Language and Cognition (2025), 17, e54, 1–26
doi:10.1017/langcog.2025.20

ARTICLE

Do hand gestures increase perceived prominence in naturally produced utterances?

Patrizia Paggio[1,2], Holger Mitterer[3,4], Greta Attard[1] and Alexandra Vella[1]

[1]Institute of Linguistics and Language Technology, University of Malta, Msida, Malta; [2]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark; [3]Department of Cognitive Science, University of Malta, Msida, Malta and [4]Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University, Seoul, South Corea
Corresponding author: Patrizia Paggio; Email: patrizia.paggio@um.edu.mt

(Received 30 October 2022; Revised 07 December 2024; Accepted 09 April 2025)

Abstract
This study investigates the effect of visually perceived gestures on the overall (multimodal) prominence of naturally occurring stimuli extracted from a multimodal corpus of Maltese conversations. Experiment participants were required to rate the prominence of target words in sentences presented to them as audiovisual and audio-only stimuli. In half of the stimuli, the target word was accompanied by a co-speech hand gesture. The results of the experiment show (i) that words produced with a co-speech gesture were rated as more prominent than words that were produced without one and (ii) that this was the case independently of whether raters could see those gestures (audiovisual condition) or not (audio-only condition). An acoustic analysis of the data shows that the presence of a co-occurring gesture has a significant effect on the pitch of the target vowel. The study suggests that gestures may provide the listener with an additional but not necessary cue to perceiving prominence.

Keywords: corpus stimuli; prominence in speech and hand gestures; spoken Maltese; visual prosody

1. Introduction

While Hockett (1960) identified the use of the vocal-auditory channel as one of the main features of language, and one leaving the rest of the body free to perform other activities, it is now well established that communication is better characterised as involving multimodality. Effects due to the interaction between speech and the visual modality go well beyond the use of visual input in the form of lip movement to decode the speech signal (McGurk & MacDonald, 1976). For instance, Holler et al. (2018) investigated how fast participants in a conversation are able to answer questions and found faster responses for answers to questions accompanied by gestures. Drijvers and Holler (2022) asked participants to shadow (i.e., repeat back as fast as possible) recordings from natural conversations and found that seeing the speaker helps even

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

when mouth movements are invisible, suggesting that the presence of gesture supports comprehension. The importance of gestural behaviour in communication is also evident in a recent study that found less creative thinking in online meetings than in physical multimodal settings (Brucks & Levav, 2022). Such findings raise the question of what kinds of multimodal information support the perception of what types of linguistic information.

In the current paper, we investigate the extent to which prominence perception may benefit from gestural cues that are produced during face-to-face interaction. Prominence is understood here as a relational property whereby a linguistic element *stands out* from other elements in an utterance (Grice & Kügler, 2021), and we speak of *multimodal* prominence as a property combining visual and prosodic characteristics.

It is well established that gesture is tightly connected with speech in human communication (Kendon, 2004; McNeill, 2008) and cognition (de Ruiter, 2000; Gentilucci & Dalla Volta, 2007; Iverson & Thelen, 1999; Willems & Hagoort, 2007), and that one of the ways in which this tight relationship manifests itself is in the temporal coordination between gestures and units of speech, in particular units of prosodic structure (see Wagner et al., 2014, for an overview).

Alignment between speech and the gestural modality, including not only hand gestures but also facial expressions and head movements, has been studied from the perspective of language production (Cavé et al., 1996; Loehr, 2007; McClave, 1998), language development (Esteve-Gibert et al., 2022) and in several languages (Chui, 2005; Ferré, 2010). These findings converge on a view of gesture and speech as temporally aligned via their prominence cues and provide support for an understanding of communication as involving a common cognitive origin of the two modalities.

It has been debated whether gestures play a role for the speaker or the listener. For instance, Krauss and Hadar (1999) claim that their main function is to help the speaker in managing their communicative flow, for example, for lexical retrieval. Other studies, however, suggest that the presence of gestural behaviour has an effect on the way interlocutors perceive speaker utterances (Dohen & Loevenbruck, 2009; Glave & Rietveld, 1979; House et al., 2001; Prieto et al., 2015), and that beats in particular contribute to the perception of prominence of target words (Bosker & Peeters, 2021; Krahmer & Swerts, 2004, 2007). The latter studies are based on highly controlled Dutch stimuli. The goal of the present one, by contrast, is to investigate how gestures contribute to the perception of overall (multimodal) prominence in naturally occurring stimuli extracted from conversations. This extension to conversational data requires explicitly addressing challenges such as the large variability in both speech and gesture, which in turn makes it harder to disentangle the factors that affect the perception of different elements in multimodal signals. Furthermore, we are the first to look at multimodal prominence perception in Maltese.

