Laboratory Phonology
Journal of the Association for Laboratory Phonology

Türk, O., & Calhoun, S. (2023). Multimodal cues to intonational categories: Gesture apex coordination with tonal events. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), pp. 1–50. DOI: https://doi.org/10.16995/labphon.6432

Open Library of Humanities

Multimodal cues to intonational categories: Gesture apex coordination with tonal events

Olcay Türk, Phonetics Workgroup, CITEC, Bielefeld University, DE, olcay.tuerk@uni-bielefeld.de

Sasha Calhoun, Te Herenga Waka – Victoria University of Wellington, NZ, sasha.calhoun@vuw.ac.nz

This study argues for a multimodal view of the identification, representation, and implementation of intonational structure, with evidence from gesture apex-tone coordination in Turkish. Many studies have reported consistent synchronisation of atomic prominence markers across modalities (i.e., pitch accents and gesture apexes). This is prima facie evidence that gesture and prosody are implemented together, and therefore the former can play a role in the identification and perception of the latter through apex-tone synchronisation. However, only few studies considered the full intonational context when investigating synchronisation (e.g., potential alignment of apexes with boundary tones). This is particularly relevant for Turkish as there is disagreement in the literature about whether all words in Turkish bear a pitch accent. In this study, we test the synchronisation of apexes with all intonational events in Turkish natural speech data annotated for gesture and prosody, resulting in 820 gesture apex and 3697 tonal event annotations. The study uses syllable duration (160ms) to determine synchronisation between these anchors via equivalence tests while also integrating gestural and prosodic context as factors that can affect the temporal distance between these units through mixed-effects linear regression. The findings showed that apexes were chiefly synchronised with pitch accents (71%), indicating that prominence was the primary constraint for synchronisation. However, analysis of cases with no prosodic prominence provides the first evidence for a hierarchical constraint on synchronisation, since apexes were preferentially synchronised with the tones marking prosodic words (76%) and not with the markers of prosodic constituents higher in the hierarchy. This finding supports the claim that there may be accentless words in Turkish since the absence of prominence caused a systematic shift in the synchronisation behaviour of apexes. More generally, the study shows how multimodal evidence from gesture can be used in the identification of phonological categories, and that prosodic structure is likely to be expressed through multimodal cues as a composite signal.

1. Introduction

Speech and gesture are implemented together (McNeill, 1992; Kendon, 2004). This joint implementation is evidenced by various highly adaptive interactions between speech and gesture (see Wagner, Malisz, & Kopp, 2014), which have been used to theorise unified speech and gesture production models (McNeill & Duncan, 2000; Kita, 2000; Krauss, Chen, & Gottesman, 2000; De Ruiter, 2000; Kita & Özyürek, 2003; Hostetter & Alibali, 2008). In light of these models, research has revealed much on how and why people gesture as they speak; however, relatively less is known about when people gesture in relation to speech. There are still many open questions about the exact nature of temporal coordination (i.e., synchronisation) of gesture with its co-speech.

Numerous studies have explored synchronisation, and these studies have generally linked prosody to gesture as the main speech component that regulates speech-gesture synchronisation (Wagner et al., 2014). The present study aims to contribute to this body of work by investigating gesture-prosody synchronisation at the smallest possible unit level (i.e., points in time) in Turkish natural speech data. For this purpose, we identify gesture apexes and tonal events as the smallest possible units (Sections 1.1 and 1.2) and test whether apexes are synchronised with any of these events systematically. Based on our findings, we make a case that there are consistent multimodal cues to intonational phonological categories in the speech stream, and that this cueing behaviour can be used in the identification of these categories through a case study in Turkish.

In what follows, we first outline the claims of previous work, highlighting what in gesture and prosody is considered to be synchronised and approaches to the quantification of synchronisation (Section 1.1). We note that in general, the selection of anchors for synchronisation has been independent from phonological considerations. We also emphasise that whether synchronisation is affected by gestural and prosodic context has been generally overlooked. In Section 1.2, we then detail our research goals, stating our view of what constitutes synchronisation, and what potential synchronisation scenarios imply for the intonational phonology of Turkish (Section 1.2.1).

1.1 Background

Early studies on gesture-prosody synchronisation were interested in the synchronisation of intervals in these modalities such as strokes (meaningful gesture parts with maximum dynamic effort) and (stressed) syllables (Creider, 1986; McClave, 1991; McNeill, 1992; Tuite, 1993, amongst others). McNeill's (1992) phonological synchrony rule summarises the general claim of these studies: "…the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech" (p. 26). In other words, prominences in gesture and prosody were claimed to be synchronised. Later, re-analysing McNeill's data, Nobe (1996) refined this rule

into a stroke-acoustic peak synchrony rule, which stated that strokes co-occur with (or precede) F0 and intensity peaks. Following on from this refinement, many later studies have focused on the synchronisation of atomic landmarks (i.e., smallest possible anchors) rather than intervals, testing whether prominent points in time in gesture and prosody are synchronised (see Türk, 2020, for an overview).

In these studies, the prominent point in time in gesture has been commonly identified as the *apex*, which is dynamically the most prominent instant corresponding to the target of gesture (e.g., endpoint as in **Figure 1**), and/or abrupt directional changes in the gesture stroke (Loehr, 2004; Shattuck-Hufnagel, Yasinnik, Veilleux, & Renwick, 2007) (see Section 2.4.1).

(a) Mid-stroke movement (b) Stroke-final maximum extension - Apex

Figure 1: Two frames showing a deictic gesture apex.

Defining prosodic prominence and selecting measurable cues, on the other hand, have been fundamental problems for studies on apex synchronisation, and in many cases, the cues used were not informed by phonological theories of prominence. Experimental studies have often assumed phonetic definitions of what prominence is. They characterised it with single acoustic parameters measured over stressed (prominent) syllables such as articulatory vocalic targets (Roustan & Dohen, 2010), jaw openings (Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008), or vowel onsets or midpoints (Leonard & Cummins, 2011; Rusiewicz, 2010; Rusiewicz, Shaiman, Iverson, & Szuminsky, 2013) without providing much insight into why those parameters were chosen. This view is in conflict with prosodic research, which has established a complex relationship of acoustic cues to prominence where multiple cues are shown to contribute to the perception of prominence (Breen, Fedorenko, Wagner, & Gibson, 2010; Cole, 2015; Arnhold & Kyröläinen, 2017; Baumann & Winter, 2018; Kügler & Calhoun, 2020). The cues to prominence depend on multiple factors such as the language in question and the position in prosodic structure. Accordingly, if particular points in time need to be measured for apex synchronisation tests, this measure should not ignore prosodic structure and how different acoustic cues are used in the language in question.

In addition to using different units of measure, previous studies have also adopted different methodologies to define and test synchronisation. In studies concerned with the synchronisation of intervals, the temporal overlaps of intervals were interpreted as synchronisation, regardless of the durational difference between the intervals and the duration of the overlap (e.g., syllables and strokes as in Jannedy & Mendoza-Denton, 2005; Shattuck-Hufnagel & Ren, 2018). Studies on the synchronisation of precise time points tended to use actual time distance measurements (Loehr, 2004; Rochet-Capellan et al., 2008; Roustan & Dohen, 2010; Rusiewicz, 2010; Leonard & Cummins, 2011; Rusiewicz et al., 2013). In some of these, synchronisation was determined as a relative temporal proximity in binary conditions where production of speech and gesture was heavily controlled (e.g., Rochet-Capellan et al., 2008; Rusiewicz, 2010). Namely, the time distances between apexes and phonetic/prosodic anchors were measured over words with varying stress positions (e.g., stress on the first vs. the second syllable). If these distances were significantly less in one condition and more in the other, then this reduction was considered as synchronisation. In other studies, the closest pre-selected phonetic/prosodic anchor to an apex was considered to be in synchrony with it (e.g., Roustan & Dohen, 2010). Such an approach, in general, is indicative of synchronisation but less absolute as it disregards the actual time distance between anchors (e.g., two units may be closest but in fact seconds away from each other). Loehr's (2004) approach counters this by considering two closest events (i.e., apexes and pitch accents) synchronised only if they occur within a pre-defined duration (i.e., 275 ms). This approach seems to be more plausible in that it provides a standard timeframe to base synchronisation interpretations on. It is also tolerant of timing differences potentially caused by other gestural and prosodic factors, which in general has not been explored in depth so far.

In fact, emerging evidence on synchronisation suggests that prosodic phrasing might be one such prosodic factor affecting gesture synchronisation. In English, gesture phrases (i.e., single meaningful unit of bodily action, Kendon, 2004) were found to be synchronised with single intermediate phrases, meaning that gestures are sensitive to prosodic boundaries (Loehr, 2004). Krivokapić, Tiede, and Tyrone (2017) showed that manual gestures lengthen under prominence and at prosodic boundaries, which potentially implies an effect on synchronisation with co-occurring prosodic units. Esteve-Gibert and Prieto (2013) and Esteve-Gibert, Borràs-Comes, Asor, Swerts, and Prieto (2017) showed that prosodic structure shapes the patterns of synchronisation as the positions of apexes and pitch accents were found to change in tandem, depending on the prosodic phrasing. Rohrer, Prieto, and Delais-Roussarie (2019) showed that in French, apexes were synchronised with pitch accents and that this synchronisation was more with (accentual) phrase-final pitch accents than phrase-initial ones, although both can be used to mark prominence. Interestingly, they also noted that synchronisations with apexes were at much lower rates compared to the rates reported in studies in English (Jannedy & Mendoza-Denton, 2005; Shattuck-Hufnagel & Ren, 2018). Moreover, in cases of asynchronisation,

apexes coincided with phrase-initial positions that contain no pitch accentuation. Their interpretation of this pattern was that apexes might be synchronised with phrase onsets marking a prosodic domain independently from the encoding of prominence. The prospect of apexes synchronising with boundaries is interesting and requires further exploration as many previous studies tested apex synchronisation with the prior assumption that apexes are only synchronised with prosodic prominence, as mentioned above. Following on from this, it might also be that in addition to phrasing, prosodic hierarchy and the relative positioning of phrases with regard to sentence prominence (e.g., nuclear vs. pre-nuclear, see Section 1.2) may also affect synchronisation patterns of apexes, which has not been investigated systematically so far.

Similarly, there might be gestural factors that can affect apex synchronisation. For instance, the semantic functions of gesture (i.e., iconics, metaphorics, deictics, and beats, see Section 2.4.1) have been mostly overlooked in many previous studies on apex synchronisation. They either did not distinguish between any gesture types in their analysis (e.g., Loehr, 2004; Jannedy & Mendoza-Denton, 2005), or only investigated apex synchronisation for one or two gesture types (Rochet-Capellan et al., 2008; Roustan & Dohen, 2010; Rusiewicz, 2010; Leonard & Cummins, 2011; Rusiewicz et al., 2013; Esteve-Gibert & Prieto, 2013; Esteve-Gibert et al., 2017). Most of these studies have analysed the synchronisation of either deictics (i.e., pointing gestures) or beats (i.e., simple flicks of hand) separately with the exception of Roustan and Dohen (2010). They compared apex synchronisation patterns of beats and deictics and found different synchronisation anchors for these where deictic apexes were synchronised with articulatory vocalic targets and beat apexes with prosodic peaks. This implies that gesture type can impact apex synchronisation. However, to our knowledge, there are not any other studies that focus on comparing apex synchronisation across different gesture types, especially for representational gestures (i.e., iconics and metaphorics – gestures that represent concrete or abstract concepts in speech). Overall, all of these highlight that gesture and prosody may be connected in more ways than previously assumed. These also imply that in terms of synchronisation, gesture and prosody must be treated as complex systems where the continuous stream of gestural and prosodic events and their organisations can influence each other.

Despite their methodological differences, however, most of the previous studies have reported a prominence-based synchronisation between gesture and prosody. That is, in these studies, dynamically prominent apexes were claimed to be synchronised with certain events that were considered to bear prominence (though see Rusiewicz, 2010; Rusiewicz et al., 2013). This indicates a consistent relationship between categories in phonology and gesture. Given this synchronisation, multimodal cues are likely to be part of the implementation of phonological structure at different levels of planning (Türk, 2020). Therefore, they could play a role in how intonational categories are produced and represented by speakers, and identified and perceived

by listeners (e.g., Krahmer & Swerts, 2007 for evidence of the effects of visual beat gestures on the perception of prosodic prominence). Furthermore, this strong relationship can also aid us as analysts. Taking advantage of tight synchronisation patterns, multimodal cues can be used to assist phonological analyses in the identification of phonological categories. In particular, if apexes are systematically synchronised with only prominence lending events, then this can be used for the identification of such phonological events in cases where there may be disagreement in literature.

The present study builds on these interpretations and predictions. Its primary goal is to show there are consistent multimodal cues to intonational phonological categories in Turkish. For this purpose, the present study interrogates the claim that atomic prominent events in gesture (i.e., apexes) and in prosody are synchronised. This investigation diverges from others in multiple ways: (1) How it selects prosodic anchors for synchronisation, (2) by accounting for the effect of gestural and prosodic context on synchronisation, and (3) how it determines synchronisation. Further, apex synchronisation has only been tested for a limited number of languages with similar intonational structures, excluding Turkish. Therefore, investigating synchronisation in Turkish enables cross-linguistic comparison with previous studies. Turkish also provides an interesting challenge to the prominence-based synchronisation claim as not every prosodic phrase is claimed to contain a prominent event (Section 1.2.1), making apex synchronisation in these cases unclear. Finally, the present study also demonstrates how any consistent multimodal cues to intonational categories can be exploited to inform phonological analyses (Section 1.2.2). In the next section, all of these points are detailed.

1.2 Present study

The present study aims to test whether gesture and prosody are synchronised at their prominent units, as claimed previously, or whether there are other anchors in prosody that gesture can be synchronised with in natural speech data. Therefore, the way in which anchors in gesture and prosody are defined is crucial for establishing meaningful synchronisation relationships. The present study diverges from previous studies in how it selects these anchors. It does not presume that there can be a synchronisation relationship only between prominent units, so it does not test synchronisation just between two pre-selected anchors. Instead, it adopts the apex as the prominent gestural anchor (Section 2.4.1) and aims to find the best synchronisation anchor amongst a set of prosodic events that have prosodic functions within Turkish prosody. The advantage of this approach is that any prosodic event, not just prominent ones, can be synchronised with apexes, which enables the discovery of other potentially crucial synchronisation patterns.

Research has shown that multiple acoustic cues contribute to the perception of prominence. It has also been shown that these cues cannot be expected to be the same for every syllable, regardless of (1) position in prosodic structure, and (2) how those acoustic cues are used in that

language (Breen et al., 2010; Arnhold & Kyröläinen, 2017; Baumann & Winter, 2018; Kügler & Calhoun, 2020). Therefore, the prosodic anchors that are tested for synchronisation should be sensitive to these factors (Section 1.1). The present study uses F0 minima and maxima (i.e., F0 turning points or tonal events) as prosodic anchors. These have been shown to be sensitive to both of these factors, and to be associated with syllables even though they are not necessarily aligned within them (Prieto, Van Santen, & Hirschberg, 1995; Arvaniti, Ladd, & Mennen, 1998; Xu, 1998; Arvaniti, Ladd, & Mennen, 2000; Atterer & Ladd, 2004; Prieto & Torreira, 2007; Ladd, 2008).

Tonal events have different functions within prosodic structure. Within the Autosegmental-Metrical (AM) framework of intonational phonology, intonation contours are construed as a sequence of high (H) and low (L) tonal targets (i.e., tonal events). These tonal events either mark prominence (i.e., pitch accents) or mark the boundaries of prosodic phrases (i.e., phrase accents or boundary tones). Within the present study, the analysis of synchronisation considered any tonal event, regardless of function, as a potential anchor for an apex. This meant that any consistent observation of synchronisation would be the result of a genuine systematic behaviour. The inventory and functions of tonal events are language specific. The next section introduces the intonational structure of Turkish.

