 69–89. https://doi.org/10.1023/A:1023274823974.
McNeill, D., 1992. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago, IL.
Nolan, F., 2022. The rise and fall of the British school of intonation analysis. In: Barnes, J., Shattuck-Hufnagel, S. (Eds.), Prosodic Theory and Practice. The MIT Press, Cambridge, MA, pp. 319–349. https://doi.org/10.7551/mitpress/10413.003.0012.
Passonneau, R., 2006a. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: Calzolari, N., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), pp. 831–836.
Passonneau, R., 2006b. Measuring Agreement on Set-Valued Items (MASI) for semantic and pragmatic annotation. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06). European Language Resources Association (ELRA), Genoa, Italy, pp. 831–836.
Pierrehumbert, J., 1980. The phonetics and phonology of English intonation (Doctoral dissertation). Massachusetts Institute of Technology.
Portes, C., D'Imperio, M., Lancia, L., 2012. Positional constraints on the initial rise in French. In: Ma, Q., Ding, H., Hirst, D. (Eds.), Speech Prosody 2012. ISCA Archive, Shanghai, China, pp. 563–566.
Pouw, W., Dixon, J.A., 2019. Quantifying gesture-speech synchrony. In: Rohlfing, K., Grimminger, A., Mertens, U. (Eds.), Proceedings of the 6th Gesture and Speech in Interaction Conference (GESPIN 6). Universitätsbibliothek Paderborn, Paderborn, pp. 75–80. https://doi.org/10.17619/UNIPB/1-815.
Prieto, P., 2011. Tonal Alignment. In: van Oostendorp, M., Ewen, C.J., Hume, B., Rice, K. (Eds.), The Blackwell Companion to Phonology, vol. 2. Wiley Blackwell, Malden, MA, pp. 1185–1203. https://doi.org/10.1002/9781444335262.wbctp0050.
R Core Team, 2021. R: A language and environment for statistical computing. Austria, Vienna, Retrieved from https://www.R-project.org/.
Rochet-Capellan, A., Laboissière, R., Galván, A., Schwartz, J.-L., 2008. The speech focus position effect on jaw-finger coordination in a pointing task. J. Speech Lang. Hear. Res. 51 (6), 1507–1521. https://doi.org/10.1044/1092-4388(2008/07-0173).
Rohrer, P. L., Vilà-Giménez, I., Florit-Pons, J., Gurrado, G., Esteve-Gibert, N., Ren-Mitchell, A., Shattuck-Hufnagel, S., Prieto, P., 2021. The MultiModal MultiDimensional (M3D) labeling system. https://doi.org/10.17605/OSF.IO/ANKDX.
Rohrer, P.L., Prieto, P., Delais-Roussarie, E., 2019. Beat gestures and prosodic domain marking in French. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (Eds.), Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc., Canberra, Australia, pp. 1500–1504.
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.
Selkirk, E., 1981. On prosodic structure and its relation to syntactic structure. In: Fretheim, T. (Ed.), Nordic Prosody II: Papers from a Symposium. Tapir, Trondheim, Norway, pp. 111–140.
Shattuck-Hufnagel, S., Ostendorf, M., Ross, K., 1994. Stress shift and early pitch accent placement in lexical items in American English. J. Phon. 22 (4), 357–388. https://doi.org/10.1016/S0095-4470(19)30291-8.
Shattuck-Hufnagel, S., Ren, A., 2018. The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. Front. Psychol. 9. https://doi.org/10.3389/fpsyg.2018.01514.
Silverman, K., Beckman, M.E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J.B., Hirschberg, J., 1992. TOBI: A standard for labeling English prosody. In: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 1992). ISCA, pp. 867–870. https://doi.org/10.21437/ICSLP.1992-260.
Turk, O., 2020. Gesture, prosody and information structure synchronisation in Turkish (Doctoral Dissertation). Victoria University of Wellington.
Veilleux, N., Shattuck-Hufnagel, S., Brugos, A., 2006. Transcribing Prosodic Structure of Spoken Utterances with ToBI. MITOpenCourseware. Retrieved from https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm.
Voeten, C., 2022. buildmer: Stepwise elimination and Term Reordering for Mixed-Effects Regression. R package version 2.4, https://CRAN.R-project.org/package=buildmer.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006. ELAN: a professional framework for multimodality research. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy, pp. 1556–1559.
Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S., 2004. The timing of speech-accompanying gestures with respect to prosody. In: Proceedings of the International Conference: From Sound to Sense. MIT, Cambridge, MA, pp. C97–C102. https://doi.org/10.1121/1.4780717.