Assuming that speakers use gestures if a word is particularly important to them, we expect that words accompanied by gestures will be produced with stronger prominence in the speech signal than words not accompanied by a gesture. Therefore, we hypothesise that the perceived prominence of a target word in an utterance will be increased if the word is accompanied by a manual gesture. We further hypothesise that the perceived prominence of the co-occurring word will be higher if listeners see the speaker produce the manual gesture compared with if they do not. In other words, the first hypothesis concerns the listener's perception of utterances that are produced with *vs.* without a co-speech gesture. The second concerns the listener's perception of the same two types of utterance while viewing *vs.* not viewing

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

the gestures and, therefore, investigates whether seeing the gesture actually affects prominence perception. The choice of positing two separate hypotheses is based on the fact that several studies (Ambrazaitis & House, 2023; Berger & Zellers, 2022; Esteve-Gibert et al., 2022; Krahmer & Swerts, 2007; Krivokapić et al., 2017; Pouw et al., 2021) have provided evidence for stronger articulated verbal prominence cues in connection with accompanying gestures. Therefore, if gestures affect prominence perception, this may also happen if the gestures are not seen (first hypothesis). However, if the visual cues provided by the gestures add to the perception of prominence, beyond their effects on the speech signal, then the effect on prominence perception should be stronger when gestures are visible than otherwise (second hypothesis).

In this study, we compare the acoustic cues to prominence in words produced with *vs.* without co-occurring gestures in utterances extracted from Maltese dialogues. We then test whether participants listening to the utterances perceive words accompanied by gestures as more prominent. Finally, we test whether seeing the accompanying gestures also increases the perceived prominence.

We begin with a brief review of research investigating the relationship between gestures and prosodic features in Section 2.1, followed by an account of how prosodic prominence is expressed in Maltese in Section 2.2. In Section 3, we explain the methods used in the experiment, including gestural and acoustic properties of the stimuli used, particularly the effect of gesture co-presence on intensity and pitch values of the stressed vowels in the target words. The results of the perception experiment are reported in Section 4 and discussed in Section 5. Section 6 draws the conclusions.

2. Background

2.1. Related work

Systematic relationships between gestures (not necessarily of the hand) and prosodic features have been observed in several studies. In general, there is broad agreement in the literature that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries (Bolinger, 1986; Kendon, 1980; Leonard & Cummins, 2010; Loehr, 2007, 2004), and several studies claim that hand gesture strokes – a stroke being defined as the most dynamic part of a gesture – are temporally aligned with (or slightly precede) the main sentence accent (Alahverdzhieva & Lascarides, 2010; McNeill, 1992). In their empirical study of German data (276 examples), Ebert et al. (2011) confirmed this generalisation by observing that gesture strokes tend to precede sentence accent by 0.36 s on average. Esteve-Gibert and Prieto (2013) investigated the alignment of speech and pointing gestures in Catalan and found that the positions of intonation peaks and gesture apexes were correlated and influenced by prosodic structure. In a further study on Catalan data, Esteve-Gibert et al. (2017) looked at coordination between prosodic elements and head movements and demonstrated that the position of prosodic heads (accented syllables) and prosodic edges (prosodic word and intonational phrase boundaries) has an impact on the timing of head movements. Ambrazaitis et al. (2020b) found a tendency for stressed syllables in Swedish compounds to be accompanied by gestural movements.

It has been suggested that the choice of which words will be accompanied by gestures could be explained in terms of information structure (IS), which in turn

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

affects the degree of prominence of linguistic expressions in a sentence. Thus, Ebert et al. (2011) observed that the onsets of hand gesture phrases in their German data align with new-information foci but not contrastive ones. Krahmer and Swerts (2004) studied the way eyebrows contribute to the perception of focus in Dutch and Italian. Their conclusion, however, is that eyebrow movements seem to play little to no role in the perception of focus in spite of the fact that speakers of both languages prefer an eyebrow movement to coincide with the most prominent word. Paggio et al. (2018) investigated the prosodic characteristics and occurrence of gestures in Maltese fronted complements. Complement fronting in Maltese is a marked construction in IS terms, with the fronted complement taking the role of contrastive topic or focus depending on the context. The study showed that in Maltese, fronted complements with a falling pitch accent are often accompanied by hand gestures. Debreslioska and Gullberg (2020) found, on the basis of German narrative spoken data, that the information status of a referent, whether new or inferable, has an effect on the type of gesture it is accompanied by. Esteve-Gibert et al. (2022) showed that French-speaking children use head movements to mark the informational status of discourse referents.