1.2.1 Turkish phonology

To the authors' knowledge, there are no current complete intonational descriptions for Turkish within the AM framework but only partial descriptions (Özge & Bozsahin, 2010; Kamali, 2011; Ipek & Jun, 2013; Güneş, 2013, 2015). These earlier studies employed read-out tasks of pre-set sentences in their analyses, and they differ clearly from each other in their representation of Turkish tonal inventory. In the present study, it was not possible to follow any of these descriptions entirely because the study uses spontaneous natural speech data. There were many cases where the existing descriptions fell short or starkly conflicted with our naturalistic data. Therefore, the lack of consensus on the descriptions of Turkish phonology and uncontrolled variability introduced by our data led us to develop a new annotation scheme. The scheme takes earlier studies in consideration but diverges from them in the form of simplifications and generalisations that explained our data best (these will be touched upon in Section 2.4.2). In particular, where there were disagreements between previous descriptions, and/or our data did not fit any of the previous schemes, we adopted a broad phonetic approach to be reasonably transparent to the phonetic cues at the surface level (see Hualde & Prieto, 2016 and references therein). This approach can be seen as a form of transcription "…that includes a certain amount of redundant, phonologically non-contrastive detail that is nevertheless a systematic aspect of the language" (Hualde & Prieto, 2016, p. 3). It aims to be surface-transparent, and is precisely useful in cases like Turkish where the correct phonological analysis is not yet known or is disputed.

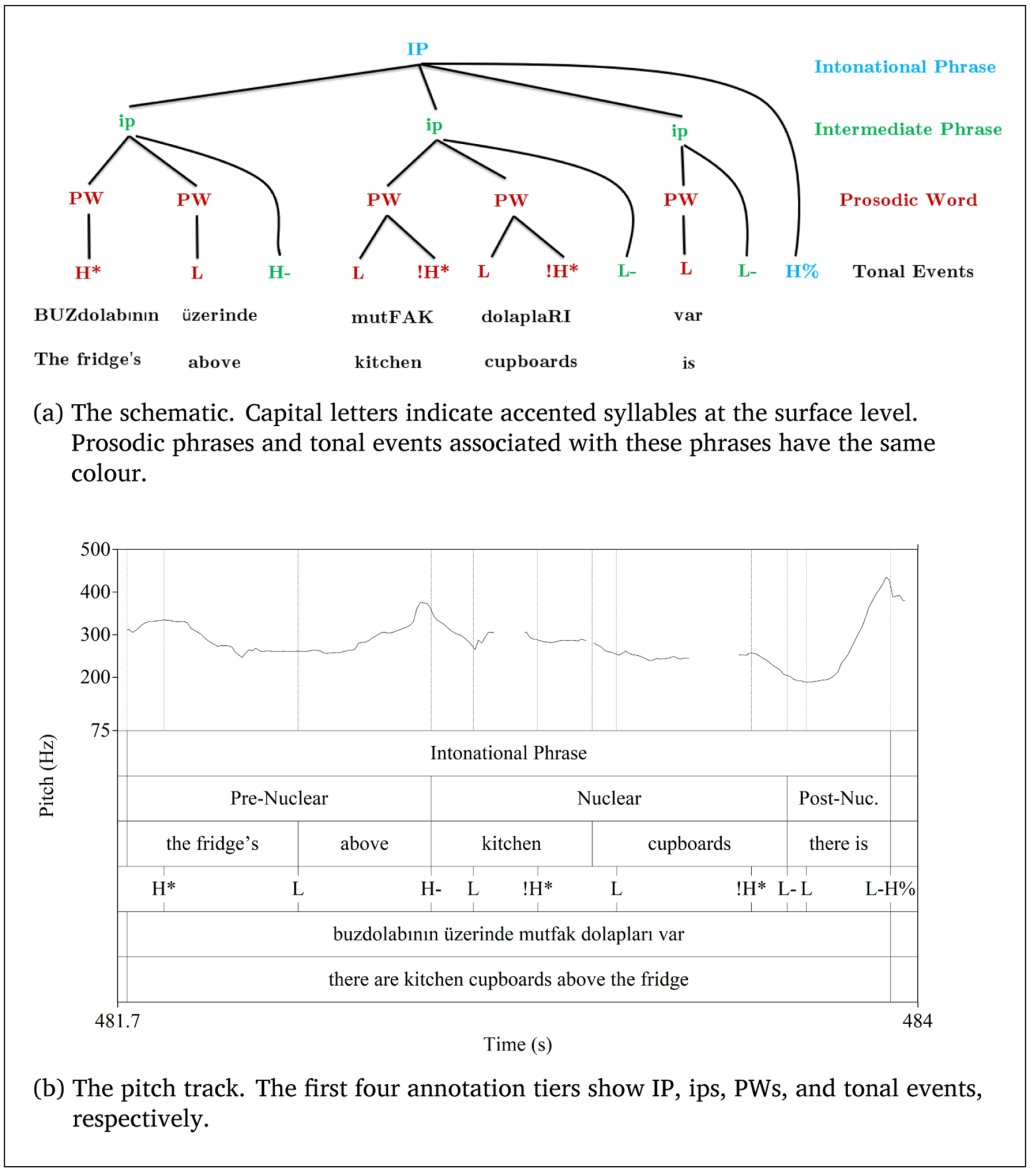
We follow this approach also in order to account for non-predictable phenomena introduced by the nature of our data. Our annotation scheme can be considered as a low-level prosodic transcription that does not aim to describe the full range of phonological units and contrasts, but to capture prosodic representations that are transparent at the surface/phonetic level (see Kügler et al., 2015 for a scheme for German with a similar motivation). We do not claim to be exhaustive in our description nor is it the only analysis possible. Our focus was on representing the common intonational phenomena found in our data even if they did not fit any of the existing descriptions. In general, however, the scheme is more similar to Ipek and Jun's (2013) descriptions than others. In the rest of this section, Turkish intonation is introduced briefly (see further Section 2.4.2) in order to demonstrate the potential tonal event candidates for apex synchronisation and what consistent synchronisation patterns with these candidates imply for Turkish intonation (Section 1.2.2).

In Turkish, there are three levels of prosodic phrasing, each usually associated with a corresponding level of morphosyntactic structure: (1) Prosodic word (PW) (≈ morphological word), (2) intermediate phrase (ip) (≈ syntactic phrase), and (3) intonational phrase (IP) (≈ syntactic clause). Each of these is marked by specific tonal events. The IP is the largest phrase in the hierarchy, and it is marked by either a low or high boundary tone at the right edge (L% or H%). Similar to the IP, the ip is marked at the right edge with a phrase accent (L- or H-). The PW, on the other hand, is marked with a L tone at its left edge. The only other tonal event within the PW is the pitch accent (H*, !H*, or L*). Pitch accents are associated with the prominent syllables of PWs. A basic schematic for the prosodic structure of Turkish and an example pitch track are shown in **Figure 2** (Section 2.4.2 for details).

Figure 2 shows that in Turkish, there can be multiple prominence-lending and boundary marking tonal events occurring near each other (PW initial L tones, pitch and phrase accents, and boundary tones). Moreover, the realisations of these tonal events are sensitive to the organisation of prosodic structure. This is observed in the differences in tonal marking due to the position of ips relative to the ip that contains sentence level prosodic prominence, the nuclear ip (nip). For instance, pre-nuclear ips (prips) are usually marked with a H- phrase accent, whereas nuclear and post-nuclear ips (ptips) are usually marked with a L-, and the pitch range of pitch accents in nips is substantially compressed (!H*) (see **Figure 2b**).

The present study tests whether apexes are synchronised with any of these events with different functions and positions in the prosodic organisation. By adopting tonal events as anchors for apexes, we are able to test whether apexes are indeed synchronised with prominent pitch accents when other types of tonal events (e.g., boundary-marking events) are included in the analysis. We also test whether apex synchronisation shows any difference depending on the organisation of prosodic phrases, which has not been reflected in earlier studies on gesture-prosody synchronisation. Turkish presents an interesting prosodic context to test synchronisation

as there can be multiple tonal events occurring close to each other within a short duration in PWs. This crowding is potentially a curious challenge for synchronisation as it generates noise for the mechanism that manages synchronisation, especially in naturalistic data such as is used in the present study. However, consistent synchronisation with one of these tonal events over others in such conditions would indicate a systematic selection of particular tonal events for apex synchronisation, which resists contextual effects.



(a) The schematic. Capital letters indicate accented syllables at the surface level. Prosodic phrases and tonal events associated with these phrases have the same colour.

(b) The pitch track. The first four annotation tiers show IP, ips, PWs, and tonal events, respectively.

Figure 2: A schematic for the prosodic structure of Turkish shown on the utterance "There are kitchen cupboards above the fridge" and its pitch track.

Such systematic selection of anchors for synchronisation is inherently a cueing behaviour. For instance, if apexes are synchronised with pitch accents, as earlier studies would predict, then the locations of pitch accents can be predicted from the visual signal alone provided that there is a gesture accompanying speech. In other words, the systematic behaviour of apexes can be used in the identification of tonal events they are associated with, assisting phonological analyses. It is possible to show how apex synchronisation can shed light on intonational categories in a case study of a disputed feature related to accenting in Turkish.

1.2.2 A case study on accentlessness in Turkish

As noted in Section 1.1, earlier studies on apex synchronisation claimed a synchronisation of prominences in gesture and prosody. This is interrogated in Turkish in the present study. Turkish presents an interesting ground for this interrogation since its prosodic structure poses a challenge for this claim — there can be prosodic phrases without a pitch accent in Turkish. In such cases, there would be no tonal events that encode prominence, and therefore, there are no targets for apexes to be synchronised with. For instance, post-nuclear ips in Turkish do not contain pitch accents (see İşsever, 2003; Özge & Bozsahin, 2010; Section 2.4.2; and also Kabak & Vogel, 2001 for other prosodic processes that can lead to deaccentuation). Another example of phrases with no prominence pertains to lexical stress in Turkish. Two types of lexical stress are prescribed in Turkish: (1) Regular stress where the stress is on the PW-final syllable, and (2) irregular stress where the stress is on a non-final syllable (Sezer, 1981; Kabak & Vogel, 2001). In some studies, a pitch accent is associated with both types of lexical stress (Ipek & Jun, 2013). However, in others, words with regular stress are claimed to be accentless (Levi, 2005; Kamali, 2011; Güneş, 2013). In this case study, we concentrate on this claim and show how apex synchronisation can be employed to inform us about the presence or absence of accents in pre-nuclear ips in Turkish.

Pre-nuclear ips (prips) typically end in a pitch rise in Turkish (**Figure 2b**). In cases of regular stress, this rise is hard to dissociate from the hypothesised word-final pitch accent, which is also associated with a pitch rise. In other words, word-final pitch accents and phrase-final phrase accents coincide in words with regular stress, and therefore, it is unclear what the status is of the rise in pitch (i.e., whether the H tone is a part of the pitch accent or the phrase accent). There are two views regarding this issue. Ipek and Jun (2013) suggests that the H tone functions both as a part of the pitch accent and as part of the phrase accent. On the other hand, Kamali (2011) claims that the H tone is a property of the ip only, which in turn proposes that words with regular stress are accentless in Turkish (see Section 2.4.2 for details).

We claim that apex synchronisation can be used to shed light on this disagreement. In these prips, the synchronisation options (if at all) are limited to the PW-initial L and the H tone (see **Figure 3** for an illustration). The prominence-based synchronisation claim, which the present study interrogates, predicts that apexes would be synchronised with prominent pitch accents

since the L tones are only demarcative and do not lend prominence (see Section 2.4.2). *Assuming* that this claim is true for Turkish as well (which is tested first in this study), there are three possibilities:

1. If apexes are synchronised with the H tone (see **Figure 3a**), this implies that the H tone may actually be a part of the pitch accent as well as the phrase accent, supporting Ipek and Jun's (2013) double function claim. The presence of a prominent pitch accent at the location of the rise attracts apexes.
2. If apexes are synchronised with PW-initial L tones instead (see **Figure 3b**), this synchronisation pattern would also imply that synchronisation is not managed by prominence only. In the absence of pitch accents associated with PWs, apexes may be synchronised with PW-initial L tones which are also associated with PWs. This would suggest that apex synchronisation is sensitive to the prosodic hierarchy because synchronisation occurs only between the members of the PW. Namely, apexes do not synchronise with tonal events that are markers of higher level constituents such as phrase accents and boundary tones.
3. If apexes do not show any discernible synchronisation pattern (either the L tones or the H tone), this implies that there is no prominent pitch accent with which the apex can be synchronised. Therefore, synchronisation cannot be reliably predicted. This would support that the H tone is part of the phrase accent only, and not prominence-lending (Kamali, 2011; Güneş, 2013, 2015). This would also suggest no effect of prosodic hierarchy, as apex synchronisation floats between boundary events marking constituents at different levels.

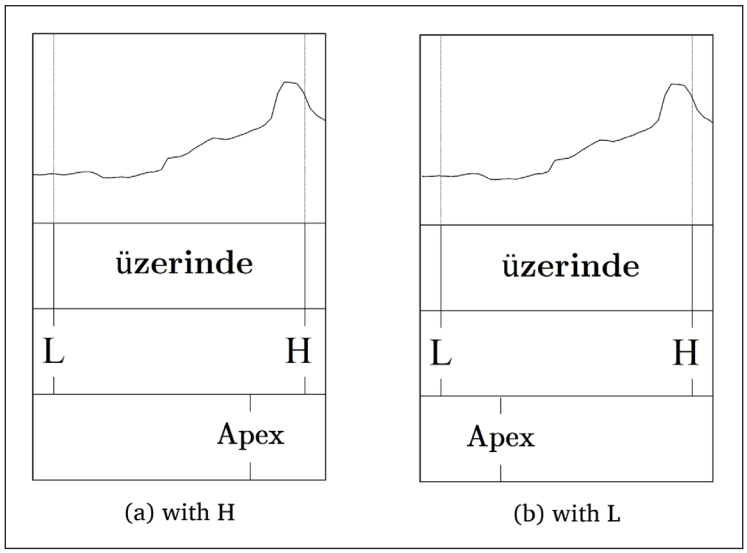


Figure 3: Two hypothetical gesture apex synchronisation cases with L or H on the word *üzerinde* 'above' in Figure 2b.

As can be inferred from these, consistent synchronisation with either of these tonal events (or lack thereof) can be used to identify whether words with regular stress are accentless or not in Turkish. Moreover, (a) synchronisation has implications for which levels in their organisation gesture and prosody are coupled with while also exploring apex synchronisation behaviour in cases where there is no prescribed anchor.

To summarise, the present study tests whether apexes are synchronised with pitch accents or with other tonal events in Turkish. The analysis examines synchronisation in different prosodic (phrase types such as pre-nuclear and nuclear) and gestural contexts (gesture types such as deictic and iconic) accounting for their potential effects on apex-tonal event timings (see Section 1.1). Following, systematic coordination of apexes with these events is used to determine what their role is as multimodal cues in the representation and identification of intonational categories in Turkish. Based on its findings, the study then moves on to demonstrate how these multimodal cues can be utilised in phonological analyses focusing on the disagreement about accentless words in Turkish.

1.3 What constitutes synchronisation?

Section 1.2 stated the aims of the present study and clarified the anchors in gesture and prosody for synchronisation. In this section, we introduce what constitutes *synchronisation* within this study while making comparisons with approaches adopted in earlier studies.

First, unlike many earlier studies (Section 1.1), the present study does not define a single pre-specified target (e.g., a target accented/edge syllable) with which apexes are assumed to be synchronised. Instead, apexes are free to be synchronised with any event in the prosodic signal, enabling us to capture the range of synchronisation behaviours in spontaneous data. Second, the study is interested in the synchronisation of points in time, not intervals (e.g., syllables and strokes). The analyses of synchronisation through overlaps of intervals with potentially incomparable durations can lead to less precise conceptualisation of synchronisation. For instance, the stroke of an iconic gesture can easily reach one second in duration, and comparing the overlap of any one syllable with such a stroke is not likely to return a meaningful result in general. This is further highlighted if any amount of overlap between intervals is considered as synchronisation. For example, with such an approach, only 1ms of overlap between a syllable and a stroke would count as synchronisation, although most of the syllable would fall outside of the stroke (the stroke would align with many other syllables, too). In contrast, an analysis of synchronisation of gestural points in time (i.e., apexes) with points in time that are associable with (but do not fully equate to) intonational categories (e.g., tonal events) provides a more refined look into gesture-prosody synchronisation at the smallest unit level, as intended in the present study.

Apexes and tonal events only need to consistently co-occur near each other as synchronisation is generally defined as the systematic co-occurrence of units in time. This, on the other hand, necessitates a measurable definition of "near" — how near these units should be in order to be considered synchronised. There is no set number in the literature that shows how near apexes and tonal events should be in order to be considered synchronised. However, there have been other studies that approached synchronisation in a similar manner (Section 1.1). For instance, Loehr (2004) considered two anchors synchronised only if they occurred within a pre-defined duration, 275ms, which was the average duration between gestural and prosodic markings in his data.

The present study adopts the same approach to synchronisation but uses the average syllable duration, 160ms, as its synchronisation criterion. This means that if the time distance between an apex and the nearest tonal event to it is less than the average syllable duration then they are considered synchronised. This duration was selected because it is phonologically meaningful: The syllable is the smallest phonological unit that can carry prominence, and prominence is the main concept that gesture synchronisation has been built on so far (Section 1.1). Therefore, it is plausible to look for temporal relationships within this duration. The use of such a duration is also plausible from the perspective of studies on the perception of synchronisation. For instance, Leonard and Cummins (2011) reported that apexes that occur up to 200ms before target syllables can still be perceived as prosodically aligned, which supports using an even shorter duration in the present study. To reiterate, 160ms is a duration that only describes what "near" is in the present study and not a claim of synchronisation within any one syllable. We argue that apexes and tonal events need not occur within the same syllable to indicate a cueing behaviour. Otherwise, an analysis can unnecessarily predict asynchronisation when the actual time distance between two anchors is as short as 10ms if there happens to be a syllable boundary between them. Moreover, tonal events themselves do not need to align within the syllables to which they are related (Prieto et al., 1995; Arvaniti et al., 2000). Therefore, points in time in another modality should also not be strictly expected to align within syllables in order to signal (temporal) relationships, especially in spontaneous speech.

Having a fixed duration as a synchronisation criterion enables statistical analyses (Section 2.5) where this duration sets a tolerance zone only large enough to predict meaningful timing relationships, ignoring spurious effects that might cause slight asynchronies. In addition, this makes it possible to determine the effect of gestural and prosodic contexts on synchronisation. As implied in Sections 1.1 and 1.2, we hypothesise that tone types (e.g., L, H*), prosodic phrase types (e.g., pre-nuclear, nuclear), and gesture types (e.g., iconic, deictic) may affect the time distances between apexes and tonal events. Further, the present study also considered information structure as a speech component that might have an effect on apex-tonal event synchronisation due to the strong link between information structure and prosody (see Kügler & Calhoun, 2020

for an overview). Accordingly, the study also tested whether apex synchronisation is sensitive to information structural categories such as topic, focus, contrast, and givenness (see Götze et al., 2007 for definitions). However, no such effects were observed; therefore, this will not be reported further in this paper.

2. Methods

The present study investigates apex and tonal event synchronisation using natural speech data. One issue with earlier studies in general has been that their designs were highly constrained in order to meet their experimental demands (see Rochet-Capellan et al., 2008; Roustan & Dohen, 2010; Leonard & Cummins, 2011; Rusiewicz, 2010; Rusiewicz et al., 2013; Esteve-Gibert & Prieto, 2013; Esteve-Gibert et al., 2017). The elicitation of gesture, in particular, can be considered as a major point of argument against the generalisability of findings in these studies. For instance, participants in such studies were instructed to gesture in a pre-selected form at a target on a screen, as well as to align their gestures with their verbal productions which were single sentences in isolation. In contrast, gestures that humans perform and observe everyday are typically produced unselfconsciously with clear communicative intentions. Consequently, these studies lacked ecological validity due to altering the spontaneous and unrestricted nature of co-speech gesture through heavy experimental manipulations. The present study addresses this concern by using natural speech data acquired through a narrative task.

2.1 Participants

Four (two female, two male) participants and one male confederate participant were recruited for the study in Ankara, Turkey. These participants were 18–26 year-old native speakers of Turkish who did not report they could speak any other language. The participants were naive as to the purpose of the study.

2.2 Design

The overall design of the study involved the participants watching five pairs of short videos (see supplementary files) and then one-by-one, recounting what they saw in the videos to the confederate who they believed had not seen the videos. The videos the participants watched contained basic daily activities, acted out by actors, such as reading a newspaper, which added up to form a story. Each video had a scenario and told a full story with a sense of completion. The videos were designed to elicit narratives and did not contain any elements that are likely to propagate other discourse genres. Whether different scenarios affected production and synchronisation of gestures was also tested in the present study (Section 2.5).

2.3 Procedure

The participants watched the videos one-by-one, and then recounted what they saw to the confederate listener immediately after watching the video. The participants were told that the confederate would have a task after the session based on what he understood from the participant's recounts of the videos. This was done to make the participant's task more meaningful by giving the task a purpose and to encourage them to include as much detail as possible in their recount in order to help the confederate with his task. The confederate also functioned to offer the participant a natural communication target instead of just asking them to talk to a camera. The interactions were video-recorded, and the confederate's speech was transcribed but not included in the analysis. Each filming session took about 30–40 minutes.

2.4 Annotation

Three hours of multimodal speech data were collected in total, which then was transcribed, translated, and manually annotated for gesture (in ELAN; ELAN, 2019), prosody, and information structure (in Praat; Boersma & Weenink, 2018). Not every utterance in the data was annotated, given the enormity of the annotation task. First, the utterances that were accompanied by gestures were identified. Amongst these, the utterances that had clear verbal disfluencies and those that were accompanied by aborted gestures were excluded. The full utterances to be annotated were sampled using a simple random sampling method (i.e., lottery method). The utterances selected by the random sampling were extended if needed so that they did not break in the middle of gestures (e.g., utterance annotations continued if the hands were still in gesture). Since an utterance can contain many more tones than a gesture can contain apexes, there was a natural imbalance between the number of tones and apexes in the data — not every tone had an apex accompanying it (Section 3.1). Overall, we transcribed roughly 25 minutes of video (out of three hours) containing 568 manual gestures. No connection between gesture, prosody and information structure was assumed while annotating them. Each set of annotations was carried out separately and without access to the other annotations so that any findings of correlation that might be found were genuine. Namely, the marking of tonal events was independent of the marking of apexes. Therefore, any associations between these could only fall out of the analysis of synchronisation (Section 2.5).

2.4.1 Gesture annotation

The annotation of gesture involved the annotation of gesture strokes expressing the semantic function of gestures (i.e., gesture type), and apexes within the strokes. These annotations were done making use of widely-used guidelines and descriptions in McNeill (1992), Kita, Van Gijn, and Van der Hulst (1998), and Loehr (2004).

A gesture stroke carries the meaning of gesture and is executed with maximum effort (McNeill, 1992). Gestures can be categorised into four types, depending on whether or not they exhibit a discernible meaning in their stroke and if so what kind (McNeill, 1992). *Iconic gestures (iconics)* have a close semantic relationship with co-expressive speech in that they represent the physical aspects of the information encoded in the speech (e.g., gesturing to describe the shape of a table as in **Figure 4a**). *Metaphoric gestures (metaphorics)* function in the same way as iconics, except they represent abstract contents (e.g., gesturing with open palms facing up to show "empty hands" which indicates uncertainty in **Figure 4b**). *Deictic gestures (deictics)* are pointing gestures indicating the locations of entities in space (e.g., pointing to side with the index finger as in **Figure 4c**). *Beat gestures (beats)* are flicks of the hand. They do not bear any semantic content themselves (e.g., a quick sideways flick as in **Figure 4d**), but have instead been shown to function as a visual highlighter (Loehr, 2004; Leonard & Cummins, 2011; Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016; Shattuck-Hufnagel et al., 2016; Shattuck-Hufnagel & Ren, 2018). Generally, earlier studies analysed only one type of gesture (Section 1.1). It remains unknown whether synchronisation with prosodic anchors is the same for apexes of different types of gesture. This possible effect is accounted for within the present study where synchronisation of apexes is compared across all gesture types (Section 3.2.5).

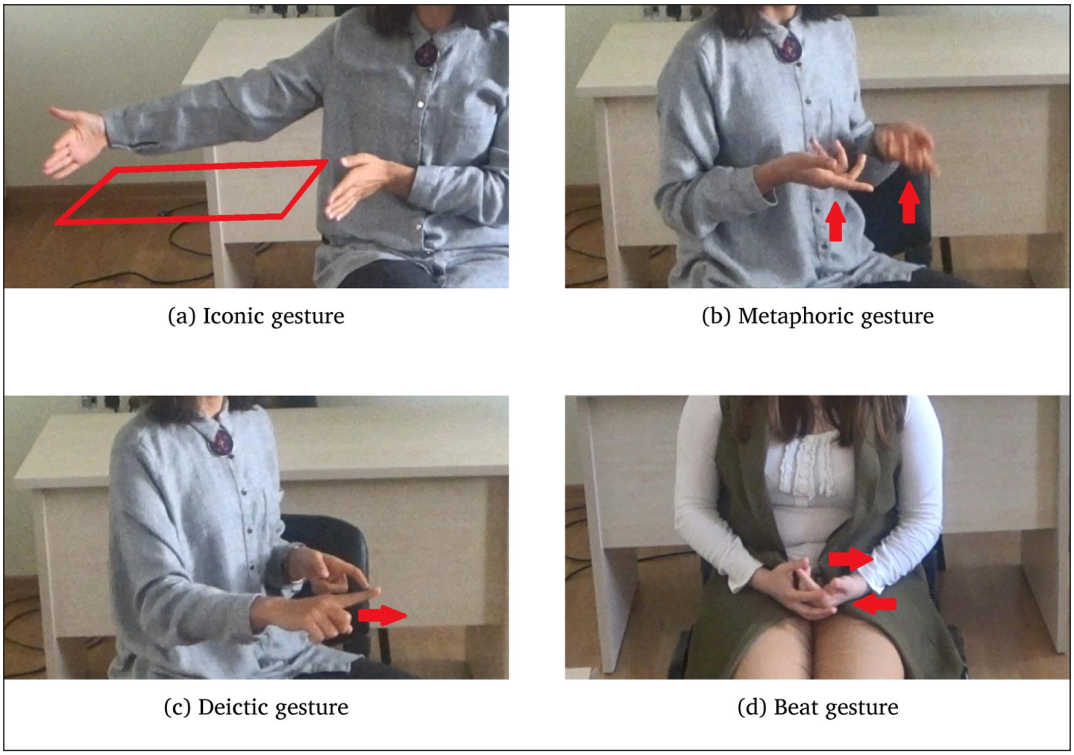


Figure 4: Categorisation of gestures according to their semantic function.

Apex Every stroke, regardless of its semantic function, contains an apex. The apex is "a single instant which could be called *the apex* of the stroke, the *peak of the peak*, the kinetic *goal* of the stroke" (Loehr, 2004, p. 89; also see Shattuck-Hufnagel et al., 2007). It is the dynamically most prominent point in time within the stroke. Within the present study, changes in direction within the stroke and the endpoint of the stroke were considered as dynamically prominent and marked as apexes (see one-segmented versus multi-segmented stroke distinction in Kita et al., 1998). For example, in a pointing gesture where the hand reaches its target location without any changes in its direction, the apex is the offset of the stroke (the endpoint of the movement) where the hand completes its extension. Kinematically more complex strokes can contain multiple directional changes followed by changes in speed. In these cases, each change in direction is considered as an apex (e.g., the index finger drawing an object with four corners to describe the shape of a table; Kita et al., 1998). At least one apex was identified for every stroke in the present study.

2.4.2 Prosody annotation

The prosodic structure of Turkish was introduced in Section 1.2.1, which also stated why the present study used a novel scheme for its annotations — our naturalistic data did not fit earlier descriptions well enough. The divergences from these descriptions based on read speech followed a broad phonetic approach and aimed to account for the variations we observed and to enable our analysis to delve into synchronisation patterns more comprehensively. Here, we describe this scheme which uses Tones and Break Indices (ToBI) conventions (Beckman & Ayers, 1997).

Prominence Prominence is signalled by pitch accents, which are typically associated with (although not always aligned with) the stressed syllables of prosodic words. Available pitch accents were H*, !H*, and L*. Pitch accents overwhelmingly surfaced as H* in our data. A H* was marked when there was a local pitch peak on the perceptually prominent syllable (regardless of regular/irregular stress distinction, see further below). This is similar to Ipek and Jun (2013). However, this differs from Kamali (2011) who argues that only words with irregular (non-final) stress get pitch accents which are H*L. In our data, there were very few cases where the fall after the pitch accent could not be explained by a following PW-initial L or L- phrase accent. These could be independently justified, and therefore, it is more parsimonious not to annotate a bitonal accent (also see Ladd, 2008, Ch. 3 for related theoretical disagreements in the use of H*(+)L vs. H*L- within and across a number of languages). This decision is further supported by cases where the Ls (of the claimed H*L) did not surface at all when there was a following H- (see non-finally stressed *radyoyu* in **Figures 7** and **9**). Accordingly, we did not annotate any bitonal pitch accents in order to achieve a more surface-transparent and minimalistic system of annotation. L* was marked when the pitch movement on the perceptually prominent syllable created a dip or low flat trend, which was often the local minimum (see Özge & Bozsahin, 2010 for different L* markings). We observed only a few examples of this on nuclear utterance-final verbs (Kamali,

2011 also mentions cases where the main prominence was produced on the verb as opposed to the pre-verbal element, p. 69).

The pitch range of pitch accents in the nuclear region was usually compressed compared to the pre-nuclear region (see Ipek & Jun, 2013). Often no peaks were observed, and a flat pitch plateau was maintained over the nuclear ip (e.g., the nuclear ip marked as "nip" in **Figure 5**). This lowering was marked with !H* at the pitch peak if there was a discernible peak; if not, it was marked at the midpoint of the prominent syllable's vowel. !H* could also be seen in the pre-nuclear area, and downstepping was considered as an event that can take place within and across intermediate phrases (see Ladd, 2008; see the pre-nuclear area marked as "prip" in **Figure 5**).

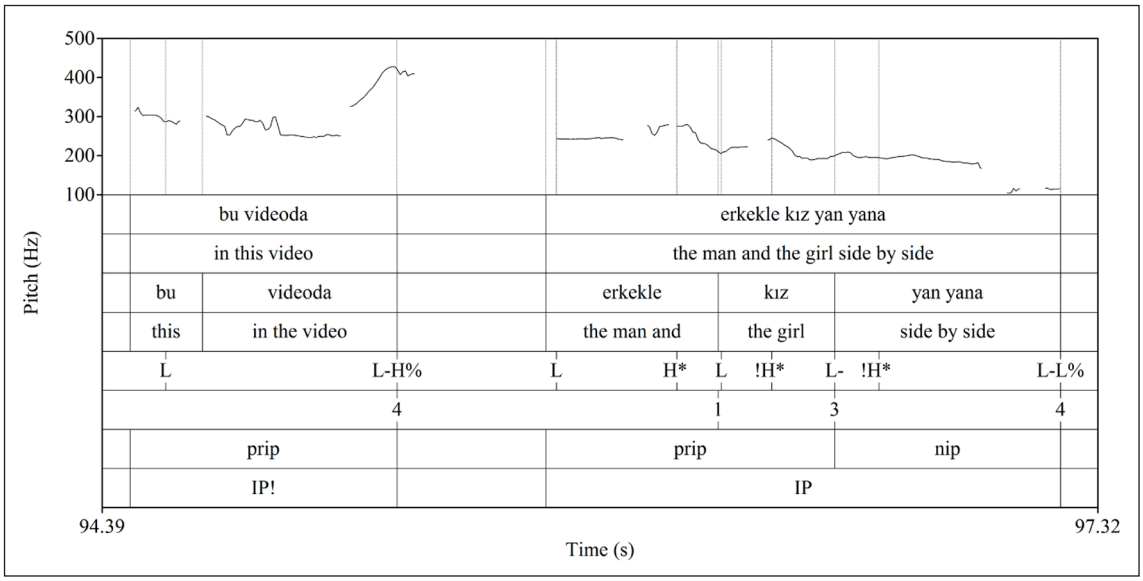


Figure 5: A pitch track showing a !H* in the nuclear ip with flat contour. !H* in "the girl" shows the lowering of the pitch peak from the preceding PW in the same pre-nuclear ip.