The role of co-speech gesture – again, not just hand gestures – has also been studied by testing its effect on listeners. Several studies provide experimental evidence suggesting that beats contribute to both the realisation and perception of prominence. Prieto et al. (2015) explored the role of visual cues in the perception of focus in Catalan using a 3D animated character and found that both head and eyebrow movements contributed to the perception of contrastive focus, whereas head nods were more informative than eyebrow movements for focus identification. In House et al. (2001), listeners were tasked with identifying the most prominent of two words in test sentences uttered by a talking face in Swedish. The study found that eyebrow and head movements contributed to prominence perception when synchronised with the stressed vowel of the potentially prominent word. Krahmer and Swerts (2007) used highly controlled Dutch stimuli to investigate the effect of gesture on prominence perception. The results of the study suggest that visual beats (hand, face, and eyebrow) have a significant effect on the perceived prominence of the target word (in this study, either the word *Amanda* or the word *Malta*; in other words, only two possible words, each containing the same target vowel). Seeing a speaker realise a visual beat on a word increases its perceived prominence compared with not seeing the beat. In addition, the presence of an accompanying beat gesture has an effect on some of the acoustic characteristics of the target word, that is, duration and the F_2 formant. The combined effect of acoustic and gestural cues on multimodal perception of prominence is discussed in Jiménez-Bravo and Marrero-Aguiar (2020) based on audiovisual stimuli created from recordings of conversations in an online Spanish talent show. Stimuli for this study were generated by manipulating both F_0 and intensity to determine their contribution to perceived prominence in the presence of gestural cues. It was found that duration had a stronger effect than F_0.

Bosker and Peeters (2021) showed that beat gestures can influence stress perception. The authors used Dutch minimal stress pairs (e.g., *KAnon vs. kaNON* 'canon' *vs.* 'cannon' in English), removed the pitch and intensity cues to stress and had a hand gesture produced on the first or second syllable. Participants perceived the word to be stressed on the syllable that was aligned with the hand gesture.

Ambrazaitis et al. (2020a, 2022) studied prominence perception with data from Swedish TV news. Similarly to Krahmer and Swerts (2007), they found that words

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

realised with a pitch accent and head movement tended to receive higher prominence ratings than words with only a pitch accent in the audiovisual condition, whereas words with low prominence tended to be rated slightly higher in the audio-only condition. These results were further analysed and discussed in Ambrazaitis and House (2022), where cumulative effects were observed between F_0 rises, the presence of accompanying head movements and the occurrence of eyebrow movements.

In summary, empirical studies of multimodal language data in several languages have shown systematic relationships between prosodic features of speech and the timing as well as the function of accompanying gestures. One aspect of this relationship is the way gestures, both of the hands and the head, affect the perception of prominence by listeners. Several studies have demonstrated the effect of gestures using highly controlled data, or data from reading TV news. The only study we know of which uses naturally occurring conversational data is Jiménez-Bravo and Marrero-Aguiar (2020). The authors themselves, however, consider the study a pilot project due to the limited number of experiment participants (12). More research is necessary to understand how listeners perceive multimodal prominence in naturally occurring unscripted data. The current study is a contribution in this direction.

2.2. Prosodic prominence in Maltese

Determining the location of prosodically prominent elements is an important prerequisite for a study on the perception of prominence of gestures that may co-occur with such elements. Research on prosodic prominence in Maltese, though not extensive, suggests that there is a strong tendency for the main sentence accent that marks prominence to gravitate towards the right edge of the relevant prosodic domain (Vella, 2009a). The lexically stressed syllable that is assigned the main sentence accent also serves as the anchor for a pitch accent. More specifically, Vella (1995) claims that prosodic structure in Maltese is organised in such a way that it is the phonological phrase which is the domain of focus, defined in Gussenhoven (1983, p. 18) as 'mark[ing] the speaker's declared contribution to the conversation, while [−focus] constitutes his cognitive starting point'. The focused element within the (final) phonological phrase (in a sequence) is designated as nuclear within the higher-level intonational phrase domain and assigned one of a number of pitch accents identified for Maltese. Additional material can follow this focused element outside the phonological phrase but within the same intonational phrase. This gets assigned a phrase accent rather than a pitch accent.