To reiterate, we observed that the word-final and non-final stress distinction in Turkish did not affect the type of pitch accent realised in the data. In addition, the annotation of pitch accents did not strictly depend on the prescribed lexical stress locations of words. That is, pitch accents were annotated exactly where they appeared on the pitch tracks, allowing shifts from their canonical locations (also shown in Özge & Bozsahin, 2010). For instance, a word may have regular word-final stress, but if the intonational cue indicating a pitch accent was on another syllable for any reason, the pitch accent was marked where the intonational cue was (see **Figure 6**, compare **Figure 5**). Similarly, if a word did not contain any cues to prominence on its syllables, a pitch accent was not marked, regardless of its prescribed stress location (see *bu*

videoda in **Figure 5**). This should not be interpreted as a claim against the dichotomy of lexical stress position in Turkish, but as a claim that cues to stress might be neutralised or shift away from the lexical stress positions. This is in line with studies showing that tonal events may not be aligned with syllables with which they are associated (Prieto et al., 1995; Arvaniti et al., 2000). The misplacement or omission of accents could happen through processes of deaccentuation (Kabak & Vogel, 2001), or other process related to the nature of the data. These are not detailed here as they are out of the scope of the present study. Overall, our practice only aimed to achieve surface-transparent annotations to capture apex synchronisation with intonational cues that are actually present in the signal.

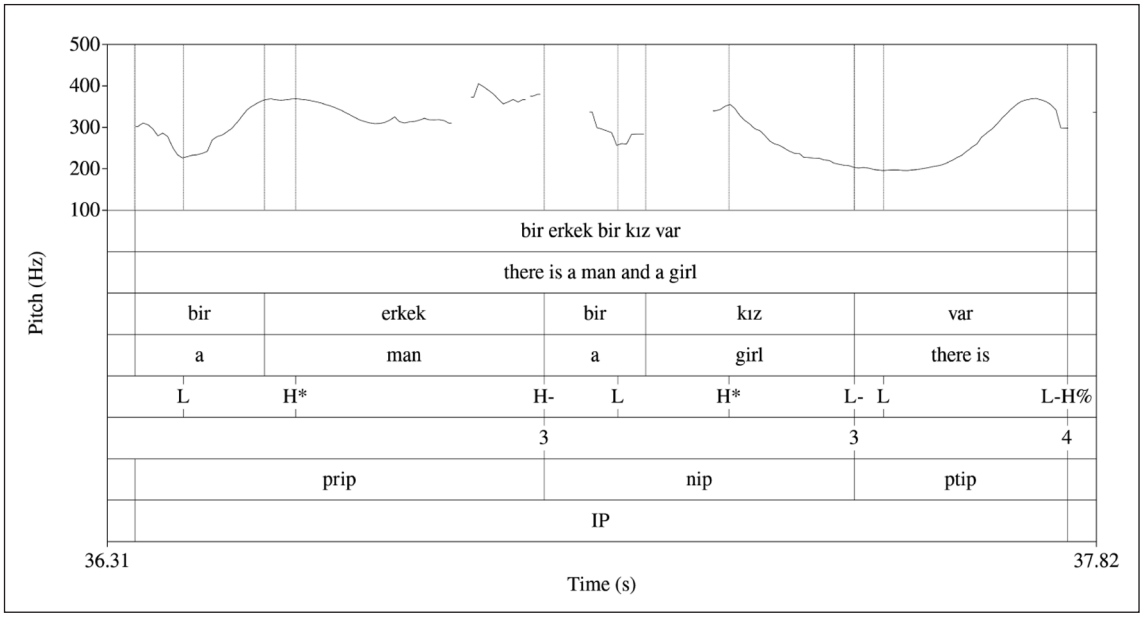


Figure 6: A pitch track showing an example for a misaligned non-final pitch accent on a word with regular (final) stress, *erkek* 'man'.

Phrasing Three levels of prosodic phrasing were adopted in the scheme: Prosodic Word (PW), Intermediate Phrase (ip), and Intonational Phrase (IP). *Prosodic words (PWs)* had typical inter-word boundaries and were the domains of word stress. PWs are marked on their left edges with an L tone at the lowest point in the first syllable of the PW (Ipek & Jun, 2013). These are similar to accentual phrase initial L tones in French (Delais-Roussarie et al., 2015). They are boundary events that do not bear any prominence and are always realised over the initial syllable, usually before any prominent pitch accent could occur. However, they may also not surface when the initial syllable is accented (also similar to French, see Delais-Roussarie et al., 2015). In Turkish, these L tones were first proposed by Kamali (2011) for the left edges of all

constituents except for the utterance initial ones (L- in her notation). In her account, a mid-range pitch level is kept throughout pre-nuclear and nuclear areas followed by a low plateau in the post-nuclear area (excluding for H*L and H-). The reset to the mid-range level after pre-nuclear final H- and the lowering into post-nuclear low-plateau are accomplished via these initial Ls. In our annotation, they are associated directly with PWs (Ipek & Jun, 2013) and annotated where there was low pitch word-initially. They serve to create a lowering contrast with the final tone in the previous word/phrase (e.g., H* to L if there is no ip boundary, H- to L at pre-nuclear and nuclear onsets, L- to L (lower) in post-nuclear onsets; see **Figures 7, 8**, and **9**).

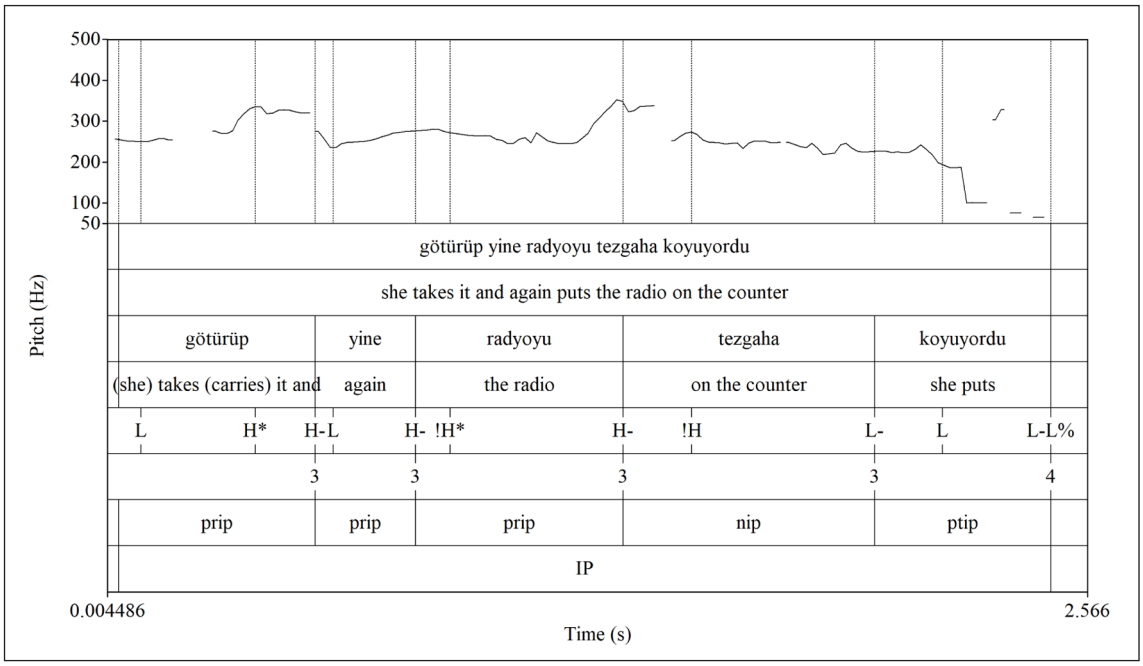


Figure 7: A pitch track that shows examples of one accentless and two accented pre-nuclear ips.

Intermediate phrases (ips) loosely correspond to syntactic constituents. They were marked on their right edges with either H- or L- phrase accents. *Pre-nuclear ips (prips)* typically exhibited a pitch rise at their right edge, which was marked with H- (see in **Figure 7** and the pre-nuclear area in **Figure 8**). Ipek and Jun (2013) also describe bitonal phrase accents, LH-, marking the right edge of prips with non-final stress (a subcategory of this, LHn, marks the left edge of the nuclear word). This bitonal marking can be compared with Kamali's (2011) annotation of H*L in words with non-final stress (in prips). That is, both Ipek and Jun (2013) and Kamali (2011) observed a sequence of H L H in these cases. Kamali (2011) seems to assume that L is part of the pitch accent (H*L H-) but Ipek and Jun (2013) assumes that it is part of the phrase accent (H* LH-).

The contrast here is abstract and not immediately clear from the contour. Therefore, further research is required. As stated previously, we did not reliably observe the L tones in these cases in our data. Therefore, given our approach and goals, these were not annotated. However, we sometimes observed steeper increases in pitch at the right edge of prips (often before the nucleus, see Ipek & Jun, 2013), which may be conceived as LH- (see **Figures 7, 8**). However, given our broad phonetic approach, we did not assume the presence of a L tone here (i.e., a bitonal phrase accent), but rather acknowledged that these phrase accents had higher pitch values compared to others in our marking (using ˆH-). This was done with the initial assumption that they could potentially reveal distinct synchronisation patterns, which was not case in our data (see Section 3.1). Therefore, they are not discussed in the rest of this paper.

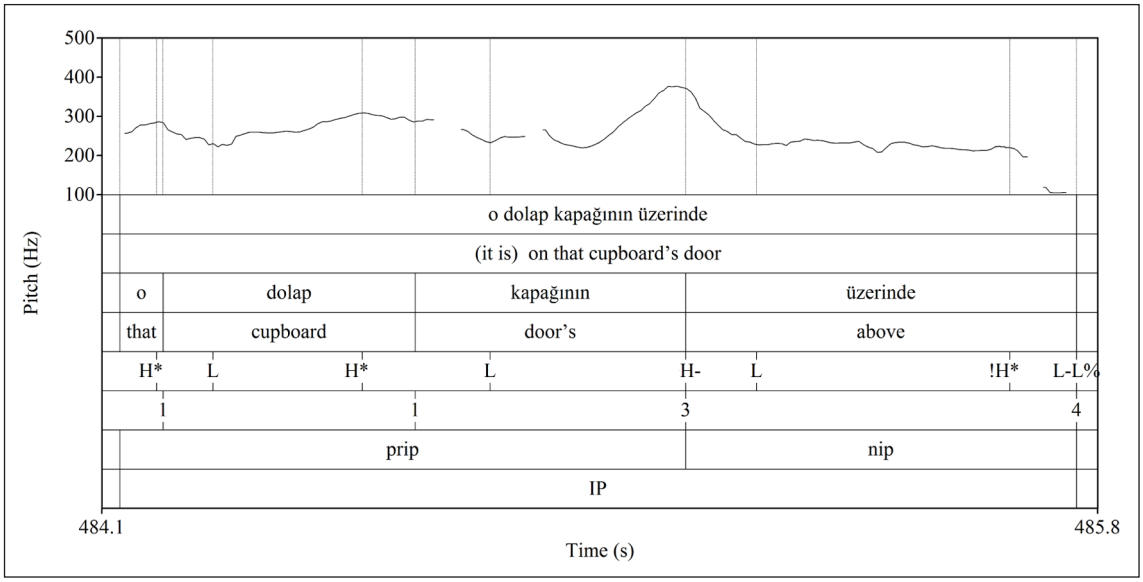


Figure 8: A pitch track showing pitch accents exhibit different pitch movement from raised H-. Pitch movement on pitch accents is not as sudden or steep.

As explained in Section 1.2.2, there were cases where a word-final H* and an ip-final H-coincided if the ip-final word had regular stress. In these cases, it was not clear if the H tone was part of the pitch accent or the phrase accent. It was also mentioned that there are two views in the literature on this issue. One claims that the H tone is a part of the pitch accent that also functions as a phrase accent, and the other claims that the H is an independent event and is a property of the ip (see Section 1.2.2), implying that regularly-stressed words in Turkish are accentless. The present study remains agnostic on this issue since our approach is surface-driven (i.e., annotations supporting either claim were possible in our scheme). In line with our pitch accent annotation, if there was no perceptually salient pitch accent in the final PW of an

ip, the final rise was marked with only H- (see *yine* in **Figure 7**). If there was a perceptually salient pitch accent, the ip was still marked with H- but H* was also marked at the pitch peak location within the final syllable (see *götürüp* in **Figure 7**) rather than together with the phrase accent at the ip boundary (cf. Ipek & Jun, 2013). In these cases, the evaluation of perceptual saliency was also based on the intensity and duration of the syllable. Intensity and duration have also been claimed to be correlates of stress in Turkish, although not as robust as pitch (Levi, 2005). However, this was the case only when averaging across final and non-final stress positions without controlling for syllable structure. In contrast, in Vogel, Athanasopoulou, and Pincus (2016), intensity appears as a strong classifier of stress (especially for final syllables) when compared across final/non-final and focus/non-focus positions (F0 is again the most reliable correlate of stress but not as successful in stress classification as in languages with unpredictable stress). Our scheme follows this: If intensity and duration values on the syllable with the H tone were distinct from those of the syllables in the immediate environment, a pitch accent was marked at the pitch peak. Regarding the marking of pitch accents in prips, we observed a pattern where if the pitch peak was higher and steeper, it was generally a phrase accent — peaks related to pitch accents are subtler (i.e., lower and more gradual increases, see **Figures 5, 7, 8**).

To summarise, the present study allows for both accented and accentless phrases. However, the reasons for this surface difference are unknown to us at the moment (but see Kabak & Vogel, 2001 for certain phonological processes at play). One explanation might have to do with stress deafness in languages with predictable stress patterns such as Turkish. That is, speakers of these languages cannot reliably perceive stress in other languages (Altmann, 2006) as well as in their own language (Domahs, Genc, Knaus, Wiese, & Kabak, 2013). Possible effects of this on production have also been postulated (Vogel, 2020). Namely, speakers of languages with predictable stress may produce reduced and less precise (thus more variable) cues to stress since the listener already knows where the stress is (in turn, the lack of reliable acoustic information leads the listeners not to have a clear basis for the identification of stress). We believe that this potential effect might be highlighted in our naturalistic data in which lexical stress cues were much more variable than the established final/non-final stress distinction in Turkish would suggest. In general, Vogel et al. (2016) summarise our observations regarding stress: "Given the particular subtlety or absence of stress cues in Turkish, it appears that Turkish may have undergone, or is currently undergoing, a loss of stress as a lexical phonological property" (p. 161).

Nuclear ips (nips) had an L- phrase accent on their right edge, which earlier studies did not describe. As mentioned above, Turkish shows a lowering terracing pattern in its nuclear to post-nuclear area, and in some earlier studies, perceived sentence-level prominence was claimed to

rely on this juncture (Kamali, 2011). Moreover, this lowering into the post-nuclear area was subsumed under the same L tone, marking the onset of pre-nuclear and nuclear phrases/words (Kamali, 2011; Ipek & Jun, 2013). In our annotation, the most prominent word in a sentence received a nuclear accent, with a compressed pitch range (Ipek & Jun, 2013). In addition, the lowering after the nuclear area did not require a post-nuclear element and might start after the nuclear accent within the nuclear ip (**Figure 8**, also see verbs with negation in Kamali, 2011). Therefore, this lowering was associated with the right edge of the nuclear ip and marked with L- in our scheme.

H*, !H*, and L* accents could all be seen in this area (cf. Kamali, 2011; Ipek & Jun, 2013; also see Özge & Bozsahin, 2010), but the most common accent was !H*. There was usually very little pitch movement over nips (Ipek & Jun, 2013). The most common contour for the nip was L !H* L-, which exhibited an intermediate level flat pitch plateau (see **Figure 8**). The identification of the perceptually salient syllable in these cases also relied on intensity and duration in cases where there is little movement. For instance, in **Figure 9**, typically non-finally stressed word *radyo* does not show a pitch movement to mark prominence on the stressed syllable, but rather a local intensity peak (dashed lines) that is higher than others (~8dB higher in this case), and therefore was marked with a !H* on the syllable aligning with the peak.

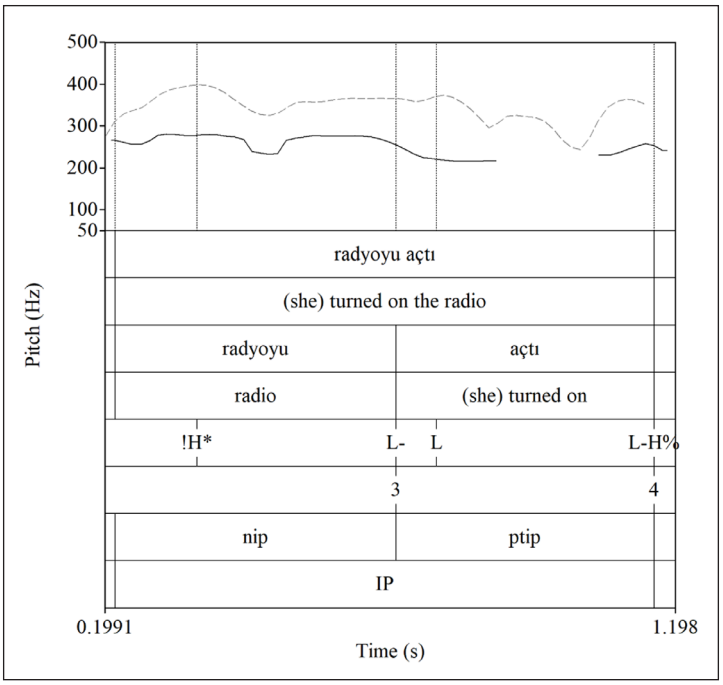


Figure 9: Non-finally stressed word *radyo* 'radio' displays an intensity peak on its stressed-syllable (dashed line).