In Maltese, two phrase accents, defined following Grice et al. (2000, p. 180) as 'edge tones with a secondary association to an ordinary tone-bearing unit', have also been identified. These phrase accents (Vella, 1995, 2003, 2009a, 2009b) have a secondary association to a lexically stressed syllable in the stretch of speech which follows nuclear material, by definition post-nuclear and post-focal [-focus] following Gussenhoven's definition (see above). Vella (1995)'s claim is that these phrase accents in Maltese are confined to [-focus] stretches of speech which are extrametrical to the phonological phrase while nevertheless occurring within the intonational phrase, and by virtue of this fact are followed by a tone associated with the intonational phrase boundary. A recent discussion of these accents is provided in Vella and Grice (2024). What is relevant in the context of this paper is that the phrase accents that occur on [-focus]

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

stretches of speech are not considered to be prominence-lending in the same way that pitch accents are.

The placement of pitch accents in Maltese is influenced by a number of factors, including both syntactic structure and IS. As a result, when formulating the criteria for selecting target utterances, it was necessary to take these factors into account to ensure that they were held constant across the sets of experimental stimuli (see also Section 3.2.2). For example, in utterances in which a pitch accent is followed by a phrase accent, prosodic prominence is signalled by the pitch accent occurring relatively early in the utterance, even in instances where this is followed by a phrase accent signalling a secondary rather than a primary prominence.

The relative constituent order flexibility that characterises Maltese, as well as other structural conditions such as the use of pronominal clitics at the end of verbs, also often lead to the [+focus] element in an utterance, which gets assigned the main sentence accent, coming relatively early in the utterance (Vella, 2003). As mentioned above, the [+focus] element, irrespective of whether it comes late or early, is the one which carries the nuclear pitch accent and is considered prominent.

Thus, for example, *SEna*, 'years' is the [+focus] element in Example A.2.12 in Appendix A.2 from the list of examples used in our study,[1] *Jien għandi dsatax il-SEna*, 'I have nineteen years' (more idiomatically, 'I am nineteen years old'). In this case, there is no material following the prominent element that comes at the end of the sentence. By contrast, the [+focus] and therefore prominent element, in Example A.1.13 in Appendix A.1, *Sal-aħħar tax-XAHAR għandhom* 'Till the end of the month they have', is *XAHAR* 'month'. This is the last word in the phonological phrase *sal-aħħar tax-XAHAR*, but it is followed post-nuclearly by *għandhom*, which, following Vella (1995), is deemed to fall outside the phonological phrase, though within the intonational phrase, thus, it is rendered ineligible to be designated as the carrier of the prominent main sentence accent. An (unmarked) SVC(omplement) version of this, *Għandhom sal-aħħar tax-xahar*, 'They have till the end of the month', without a following post-nuclear stretch, would of course also be possible. But the version of this sentence in our dataset starts with the complement rather than the verb. This is the most 'important' information, the information that the speaker wants to ensure is conveyed to their interlocutor. The pro-dropped verb *għandhom* 'they have', which follows, in other words, the complement in this sentence, is given information, which is backgrounded in terms of IS. It is [-focus] and hence gets assigned a phrase accent rather than a pitch accent that marks a secondary rather than a primary prosodic prominence (see Vella, 1995; 2003).

To conclude, while prosodic prominence in Maltese has a tendency to occur late, for example, in Example A.2.12 in Appendix A.2 discussed above, it is also possible for prominence to occur earlier than the final word in the phonological phrase. As mentioned above, Paggio et al. (2018) have shown that fronted complements in Maltese often involve both a pitch accent, which by definition implies prosodic prominence, and a hand gesture. Example A.1.13 in Appendix A.1, also discussed above, is a good illustration of a sentence in which prominence occurs relatively early.

In view of the complexity in prominence relations arising from the interplay between prosodic structure and prominence in Maltese, in selecting our stimuli, we tried to keep examples in which constituent order is not SVO or SVC to an absolute

[1]See the Appendix.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

Table 1. Mean (SD) of participants' self-rated proficiency (scale 0–4) in Maltese and English

Language	Listening	Speaking	Reading	Writing
Maltese	3.85 (0.55)	3.73 (0.71)	3.74 (0.61)	3.46 (0.82)
English	3.91 (0.33)	3.72 (0.50)	3.84 (0.42)	3.79 (0.46)

minimum. This means that sentence accent in our examples generally occurs on the final element within the sentence, although this final element could have the lexical stress on the final, penultimate or antepenultimate syllable. There are a few exceptions, such as Example A.1.13 in Appendix A.1.