Post-nuclear ips (ptips) had an L- phrase accent on their right edges, which earlier studies also did not describe. The motivation to mark L- mainly came from longer utterances with narrow focus at utterance-initial positions which were followed by tails or post-focal material (a.k.a. background) (Vallduví & Engdahl, 1996). These tails go through post-focal deaccenting (by definition, see Özge & Bozsahin, 2010), but they are not always dephrased — there could be several ptips that were marked by junctures at their right edges. In our analysis, it was possible that apexes could occur within these areas and be synchronised with such boundaries. Therefore, we included these markers in our scheme. The PW-initial L in the first ptip (following L- in nips) often brought the pitch to the bottom of speaker's range. The most common contour for these ips was L L- showing a low level plateau.

Intonational phrases (IP) are the largest phonological units (that utterances can be broken into) that have their own intonation patterns (i.e., contours). They were marked at the right edge with a boundary tone, either L% or H%, which coincided with the offset of the final intermediate phrase within the IP. This led to combinations of L-/H- with L%/H% (e.g., L-H%). All possible combinations of these were observed in the data.

2.4.3. Annotation reliability

All the annotations in the present study were done by one annotator (the first author). As a test of annotation reliability for prosodic annotation, a different annotator (the second author) did a blind annotation of ~four minutes (~20% of total annotation duration excluding the confederate's speech) following the guidelines provided in Section 2.4.2. The utterances for the blind annotation were equally sampled from all participants and were randomly selected. Then, using the kappa statistic (κ), both sets of annotations were tested for pairwise agreement corrected for expected chance agreement (Carletta, 1996). For each annotated word (N = 459), it was tested whether (1) there was a PW-initial L, (2) there was a pitch accent and on which syllable, (3) there was a phrase accent. The results are presented in **Table 1**. $\kappa > 0.8$ is claimed to show "good reliability" (Carletta, 1996), and as can be seen in the table, this was achieved for all the tonal events tested.

Tone	κ
PW-L	0.890
Pitch.acc	0.808
Phrase.acc	0.801

Table 1: κ for the pairwise agreement of tonal event annotations.

To test the annotation reliability for apex annotation, a different annotator was trained (following Section 2.4.1) and asked to do a blind annotation of apexes on 103 randomly selected

gesture strokes (at equal numbers from all participants, 20% of gesture phrases in the data). The timewise agreement of apex markings in the two sets of annotations were tested based on the proximity of markings to each other. That is, if the apex markings were within five frames of each other (80ms either way, i.e., half a syllable), this was considered as agreement. The annotators were in agreement 89% of the time, which can be considered as substantial agreement. The κ statistic could not be used here because there was no fixed external units to compare both sets of annotations to (e.g., words).

2.5 Analysis of synchronisation

The general approach to synchronisation in the present study was described in Section 1.3. In this section, we detail the exact steps of associating and determining synchronisation of apexes with tonal events. There were two main steps in the study: (1) The pairing of proximate apexes and tonal events; (2) the testing of the synchronisation of these pairings.

Before a test of synchronisation, tonal events and apexes needed to be associated with each other. This comprised the first step of the analysis. Here, the nearest tonal event to each apex was identified — these formed a *pairing* (note that not every tonal event event had an apex accompanying it). The pairing process was also sensitive to the semantic relation of gesture to speech. The semantic relatedness was sought because it has been shown that semantic relations between gesture and speech can affect gesture synchronisation, constraining the synchronisation of other speech components such as prosody (Bergmann, Aksu, & Kopp, 2011). This is what is predicted in most psycholinguistic models of integrated speech-gesture production as well (Wagner et al., 2014). In these models, there is no requirement that semantic association happens strictly at the lexical level. In fact, since speech and gesture are typically represented to come from a common origin at a higher conceptual level, the semantic association should happen before any lexicon-related process (see Kita & Özyürek, 2003 for an example model). This is also supported by findings that gestures do not always co-occur with their lexical affiliates, nor is it easy to identify a single lexical affiliate for a single gesture (Graziano, Nicoladis, & Marentette, 2020). In our data, the semantic content of a single gesture was often expressed across multiple prosodic phrases (ips). Therefore, another domain where gestures can be semantically associated with speech was needed. In our study, we used information structural units (e.g., topic or focus) for this purpose. These provide a plausible frame for pairing because (1) their boundaries are sensitive to prosodic boundaries (Kügler & Calhoun, 2020); (2) they often span over multiple prosodic phrases; (3) they maintain a meaningful communicative and discursive association between gesture and speech (Türk, 2020). To summarise, an apex was only paired with the nearest tonal event within the information structural unit carrying the same semantic content as the accompanying gesture. Since beats do not bear any semantic content, beat apexes were paired with the nearest tonal event without seeking semantic relatedness.

The second step involves testing whether paired apexes and tonal events achieved synchronisation as per the synchronisation criterion where they were only considered synchronised if they occurred within an average syllable duration of each other, i.e., 160ms (Section 1.3). For this, the analysis first calculated the time distance between the paired apexes and tonal events. It was predicted that there could be potential effects of gestural, prosodic, and information structural contexts on these time distances (Section 1.3). To test this, the study employed linear mixed-effect regression modelling (*lme4* package in R, Bates, Mächler, Bolker, & Walker, 2015). The model tested whether the calculated time distances were affected by (1) tone type, (2) ip type, (3) gesture type (as well as two-way interactions between these factors), including participant and scenario as random effects. Insignificant effects from this model were then eliminated using backwards elimination (using *lmerTest* and *rms* packages, Kuznetsova, Brockhoff, & Christensen, 2017; Harrell Jr, 2019 in R), resulting in a final model.

This final model was fitted to get the estimates of time distances accounting for the effect of the factors included in the model. In order to see whether synchronisation was achieved given the significant effects of these factors, equivalence tests (TOST procedures) were employed using these estimates (Lakens, 2017). In an equivalence test, the estimate is compared against two pre-specified equivalence bounds (a lower bound and an upper bound) using two one-sided t-tests. The area between these bounds defines a tolerance zone and the observations that fall within this zone are considered statistically equivalent. In line with the synchronisation criterion introduced in Section 1.3, the present study used the average syllable duration, 160ms, to define these equivalence bounds. This meant that 160ms on either side of zero (as the perfect synchronisation condition) set the equivalence bounds (i.e., lower bound -160ms and upper bound +160ms). In the test, if the confidence intervals of the estimates occurred within ±160ms equivalence bounds, then the estimates were statistically equivalent to zero and synchronisation was achieved. Using this procedure, each estimate acquired from the regression model was tested for equivalence.

3. Results

The present study investigated whether apexes were synchronised with tonal events in spontaneous speech data in Turkish. In Section 2.5, the steps of this investigation were detailed. The presentation of its results follows the respective order of these steps. First, we report the pairing pattern of apexes and tones (i.e., which tones apexes were the nearest to in the data and whether the pattern shows variation depending on prosodic contexts (Section 3.1)). Second, in Section 3.2, we give the overall distribution of time distances between paired apexes and tones, and then test whether these time distances were affected by prosodic and gestural factors. This is followed by the results of equivalence tests showing whether synchronisation was achieved by these pairings taking into account these factors (Section 3.2.3). Finally, we focus on what these

synchronisation results mean for our case study on accenting in pre-nuclear phrases in Turkish (Section 3.3).

3.1 Which tones do apexes tend to be paired with?

A pairing consists of an apex and the tonal event that is nearest to it. In this section, we report whether apexes were paired with certain types of tones more than others, indicating a pattern. In Section 2.4.2, tonal events marking prominence or boundaries in Turkish were described (e.g., H-, H*). To give a general look into the data and the pairing behaviour, we first group them under four main categories (one for prominence and three for prosodic phrases): Pitch accents, phrase accents, boundary tones, and prosodic word-initial low tones (L hereon). These are referred to as tone types. Grouped by tone type, **Table 2** shows the total number of tones annotated for each type (Annotated N), what number/percentage of these annotated tones were paired with apexes (Paired N/%), and what percentage they constituted out of 820 pairing instances (Frequency %).

	Annotated N	**Paired N (%)**	**Frequency %**
Pitch.acc	1030	374 (36.31%)	45.61%
Phrase.tn	687	107 (15.57%)	13.05%
Boundary	679	61 (8.98%)	7.44%
L	1301	278 (21.36%)	33.90%
Total	3697	820	

Table 2: Total number of tones (Annotated N), the frequency at which they paired with apexes (Paired N/%), and what proportion they constitute out of all pairing instances (Frequency %).

In the data, 22% of tones were paired with apexes within intonational phrases that were gestured; 36% of pitch accents were paired with apexes, and these pairings constituted nearly half (46%) of all pairing instances. Most apexes were paired with pitch accents, closely followed by Ls; 21% of Ls were paired, which constituted 34% of all pairing instances. Based on these numbers, it is possible to argue that apexes tended to be paired with pitch accents as well as with Ls. Therefore, the claim that apexes are synchronised only with pitch accents (as the only prominence markers) was not fully supported here. However, in line with our arguments in Section 1.1, it is possible that the pairing patterns (hence synchronisation) is affected by prosodic and gestural contexts — there might be different modes of pairings that alternate depending on these contexts within which tones and apexes existed.

Our next analysis looked for potential modes of pairing and revealed a bimodal distribution where apexes were paired with pitch accents when a pitch accent was available in the PWs. However, when there was no pitch accent available, Ls were greatly favoured for pairings (not

every phrase that a gesture may accompany has a pitch accent, see Section 1.2.1). In fact, 80% of the pairing instances with Ls in the data occurred when there was no pitch accent in the PWs. To demonstrate, **Table 3** breaks down the distribution in **Table 2** by whether or not there was a pitch accent in the PW that the paired tone was in. The bimodal distribution sensitive to the presence of a prominence marker is evident in the table. This change in the pairing patterns establishes an effect of prosodic context on apex timing behaviour.

	No pitch accent	**Pitch accent**
Pitch.acc	NA	71.7%
Phrase.tn	15%	11.4%
Boundary	8.2%	6.1%
L	76.9%	10.8%
n	294	526

Table 3: Is there a pitch accent in the prosodic word that the paired tone is in?

Apex pairings with Ls marking PW onsets imply that apexes are sensitive to prosodic phrasing. Note that both pitch accents and Ls are associated with the PW (Section 2.4.2). The pairing preference shifts from one to the other, depending on the accentlessness of the PW — apexes were not paired with phrase accents and boundary tones, which are associated with phrases at higher levels in the prosodic hierarchy. This suggests that apex synchronisation is also sensitive to prosodic hierarchy.

Our analysis also tested whether the pairing patterns change depending on: (1) Subcategories of four tonal events (e.g., L- vs. H- and ˆH- for phrase accents, or !H* vs. H* for pitch accents including final/non-final distinction); (2) Available tones in the prosodic phrases (e.g., L H* H- vs. L H- vs. L); (3) Intermediate phrase type that the paired tone is in (e.g., pre-nuclear ip vs. nuclear vs. post-nuclear); (4) Gesture types (e.g., iconic vs. deictic). However, we observed no sensitivity to any of these factors.

A reviewer noted that the association of 21% of apexes with phrase accents and boundary tones seems surprisingly high (**Table 2**). In fact, some earlier studies reported ∼10-20% of alignment with unpredicted anchors (e.g., 17% in Loehr, 2004) and some did not report what happens in cases of non-alignment with the pre-selected anchor (e.g., ∼17% non-alignment with stressed syllables in Shattuck-Hufnagel & Ren, 2018). We argue that the number of pairings with these was too low to constitute an overall pattern and that this much variation is to be expected in studies using natural data.

To summarise, there is a bimodal pairing pattern that alternates between pitch accents and Ls depending only on the accenting of phrases. This finding has implications for accented/

accentless pre-nuclear intermediate phrases in Turkish (Sections 1.2.2 and 2.4.2). This will be addressed in Section 3.3 together with the findings of the synchronisation analysis in the next section, where we report whether the paired apexes and tonal events achieved synchronisation given the synchronisation criterion adopted in the present study.

3.2 Are apexes synchronised with their nearest tone?

The findings in Section 3.1 have shown pitch accents and Ls as the preferred targets of apex pairings, and phrase/boundary tones as dispreferred targets. As explained in Section 2.5, pairings indicate a proximity and semantic based relation of an apex with a tonal event. Synchronisation, on the other hand, deals with the actual measurements of time distances between paired units. In Section 2.5, the present study introduced a synchronisation criterion in order to define what it considers "synchronised". Namely, if the time distance between the paired units is less than the average syllable duration (160ms), then the pairing achieves synchronisation; if not, the members of the pairings are not synchronised. In this section, the present study first gives an overview of measured time distances between paired apexes and tonal events. Then, it tests whether these distances were affected by prosodic and gestural factors (Section 2.5), and finally tests whether the pairings achieved synchronisation given the effects of these factors.

3.2.1 Overview of time distances between pairs

The pairing patterns presented in Section 3.1 indicated preferences depending on the tone type. It may also be the case that the time distance between the paired units may be greater or smaller depending on the tone type. In order to get a general view, **Figure 10** shows normalised (scaled) distributions of time distances between the paired units for each tone type. Phrase accent and boundary tone pairings were grouped together because annotations of boundary tones coincide with phrase accents (Section 2.4.2), and there were not enough observations for these tones to enable meaningful comparisons separately.

In the figures, the mean distance for pitch accents and Ls were very close to zero. The distribution for phrase/boundary tones showed a higher mean and standard deviation, exhibiting a lead of tones over apexes about a syllable duration (i.e., tones occur before their paired apexes). For each type condition, most of the observations were between $\pm$200ms range. The distribution for pitch accents in **Figure 10a** showed the most compact peak followed by Ls in **Figure 10b**, which had a slight spread into the positive direction on the x-axis. The distribution for phrase/boundary tones had smaller peaks further away from the main peak, reaching time distances as far as 800ms. Following on this, it was checked whether these small peaks and spreads were indicators of bimodality using Hartigan's dip test (Hartigan, Hartigan, et al., 1985). However, no significant evidence of bimodality was found (a: $D = 0.018$, $p = 0.481$; b: $D = 0.015$, $p = 0.935$; c: $D = 0.016$, $p = 0.992$).