3. Methods

An experiment was carried out to test our hypotheses, which, to repeat, are (i) if a target word has been produced with an accompanying manual gesture, a listener will hear it as more prominent than a word that was produced without a gesture and (ii) if the listener can also see the accompanying gesture, the perceived prominence of the target word will be even stronger. In this section, we explain the methods used in the design, conduct and analysis of the experiment.

3.1. Participants

A convenience sample[2] from the student (and staff) population at the University of Malta participated in the study. A total of 95 participants (51 male and 44 female) took part in the study for a monetary compensation. Participants were asked to provide their age bracket, with 49 being 18–26 years old, 30 being 26–35 years old and a further 11, 3 and 2 participants in the age brackets 36–56, 46–59 and 60+, respectively.

Participants self-rated their proficiency in Maltese and English on a scale from 0 to 4 for the four language skills (listening, speaking, reading and writing). As Table 1 shows, participants identified themselves as highly proficient in Maltese and English, although slightly less so in Maltese writing. This reflects the fact that students mostly write in English, while Maltese tends to be their preference in other modalities (Vella, 2013).

3.2. Materials

We used the MAMCO corpus of video-recorded Maltese conversations (Paggio et al., 2018; Paggio & Vella, 2014) to find examples containing words bearing prominence that were or were not accompanied by a hand gesture. The MAMCO corpus is the first multimodal resource involving Maltese conversational data. It consists of 12 video-recorded conversations between 12 speakers (6 females and 6 males), each taking part in two different short (∼5 min) dialogues in a recording studio. All participants are Maltese-dominant speakers. They had not met prior to the experiment. They were asked to speak freely and try to get to know each other. They were recorded by three different cameras while standing facing each other. For this study,

[2]A convenience sample is not based on probabilistic methods but simply relies on convenience.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press



Figure 1. Semi-frontal speaker view from one of the MAMCO conversations.

we employed the recordings made using two cameras recording each participant from a semi-frontal angle. For each dialogue, the recordings obtained with these two cameras were later edited into the same video. An example is shown in Figure 1.

Lapel microphones were used to record the audio. After the recordings, participants were asked to fill in a survey and sign a consent form. All gave their permission for the resulting data collection to be used for research purposes. The setting and general organisation used to collect the corpus replicate those used in the Nordic NOMCO corpus (Paggio et al., 2010), a choice that was motivated by the wish to allow future comparative analyses of the way gestures are used in different languages.

All twelve conversations were orthographically transcribed with word boundaries. An acoustic analysis of the data was also carried out as required (see more in Section 3.2.5).

3.2.1. Gesture annotation in MAMCO

All head movements and a random selection of hand gestures were annotated prior to this study following the MUMIN coding scheme for gestural annotation (Allwood et al., 2007). The coding scheme defines formal and functional attributes for various types of movements. For instance, a hand movement can be *one-handed* or *two-handed*, and it can be annotated as *symbolic*, *iconic* or *indexical* following Peirce's semiotic categories (Peirce, 2009). Within the indexical category, a distinction is made between *deictic* and *non-deictic*, the latter corresponding to gestures that only display a beat-like dimension, that is, a rapid biphasic movement excursion. Gestures are annotated as a single time interval, that is, including the preparatory phase, without marking internal components.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

3.2.2. Extracting and further annotating the stimuli

We extracted 60 short sentence stimuli from the MAMCO corpus, all involving a target word carrying sentence accent; the target word mostly occurs in the final part of a sentence. In half of the stimuli, the target word is accompanied by a hand gesture, the 'withGesture' condition. This means that the stroke of the gesture co-occurs with the word. Since the stroke was not indicated in the original annotations, it had to be identified by the research assistant choosing the examples. Hand gestures that were not already coded were also annotated following the scheme described earlier. Semantic aspects depicted in the gesture did not play a role in deciding the alignment.

In the remaining half of the stimuli, the 'noGesture' condition, no hand gestures are present.

The stimuli (together with 30 fillers) were taken from speech produced by ten of the twelve speakers in the corpus. Two research assistants were tasked with choosing and extracting the example sentences, and two senior researchers validated them, both in terms of the choice of examples and clip editing. The complete sentence list is provided in the Appendix.