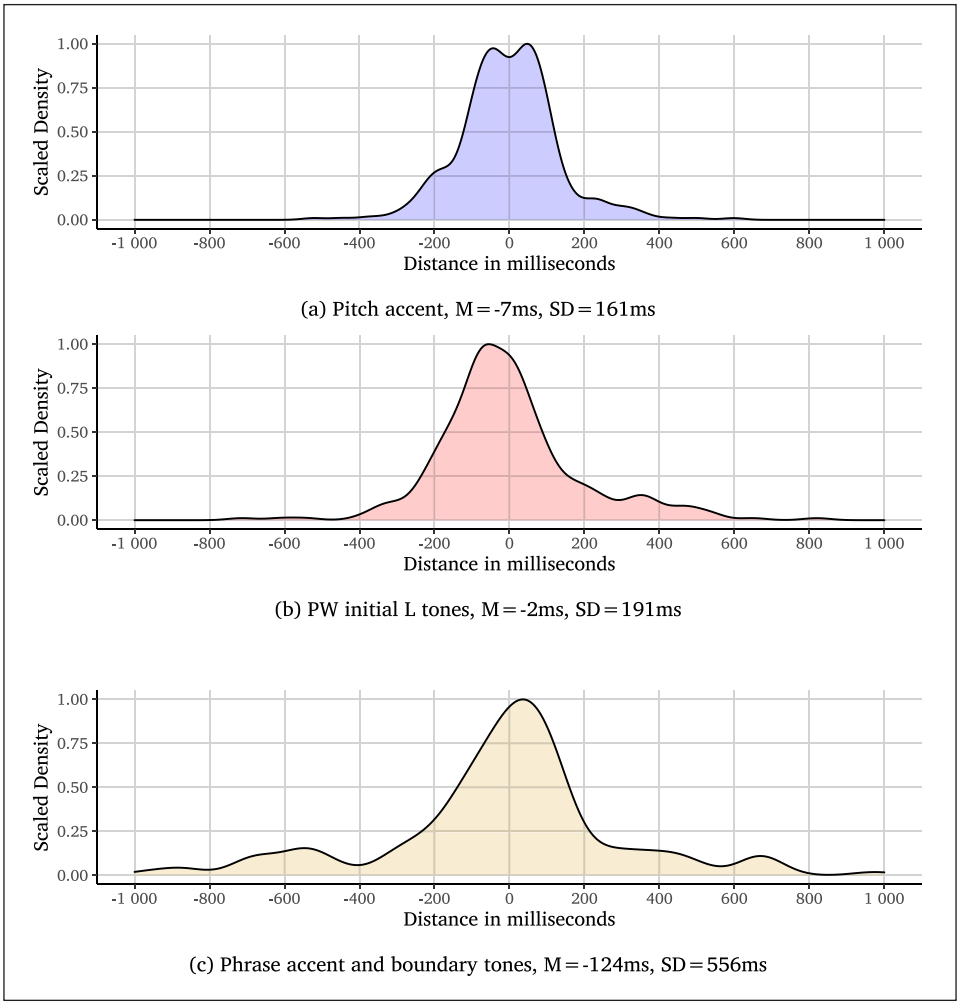


Figure 10: Time-normalized histograms of the time distances of the nearest tone from an apex.

Overall, these findings agree with the pairing preferences in that the time distances between apexes and the preferred tonal targets (pitch accents and Ls) were closer to zero than the dispreferred targets (phrase accent and boundary tones). Therefore, the type of tones involved in the pairings seems to be a factor that affects the time distances between paired units. In Section 1.1, we also predicted that there might be gestural, prosodic, and information structural contexts that can affect these measured time distances. The next section reports our findings of this analysis.

3.2.2 Factors affecting time distances between pairs

As described in Section 2.5, the next analysis tested whether (1) tone type, (2) ip type, (3) gesture type as well as participant and scenario information had an effect on the measured time distances

between apexes and paired tones. Two things must be noted regarding this full model. First, within the tone type factor, grouping of phrase accent and boundary tone pairings (as per Section 3.1) resulted in three levels of tone types: Pitch accent, L, and phrase accent and boundary tones (referred to as "phrasal tone" from hereon). Second, pairings within the ptip level from the ip type factor were excluded because of lack of data — the pairings tended not to take place within these areas. This led to an overall exclusion of 102 pairing instances in total.

Next, the full model was put through a backwards elimination process of insignificant effects (Section 2.5), which revealed the (1) tone type, (2) the interaction of tone type and ip type, and (3) the interaction of tone type and gesture type as factors that significantly affected the time distances between paired apexes and tonal events. **Table 4** shows the matrix of significant effects in the final model.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Tone type	2	1.48	0.74	13.96	0.0000
Tone type:ip type	3	0.80	0.27	5.06	0.0018
Tone type:Gesture type	9	1.17	0.13	2.45	0.0093

Table 4: The matrix of significant fixed effects on time distances between apexes and tones.

The synchronisation analysis in the present study is not directly interested in the results of this model, but in the estimates of time distance between the paired units under these significant effects. These estimates were then used in the synchronisation tests and equivalence tests (Section 2.5), which are reported in the next section.

3.2.3 Tests of synchronisation

Within the present study, and apexes and tonal events were synchronised if they occurred within the average syllable duration, 160ms, of each other (see Section 1.3), taking into account the significant contextual effects (**Table 4**). This was tested using equivalence tests (Section 2.5). In the tests, if the confidence intervals of the estimates occurred within ±160ms (i.e., equivalence bounds), then the synchronisation was achieved. The results of each significant effect in **Table 4** are presented in Sections 3.2.4, 3.2.5, and 3.2.6.

3.2.4 Effect of tone type

In **Figure 11**, the equivalence test output is plotted for the significant effect of tone type (see **Table 5**) where the dashed lines indicate the equivalence bounds at ±160ms. Note that the results of this simple effect on the time distances present a general view of synchronisation for apexes and tones, showing the overall (i.e., standard) synchronisation pattern of apexes and

tonal events. This is because the estimates plotted here were averaged over the levels of gesture type and ip type effects (Sections 3.2.5, and 3.2.6) since the tone type was the common variable in all significant terms.

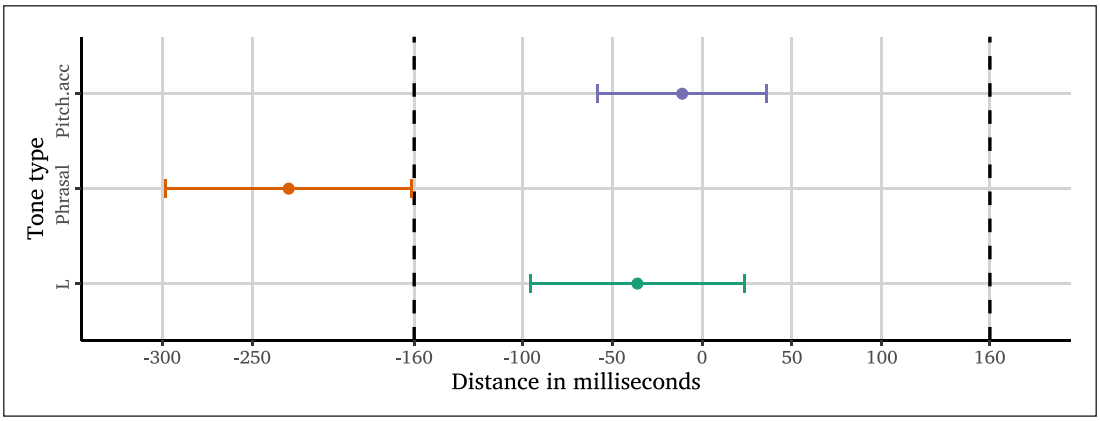


Figure 11: The estimated means of tone type term with the CIs at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

	t	df	p
Pitch.acc	6.226	373	<0.001*
L	4.109	216	<0.001*
Phrasal	–2.001	126	0.976

Table 5: The equivalence test results for apex-tone pairings for each tone type.

As can be seen in the figure, the estimates for pitch accents and Ls were one to two frames away from zero (–9ms and –34ms). However, the estimate for phrasal tones was at –229ms, falling beyond the lower bound. Consequently, the equivalence test results showed that the time distances of pairings with pitch accents and Ls were statistically equivalent to zero, and therefore synchronisation was achieved. However, the same was not true for phrasal tones. These tones tended to start more than an average syllable duration earlier than apexes. Therefore, no equivalence was observed: Apexes were not synchronised with these tones when they were paired with them.

3.2.5 Effect of gesture type

Figure 12 plots the equivalence test output for the significant interaction of tone type and gesture type. The estimates of phrasal tones in iconics (N = 225) and deictics (N = 280) showed an average lead for tones of about –252ms in iconics and –226ms in deictics not fitting

between the equivalence bounds. For pitch accent and L pairings, the tone lead was a lot less with –11ms/–43ms in iconics and –9ms/35ms in deictics. From the equivalence test, it can be concluded that for iconic and deictic gesture apexes, the time distances of pairings with pitch accents and Ls were statistically equivalent to zero (see **Table 6**), meaning that apexes were synchronised with these tonal events. However, for phrasal tones, the synchronisation was not achieved.

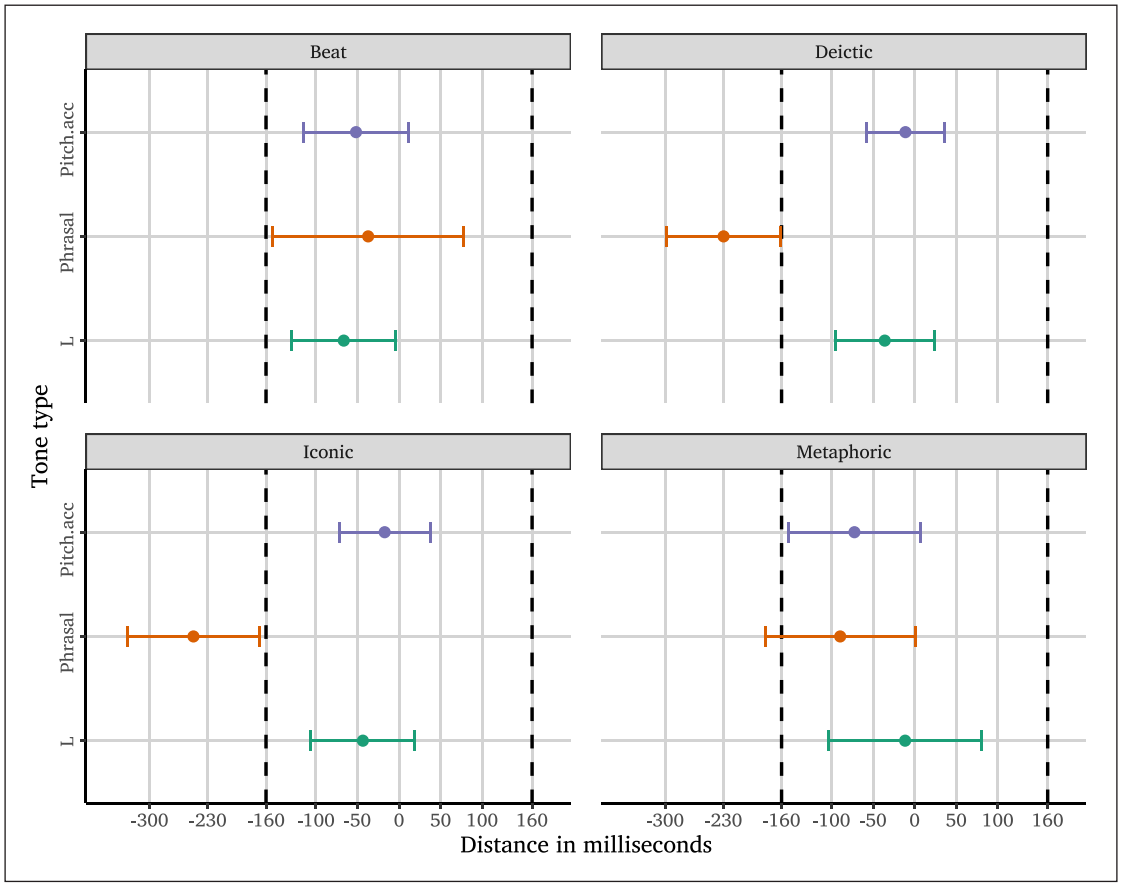


Figure 12: The estimated means of the Tone type : Gesture type interaction with confidence interval at 95%. The dashed lines are the upper (160ms) and lower (–160ms) equivalence bounds.

For metaphoric (N = 82) and beat apexes (N = 132), all estimates were within the equivalence bounds ranging between –89ms and 58ms, and the equivalence test were significant for all conditions except for the pairings of the apexes of metaphoric gestures with phrasal tones (see **Table 6**). These findings meant that beat apexes were synchronised with their nearest tone regardless of type. For metaphorics, the pairings with pitch accents and Ls satisfied the synchronisation criterion, but phrasal tones did not because the lower bound was crossed for

these (see **Table 12**). However, despite this, the present study interprets metaphoric apex pairings with phrasal tones as a successful synchronisation because the bound was crossed by only 23ms, and most of the confidence interval fell within the bounds. We deem a deviation as small as almost one video-frame length (i.e., 17ms) as acceptable in this case. Therefore, it was concluded that the apex-tone pairings for metaphoric and beat gestures achieved synchronisation in the present study.

	t	df	p
(a) Iconics			
Pitch.acc	5.139	115	<0.001*
L	3.665	74	<0.001*
Phrasal	–2.167	33	0.982
(b) Deictics			
Pitch.acc	6.226	159	<0.001*
L	4.109	66	<0.001*
Phrasal	–2.001	52	0.975
(c) Metaphorics			
Pitch.acc	2.164	31	0.018*
L	3.175	24	<0.001*
Phrasal	1.537	24	0.067
(d) Beats			
Pitch.acc	3.395	65	<0.001*
L	2.937	49	<0.01*
Phrasal	2.103	15	<0.05*

Table 6: The equivalence test results for apex-tone pairings in metaphorics and beats.

Taken together, the equivalence test results revealed two patterns of synchronisation with tone type depending on gesture type. The synchronisation in iconics and deictics mirrored the pairing preference — pitch accents and Ls were the preferred targets of apexes for pairing, and these pairings achieved synchronisation. The dispreferred targets (i.e., phrasal tones) were not synchronised with apexes when they were paired. On the other hand, this situation was not the same for the apexes of metaphorics and beats. The pairing preference was not mirrored in synchronisation since it was achieved with all tonal events.

In line with Section 3.2.4, the present study interprets the synchronisation pattern in iconics and deictics as the general apex synchronisation behaviour (i.e., in line with overall

findings in Section 3.1), and the pattern in metaphorics and beats as a different pattern that results from the rhythmic behaviour of apexes observed in metaphorics and beats in the data.

Beats occurring consecutively in clusters, as in **Figure 13** (see Loehr, 2004 and references therein), impose rhythmic apex productions. Similarly, metaphorics occurred mostly when participants wanted to express meta-narrative content (McNeill, 1992), such as uncertainty or repetition, which were expressed with jerky gestures, causing multiple consecutive rhythmic apexes. The production of apexes in such rhythmic sequences can force pairings with the dispreferred tones since apexes have to occur at a location that is imposed by the rhythm, and the nearest tone to that location will form a pairing with that apex regardless of preference. These series of apex productions follow their own rhythm, but usually at least one of the apexes in the series (often the very first one) tended to show synchronisation with a pitch accent (see McClave, 1994 for similar observations). In terms of synchronisation, the time distances between apexes and tonal events are likely to be shorter in these rhythmic productions because Turkish can offer several tonal events over short durations as potential targets. Consecutive apex productions at fixed distances are able to find a tone occurring nearby in every case, ensuring synchronisation.

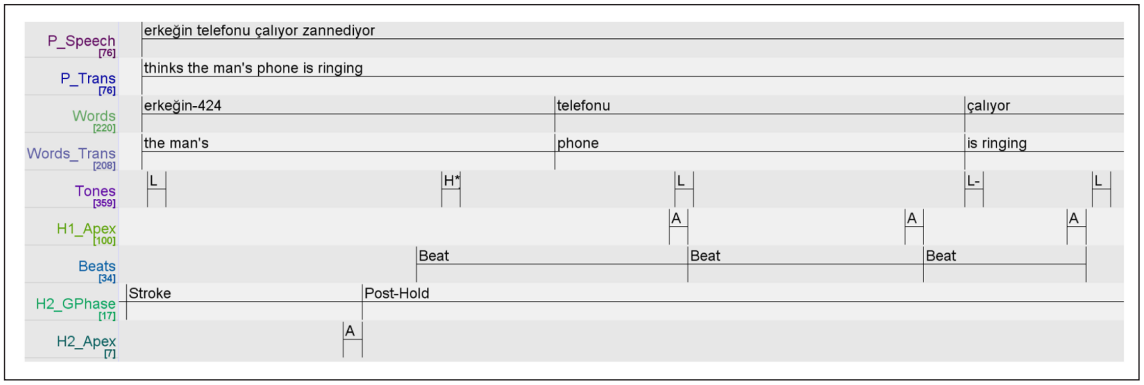


Figure 13: An ELAN screenshot showing that rhythmic pattern of apexes in consecutive beat gestures.

Overall, the metaphoric/beat apex synchronisations differed from the general synchronisation pattern observed for iconic and deictic apexes as a result of these common rhythmic formations of apexes over short periods of time. For metaphorics, we do not claim that the pattern we observed is representative of the gesture type as a whole. It is only the case that in our data, they behaved more like beats and displayed rhythmic behaviour within their formation (a single apex vs. repeating apexes in series). This implies that form-related properties of gestures can impact synchronisation as well. We leave such an analysis for a future study.

3.2.6 Effect of ip type

The final significant effect was the tone type and ip type interaction. **Figure 14** plots the equivalence test output for this effect. For nuclear ips (nips, N = 344), the estimates of pitch accents and Ls were only about one frame duration away from zero (20ms, –18ms respectively). The phrasal tone estimate was slightly further away from zero with –76ms. Accordingly, the equivalence test confirmed that all observed time distances under this condition were statistically equal to zero (see **Table 7**). For pre-nuclear ips (prips, N = 375), the estimated means of pitch accents and Ls were again closer to zero – they were only one to two frames away from it (–9ms and –34ms respectively). However, the phrasal tone estimate was much further away with –229ms, and fell outside of the lower bound. Consequently, the equivalence test was not significant for phrasal tones. However, the results were significant for pitch accents and Ls, achieving synchronisation.

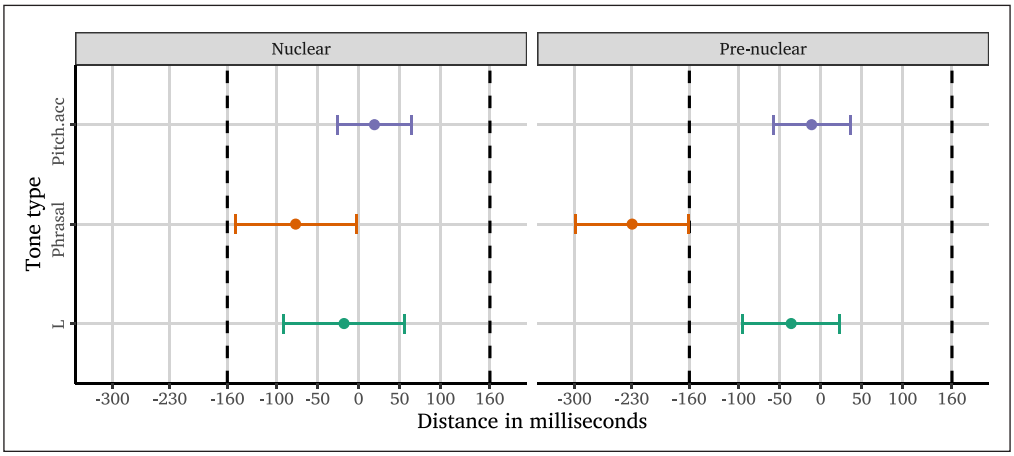


Figure 14: The estimated means of the Tone type : ip type interaction with CIs at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

	t	df	p
(a) Pre-nuclear			
Pitch.acc	6.226	151	<0.001*
L	4.109	141	<0.001*
Phrasal	–2.001	80	0.976
(b) Nuclear			
Pitch.acc	6.490	221	<0.001*
L	3.811	74	<0.001*
Phrasal	2.214	46	<0.05

Table 7: The equivalence test results for apex-tone pairings in pre-nuclear and nuclear ips.

The expected pattern was observed for prips where the synchronisation results reflected the preference indicated in the pairing pattern: Pitch accents and Ls were synchronised with the apexes they were paired with. The less common pairings with phrasal tones exhibited a lead for tones with a distance that was more than the average syllable duration, and therefore these failed to synchronise. Unlike in prips, the apex-tone synchronisation was successful for all pairings in nips. One possible explanation might be that the nuclear prominence had an effect on the distance between the members of the pairings. Regardless of the tone type, apexes were more tightly coupled with tones if they were in the nuclear area. This effect was not clearly observed for pitch accents and Ls as these were already tightly synchronised with apexes. However, the phrasal tone estimate moved almost an average syllable duration (153ms) closer to zero in nips compared to prips, achieving synchronisation. The effect of nuclearity presented here can be seen as evidence that the apex-tone synchronisation is sensitive to phrasal prominence as well.

3.3 The case of accenting in pre-nuclear ips

In Section 1.2.2, we introduced a disagreement in earlier studies on Turkish phonology about whether accentlessness is a feature of PWs with word-final stress. In the case of prips which are marked with a high tone at the right edge, some argued that this final rise is not a pitch accent and only marks the end of the prip; whereas others claimed that this rise has a double function marking both a pitch accent and the end of the prip. We argued that the pairing patterns and synchronisation of apexes can bring some insight into this discussion. In what follows, we concentrate on accented and accentless prips in our data, examining in more detail the pairing patterns in these phrases and the hypothesis they support. We will also see that the findings from this analysis are consistent with those shown in the synchronization analysis in Section 3.2.

Regarding the apex pairings in prips, the present study has hypothesised three scenarios which could inform us about the accenting of these ips. One hypothesis was that if the final rise has a double function then the apexes should be attracted to this rise, pairing with the phrase accent and the pitch accent claimed to be there (**Figure 15a**). This would mean that all prips contain a pitch accent. The second hypothesis was that there is no double function of the final rise, and therefore these phrases can lack prominence, in which case the apexes should be attracted to Ls (**Figure 15b**) as the only other tonal event associated at the level of the PW. This would also mean that apex synchronisation is sensitive to prosodic phrasing and hierarchy. The third hypothesis was that apexes will not show any clear synchronisation when the phrases were without a prominence marker — prominence is the only attractor for apexes. In what follows, we concentrate on accented and accentless prips in our data, examining pairing and synchronisation patterns in these phrases respectively, and then discuss which hypothesis they support.

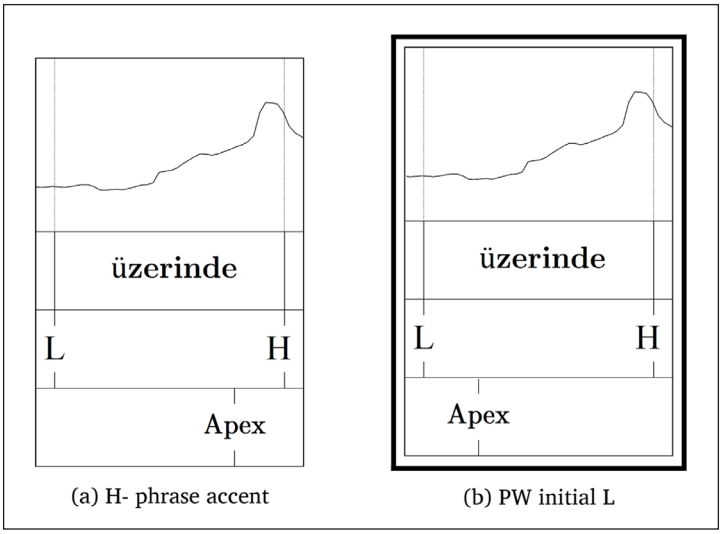


Figure 15: Two hypothetical gesture apex synchronisation cases (modified from Figure 3).

There were 405 prips containing a pitch accent in the data (out of 648), and 69% (N = 278) of these contained tonal event paired with apexes. **Table 8a** shows the pairing pattern of apexes in these prips (boundary tones were again grouped together with phrase accents). As seen in the table, the pairing preference was consistent: Apexes tended to be paired with pitch accents when the phrases were deemed to contain a pitch accent in our annotation which was independent from the annotation of apexes (Section 2.4).

(a) The types of paired tones					
	N	**%**			
Pitch.acc	153	55%			
L	68	24.5%			
L-X%	30	10.8%			
H-X%	27	9.7%			
Total	278				
(b) The effect of the location					
	Pitch.acc	**H-X%**	**L-X%**	**L**	**Total**
Final	57.4%	11.9%	11.9%	18.8%	101
Non-final	64.9%	7.8%	23.4%	3.9%	77

Table 8: Two tables showing the pairing patterns with tones in pre-nuclear ips with at least one pitch accent and whether this pattern changes depending on the location of the pitch accent in the phrase.

Here, we also checked whether the pairing pattern in accented prips was affected by the location of the pitch accent (final versus non-final syllable, see Section 1.2.2). The position of pitch accents within PWs could potentially have an effect on the distributions because in the non-final condition, pitch accents would be further away in time from the phrase accents at the edges of prips compared to the prips with word-final accents (compare the distance of the apex to the final rise in **Figures 16a** and **16b**). This increased distance in time amplifies the preferentiality of tones for pairing. That is, if apexes are paired with pitch accents even when they were further separated from the ip-final phrase accents, this would support that they were not attracted to the ip-final rises per se but to the pitch accents.

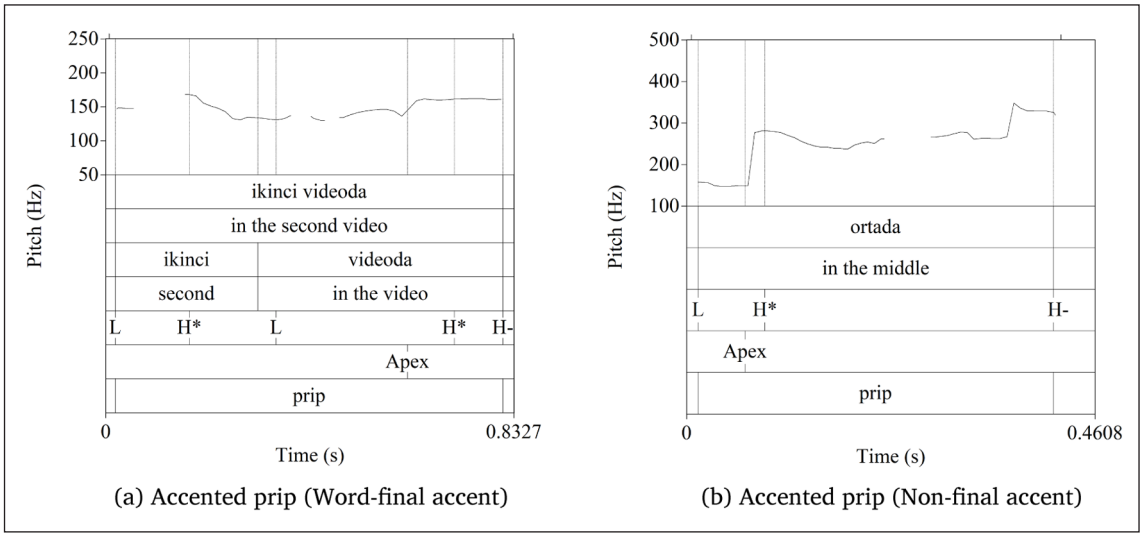


Figure 16: The apexes pairing with the tones in accented prips.

Table 8b breaks down **Table 8a** by whether the pitch accents in prips were word-final or non-final. The table shows the pairing instances where there was only one of each tone type available for pairing in the phrase (i.e., one L, one pitch accent, one phrase accent) to account for a potential effect of other available tonal events in the phrases. No major effect of pitch accent location on the pairing preference could be observed in the table — the preference was pitch accents in both conditions. As shown in the pairing examples in **Figures 16a** and **16b**, when pitch accents moved away from the phrase final rises (non-final condition), the apex locations tended to move away from the prip final rises along with pitch accents. This finding reinforces the claim that pitch accents and apexes are tightly coupled in that apex coordination is responsive to the shifts of accent locations within words. Similarly, the findings also show that apexes are not necessarily attracted to acoustic peaks themselves but to the prominence they cue (Section 1.1). Phrase accents and boundary tones in prips (and also generally) often presented higher pitch

values than pitch accents (e.g., ˆH- and L-H%, see Section 2.4.2), yet apexes were not paired with these.

In prips without a pitch accent, the phrase final rises still persist in the absence of pitch accents, marking the phrase boundary. Therefore, apexes occurring during these prips could be paired with either Ls or H- phrase accents (see **Figure 17**). There were 244 such accentless prips in our data (out of 648), and only 39% (N = 96) of these contained tones that were paired with apexes. This revealed the first difference between accented and accentless prips, which was that accentless prips were gestured less frequently than the accented ones — phrases bearing prosodic prominence attracted gestures (also observed across ip types, i.e., prips/nips vs. ptips, see Section 3.2.6).

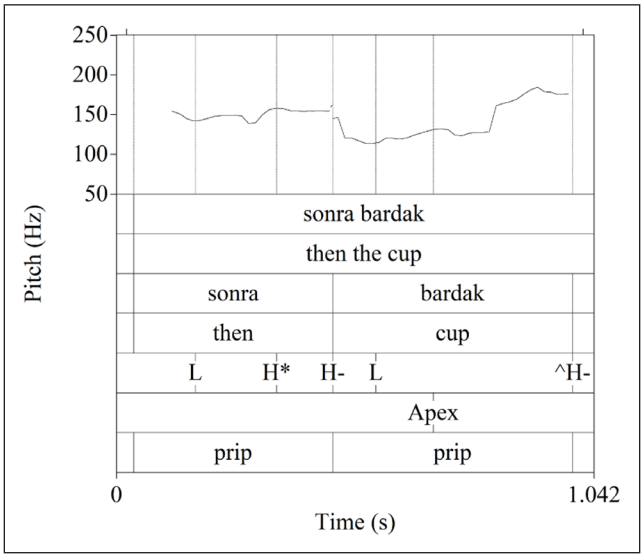


Figure 17: An apex pairing within an accentless prip.

As shown in **Table 9**, the dispreference for phrase-final rises was apparent in accentless prips. Apexes tended to be paired with Ls rather than with pitch rises (H-X%) in these prips, in line with the pairing preference reported in Section 3.1 (see **Figure 17** for an example). In these cases, the apex locations consistently shifted away from the phrase-final rises towards Ls. This is an example of a systematic behaviour of apexes, rejecting the third hypothesis which predicted that apexes would not show any consistent synchronisation when there is no marker of prominence (Section 1.2.2). Moreover, apexes did not stay anchored at the pitch rises. This indicates that there were no pitch accents at the location of the phrase final rises, which would result in more pairings with H-X%. Instead, the pairing preference shifted to Ls, as has been shown to happen

in the absence of prominence. Therefore, these findings suggest that these rises did not double function as part of a pitch accent in line with our annotation, rejecting the first hypothesis.

	N	%
L	64	66.7%
H-X%	25	26%
L-X%	7	7.3%
Total	96	

Table 9: The types of paired tones in pre-nuclear ips with no pitch accent.

Overall, the findings of the present study support the second hypothesis (**Figure 15b**). Apexes were chiefly paired with pitch accents when prips contained one. However, if prips were marked as accentless in our independent annotation of intonation (Section 2.4.2), then apexes were largely paired with L, indicating a distinct pattern in these cases. This hypothesis is further supported in the analysis of synchronisation (**Figure 14** in Section 3.2.6). There, it was shown that synchronisation was achieved when apexes were paired with both pitch accents and Ls in prips. Since in the majority of cases where apexes were paired with Ls in prips, the prip was unaccented, we can infer that apexes were both paired and synchronized with Ls in unaccented prips. On the other hand, even when apexes were paired with (i.e., nearest to) phrase accents and boundary tones, they were not synchronised with them.

4. Discussion

4.1 Summary

The present study investigated the synchronisation between gesture and prosody focusing on apexes and tonal events. There were two steps in the analysis. In the first step, the study analysed whether there were patterns in the pairings of apexes with tonal events. The second step consisted of the statistical testing of synchronisation.

The pairing analysis in Section 3.1 showed that apexes tended to be nearest to prominent pitch accents rather than other possible tonal events. If pitch accents were not available in the prosodic phrase, then apexes were paired with Ls, which are the only other tonal event at the level of the PW. Pairings with phrase accents and boundary tones which are associated with prosodic phrases above the level of the PW were avoided. This pattern was found to be consistent across different contexts. Based on these observations, pitch accents and Ls were identified as the preferred anchors of apexes for pairing, and phrase accents and boundary tones as dispreferred

anchors. In the second step of the analysis (Section 3.2), it was shown that synchronisation patterns were in line with these pairing patterns. That is, if an apex was paired with a preferred tonal event, then these were synchronised given the synchronisation criterion (Section 1.3). In other words, apexes tended to occur within an average syllable duration of pitch accents and Ls. In cases where apexes were nearest to dispreferred anchors, apex-tone synchronisation was not achieved, meaning that apexes tended to occur farther than an average syllable duration away from these tone types.

Including the possible effects of gestural and prosodic context in the analysis revealed exceptions to the general synchronisation pattern. Rhythmic apical productions in metaphorics and beats imposed synchronisation of apexes with tonal events regardless of type. Moreover, in nuclear ips, apex-tonal event pairings tended to be synchronised regardless of type. The general pairing pattern was still observed here: A great majority of apexes were paired/synchronised with pitch accents. Yet, under the effect of maximum prosodic prominence, apexes and tones were more tightly coupled in time.

4.2 Implications

Studies on gesture-prosody synchronisation have generally concentrated on prominence-based synchronisation (Section 1.1). These studies often isolated stressed syllables in the continuous event-rich prosodic signal and only checked whether apexes are synchronised with measures taken from these syllables when isolated from their prosodic context (cf. Loehr, 2004). Consequently, other possible synchronisations with anchors, such as the tonal events that mark boundaries, have been ignored. The results of the present study were in line with the prominence-based synchronisation claims for the most part (i.e., consistent apex synchronisation with pitch accent peaks).

Yet, further analyses revealed another synchronisation pattern where apexes were synchronised with boundary marking events, indicating the lack of pitch accents. Current claims about synchronisation cannot account for the pairings with Ls because they are not prominent events nor acoustic peaks. To the authors' knowledge, the possibility that apexes may also be synchronised with boundary-marking tonal events has only been mentioned recently for French by Rohrer et al. (2019) (see Section 1.1). The findings of the present study are the first to offer evidence that apexes can also be synchronised with boundary-marking tonal events. Interestingly, PWs in Turkish and accentual phrases in French show some intonational similarity. They both have default final stress and are marked with L tones at their left edge (Delais-Roussarie et al., 2015). These L tones do not encode prominence and create a contrast with phrase-final pitch rises (Ipek & Jun, 2013; Delais-Roussarie et al., 2015). Our findings were in line with Rohrer et al's (2019) prediction in that through synchronisation, apexes marked a prosodic domain independently from the encoding of prominence.

Our findings were also important as they have shown for the first time that synchronisation at the micro level was managed by the prosodic hierarchy in addition to prominence. Pitch accents and Ls are both associated with PWs in Turkish. When the highest priority target for synchronisation (i.e., a pitch accent) was not present, apexes stayed anchored at the PW level by synchronising with Ls. In fact, apex synchronisation displayed sensitivity to the prosodic hierarchy by not synchronising with the markers of other prosodic constituents (i.e., phrase accents and boundary tones) that are higher in the prosodic hierarchy (see **Figure 18**). This result is interpreted to mean that both prosodic prominence and hierarchy are active agents that play a role in the anchoring of gesture to speech. Altogether, these findings hint at a typology of synchronisation that adapts to the prosodic structure of languages. More research focusing on prosodically different languages is required to uncover different synchronisation patterns.

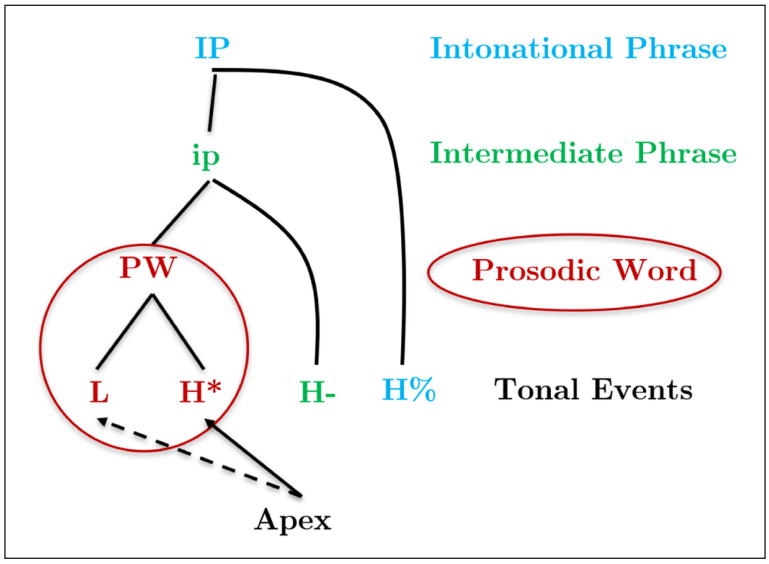


Figure 18: A schematic that shows the domain of apex synchronisation as the prosodic word.

The fact that apex synchronisation was not arbitrary and clearly informed both by prosodic prominence and phrasing shows how multimodal cues are intertwined with phonological structure. The present study has shown that in terms of production, phonological categories impact gesture production since the timing of gesture components relies on feedback from phonological encoding processes. Moreover, in terms of implementation, interlocutors methodically implement multimodal cues along with phonological ones to code communicative intent for other interlocutors to decode. It follows that multimodal cues can influence the perception of intonational categories by listeners, when they are available (Vaissière, 2008; Cruz, Swerts, & Frota, 2017; Kelly, Bailey, & Hirata, 2017). The systematic and tight synchronisation with prosodic structure shown here suggests these cues could be an even more effective cue to

prosodic events than has previously been explored, and play a role in the identification of both prominence-marking and boundary-marking prosodic events by listeners. The present study has shown evidence for this at the smallest unit level but there are also a few studies that reported temporal relationships between gestural phrases and prosodic phrases as well as information structural categories such as focus and topic (see Türk, 2020 and references therein). These findings support the general argument being made in the present study: Multimodal cues play a role in the implementation, production, and perception of intonational categories at multiple levels of the phonological structural hierarchy.

The present study has argued that such consistent behaviour of multimodal cues can be used to assist phonological analyses in identification of categories. In particular, it has shown an example case where looking at apex synchronisation can shed light on the accenting of words in Turkish focusing on pre-nuclear ips. The study made use of the bimodal apex synchronisation pattern to check whether words in pre-nuclear ips bear prominence. It was found that these phrases can be accented or accentless based on a systematic shift in the synchronisation behaviour of apexes. Note that the present study did not aim to explain why some words had pitch accents and some did not. Accentless realisations of these words can be a result of a variety of factors, which is out of the scope of this study to analyse. More research is required on the meaning and function (in terms of information structure) of these different realisations. The present study only showed that apexes were not synchronised with the pitch accent argued by some to be a constituent part – together with the phrase accent – of the tonal events at the end of the ip. Rather, apex synchronisation behaved as if there was no pitch accent in these words, offering multimodal evidence in support of the claim that words with final stress may not bear a pitch accent.

Our findings highlight the importance of a linguistically informed selection of anchors for synchronisation tests. Using measurements such as jaw displacement, vowel onsets, and acoustic peaks may provide methodological convenience; however, the understanding that gestural and prosodic units exist as members of complex systems where events, constituents, and their positioning influence each other can add valuable insight into our understanding of the relationship between multimodal cues and phonological categories, and more generally how speech and gesture are temporally coordinated. Furthermore, since the prosodic structures of languages exhibit different features cross-linguistically, more studies in different languages are needed to understand these in full. Investigations in this manner may shed light on the complex relation between acoustic and visual information, and how this relation is used by speakers and listeners as well as by us analysts.

Supplementary file

The supplementary file for this article can be found as follows:

- **Supplementary File.** Scenarios. DOI: https://doi.org/10.16995/labphon.6432.s1

Acknowledgements

This study had ethical approval from the Victoria University of Wellington Human Ethics committee (approval number: 23680). The study was supported by Didem Yaman Scholarship offered by Ministry of Foreign Affairs of New Zealand, awarded to the first author.

Competing interests

The authors have no competing interests to declare.

References

Altmann, H. (2006). *The perception and production of second language stress: A cross-linguistic experimental study* (Unpublished doctoral dissertation). University of Delaware.

Arnhold, A., & Kyröläinen, A.-J. (2017). Modelling the interplay of multiple cues in prosodic focus marking. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 1–25. DOI: https://doi.org/10.5334/labphon.78

Arvaniti, A., Ladd, D., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, *26*(1), 3–25. DOI: https://doi.org/10.1006/jpho.1997.0063

Arvaniti, A., Ladd, R., & Mennen, I. (2000). What is a starred tone? Evidence from Greek. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the lexicon* (pp. 119–131). Cambridge University Press.

Atterer, M., & Ladd, D. (2004). On the phonetics and phonology of segmental anchoring of F0: evidence from German. *Journal of Phonetics*, *32*(2), 177–197. DOI: https://doi.org/10.1016/S0095-4470(03)00039-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

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

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling (version 3). *The OSU Research Foundation*, 1–30.

Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GESPIN)*. Bielefeld, Germany.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer Software]. Retrieved from http://www.praat.org/ (version 6.0.56)

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7–9), 1044–1098. DOI: https://doi.org/10.1080/01690965.2010.504378

Carletta, J. (1996). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, *22*, 249–254.

Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1–2), 1–31. DOI: https://doi.org/10.1080/23273798.2014.963130

Creider, C. (1986). Interlanguage comparisons in the study of the interactional use of gesture: Progress and prospects. *Semiotica*, *62*(1–2), 147–164. DOI: https://doi.org/10.1515/semi.1986.62.1-2.147

Cruz, M., Swerts, M., & Frota, S. (2017). The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1). DOI: https://doi.org/10.5334/labphon.110

De Ruiter, J. P. (2000). The production of gesture and speech. *Language and Gesture*, *2*, 284–311. DOI: https://doi.org/10.1017/CBO9780511620850.018

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., di Cristo, A., Feldhausen, I., … Yoo, H. (2015). *Intonational phonology of French: Developing a ToBI system for French*. DOI: https://doi.org/10.1093/acprof:oso/9780199685332.003.0003

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269. DOI: https://doi.org/10.1162/jocn_a_00963

Domahs, U., Genc, S., Knaus, J., Wiese, R., & Kabak, B. (2013). Processing (un-) predictable word stress: ERP evidence from Turkish. *Language and Cognitive Processes*, *28*(3), 335–354. DOI: https://doi.org/10.1080/01690965.2011.634590

ELAN. (2019). [Computer Software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://archive.mpi.nl/tla/elan (version 5.8)

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. *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*, 850–864. DOI: https://doi.org/10.1044/1092-4388(2012/12-0049)

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., … Stoel, R. (2007). Information structure. *Interdisciplinary Studies on Information Structure*, *7*, 147–187.

Graziano, M., Nicoladis, E., & Marentette, P. (2020). How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Language Learning*, *70*(1), 266–304. DOI: https://doi.org/10.1111/lang.12376

Güneş, G. (2013). On the role of prosodic constituency in Turkish. In U. Özge (Ed.), *Proceedings of Workshop on Altaic Formal Linguistics* (Vol. 8, pp. 115–128). MITWPL.

Güneş, G. (2015). *Deriving prosodic structures* (Unpublished doctoral dissertation). University of Groningen.

Harrell Jr, F. E. (2019). rms: Regression modeling strategies [Computer software manual]. (R package version 5.1-3.1)

Hartigan, J. A., Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84. DOI: https://doi.org/10.1214/aos/1176346577

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514. DOI: https://doi.org/10.3758/PBR.15.3.495

Hualde, J. I., & Prieto, P. (2016). Towards an international prosodic alphabet (IPrA). *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*, 25. DOI: https://doi.org/10.5334/labphon.11

Ipek, C., & Jun, S.-A. (2013). Towards a model of intonational phonology of Turkish: Neutral intonation. In *Proceedings of Meetings on Acoustics* (Vol. 19, p. 060230). Montreal, Canada: Acoustical Society of America. DOI: https://doi.org/10.1121/1.4799755

İşsever, S. (2003). Information structure in Turkish: the word order–prosody interface. *Lingua*, *113*(11), 1025–1053. DOI: https://doi.org/10.1016/S0024-3841(03)00012-3

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. In S. Ishihara, M. Schmitz, & A. Schwarz (Eds.), *Interdisciplinary Studies on Information Structure* (Vol. 3, pp. 199–244). Potsdam: Universitätsverlag.

Kabak, B., & Vogel, I. (2001). The phonological word and stress assignment in Turkish. *Phonology*, *18*(3), 315–360. DOI: https://doi.org/10.1017/S0952675701004201

Kamali, B. (2011). *Topics at the PF interface of Turkish* (Unpublished doctoral dissertation). Harvard University.

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology*, *3*(1). DOI: https://doi.org/10.1525/collabra.76

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

Kita, S. (2000). How representational gestures help speaking. *Language and Gesture*, *1*, 162–185. DOI: https://doi.org/10.1017/CBO9780511620850.011

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. DOI: https://doi.org/10.1016/S0749-596X(02)00505-3

Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer Interaction* (pp. 23–35). Springer. DOI: https://doi.org/10.1007/BFb0052986

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

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: a process. *Language and Gesture*, *2*, 261–291. DOI: https://doi.org/10.1017/CBO9780511620850.017

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). DOI: https://doi.org/10.5334/labphon.75

Kügler, F., & Calhoun, S. (2020). Prosodic encoding of information structure: a typological perspective. In C. Gussenhoven & A. Chen (Eds.), *Oxford Handbook of Language Prosody* (pp. 454–467). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198832232.013.30

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., … Peters, J. (2015). DIMA annotation guidelines for German intonation. In M. Wolters (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences.* International Phonetic Association.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. DOI: https://doi.org/10.18637/jss.v082.i13

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511808814

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and metaanalyses. *Social Psychological and Personality Science*, *8*(4), 355–362. DOI: https://doi.org/10.1177/1948550617697177

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

Levi, S. V. (2005). Acoustic correlates of lexical accent in Turkish. *Journal of the International Phonetic Association*, *35*(1), 73–97. DOI: https://doi.org/10.1017/S0025100305001921

Loehr, D. P. (2004). *Gesture and intonation* (Unpublished doctoral dissertation). Georgetown University Washington, DC.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45–66. DOI: https://doi.org/10.1007/BF02143175

McClave, E. (1991). *Intonation and gesture* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

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

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. *Language and Gesture*, 141–161. DOI: https://doi.org/10.1017/CBO9780511620850.010

Nobe, S. (1996). *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production*. University of Chicago, Department of Psychology.

Özge, U., & Bozsahin, C. (2010). Intonation in the grammar of Turkish. *Lingua*, *120*(1), 132–175. DOI: https://doi.org/10.1016/j.lingua.2009.05.001

Prieto, P., & Torreira, F. (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics*, *35*(4), 473–500. DOI: https://doi.org/10.1016/j.wocn.2007.01.001

Prieto, P., Van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, *23*(4), 429–451. DOI: https://doi.org/10.1006/jpho.1995.0032

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1507–1521. DOI: https://doi.org/10.1044/1092-4388(2008/07-0173)

Rohrer, P. L., Prieto, P., & Delais-Roussarie, E. (2019). Beat gestures and prosodic domain marking in French. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1500–1504). Australasian Speech Science and Technology Association Inc.

Roustan, B., & Dohen, M. (2010). Gesture and speech coordination: The influence of the relationship between manual gesture and speech. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association*. Makuhari, Japan. DOI: https://doi.org/10.21437/Interspeech.2010-207

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

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of prosody and position on the timing of deictic gestures. *Journal of Speech, Language, and Hearing Research*, *56*(2), 458–473. DOI: https://doi.org/10.1044/1092-4388(2012/11-0283)

Sezer, E. (1981). On non-final stress in Turkish. *Journal of Turkish Studies*, *5*, 61–69.

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

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

Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A method for studying the time alignment of gestures and prosody in American English: Hits and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanic, E. Keller, & M. Marinaro (Eds.), *NATO security through science series E human and societal dynamics* (Vol. 18). Washington, DC: IOS PRESS.

Tuite, K. (1993). The production of gesture. *Semiotica*, *93*(1–2), 83–106. DOI: https://doi.org/10.1515/semi.1993.93.1-2.83

Türk, O. (2020). *Gesture, prosody and information structure synchronisation in Turkish* (Unpublished doctoral dissertation). Victoria University of Wellington.

Vaissière, J. (2008). Perception of intonation. In D. Pisoni & R. Remez (Eds.), *The Handbook of Speech Perception* (p. 236–263). John Wiley. DOI: https://doi.org/10.1002/9780470757024.ch10

Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, *34*(3), 459–520. DOI: https://doi.org/10.1515/ling.1996.34.3.459

Vogel, I. (2020). Fixed stress as phonological redundancy: Effects on production and perception in Hungarian and other languages. In *Approaches to Hungarian* (Vol. 16, pp. 188–206). John Benjamins. DOI: https://doi.org/10.1075/atoh.16.09vog

Vogel, I., Athanasopoulou, A., & Pincus, N. (2016). Prominence, contrast, and the functional load hypothesis: An acoustic investigation. In J. Heinz, R. Goedemans, & H. van der Hulst (Eds.), *Dimensions of phonological stress* (p. 123–167). Cambridge University Press. DOI: https://doi.org/10.1017/9781316212745.006

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. DOI: https://doi.org/10.1016/j.specom.2013.09.008

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, *55*(4), 179–203. DOI: https://doi.org/10.1159/000028432