An example where the accented word is accompanied by a hand gesture is shown in Figure 2 (Example A.1.7 in Appendix A.1): the speaker on the right says *kelli subSIdiary area* 'I had a subSIdiary area' and makes a hand gesture that is aligned with the accented syllable *SI*. In this and other illustrative examples in the text, the stressed syllable of the target word is shown in capitals. Note that in the stimuli shown to the experiment participants, faces were not blurred. They are blurred here for reasons of privacy protection.



Figure 2. Video frame from an example.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

It was not easy to find stimuli where the alignment between target word and gesture was easily identifiable, and which also met a number of syntactic constraints aimed at reducing syntactic variation (see Section 3.2.3). Therefore, we did not limit the form and function of the gestures. The final selection consisted of 18 one-handed and 12 two-handed gestures. As for the semiotic type, 17 are indexical and non-deictic, 6 are deictic, and the remaining 7 are iconic.

3.2.3. Reducing variability in the stimuli

Since our knowledge of the effects of sentence modality and IS in Maltese is, to date, quite limited, and given the small size of the dataset, we made the decision to restrict our stimuli with the aim of reducing syntactic diversity and phonological confounds. The following conditions served as a guide to the choice of stimuli:

- The stimulus is a relatively short independent clause.
- The main sentence accent is on a content word.
- The clause is not a negative statement.
- The clause is not a question.
- The clause generally has an SVO or SVC order, with few exceptions.
- There are no contrastive accents.
- The speaker does not move their hands, with the exception of the gesture accompanying the target word.
- There is no overlap between the speakers.

As mentioned above, these conditions were purposely meant to limit syntactic diversity in our dataset. The sentences, nevertheless, show wide lexical variation. In particular, the target words differ in almost all examples. This is intended since our goal was to verify, for naturally occurring speech, results that other studies obtained using controlled experimental stimuli that draw conclusions only based on a comparison of the same target vowel in different conditions (Bosker & Peeters, 2021; Krahmer & Swerts, 2007).

Care was taken, as mentioned earlier, not to include stimuli (both with and without accompanying gestures) where contrastive accents were present. We did not, however, otherwise control for differences in IS. Both sets of stimuli include examples displaying a broad focus on a final constituent or on the entire sentence, and others with a narrow focus on the phrase containing the target word.[3] The distribution of the two IS patterns, however, is very similar in the two types of stimulus, with 22 examples of broad focus and 8 of narrow focus in the 'withGesture' condition, whereas the counts are 24 and 6, respectively, in the 'noGesture' one. In addition, the stimuli were presented to the experiment participants (see below) without any preceding context, and the IS of the individual stimulus is therefore difficult, if not impossible, for them to determine out of context. We therefore did not include IS as a factor in our analyses.

A set of 30 fillers was also identified and used in the experiment. These are, in most respects, similar to the experimental stimuli. In these sentences, however, sentence accent falls on a word anywhere in the sentence with or without one or more hand gestures.

[3]Focus is viewed here from an IS perspective and is used to refer to the constituents denoting information that is presented as being new (Lambrecht, 1996).

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

3.2.4. Preparing the stimuli

Since noise levels differed between clips, we used the noise-reduction functionality in Audacity (version 2.4.2) to equalise noise levels across files. To do this, we separated the video from the audio stream using VirtualDub2 (www.virtualdub.com). In order to generate a noise profile for each speaker, we identified and used a stretch of silence of at least 1 s from the channel of that speaker from the relevant MAMCO corpus recording. This noise profile was then used in the noise reduction algorithm in Audacity. In order to avoid sudden visual onsets, we used MATLAB to generate a linear 400-ms fade-in and fade-out for the video files, and zero-padded the noise-reduced audio files with the same amount of silence. The processed files were then recombined using VirtualDub2 and then encoded as MP4 files.

3.2.5. Acoustic properties of target vowels

Here, we give an account of a number of relevant acoustic properties of the target words with a view to getting a better understanding of whether some of these properties are influenced by the presence of a co-occurring gesture. This information will be important for interpreting the results of the perceptual experiment described below.

We start by showing counts of the vowels in the accented syllables of the target words in Table 2.

We have grouped these vowels according to their degree of openness because openness is known to influence F_0 and loudness, parameters that are also relevant for prosody. We know, in fact, that listeners take into account vowel-intrinsic pitch (higher vowels have a higher intrinsic pitch) when judging the prosodic prominence of a syllable (Fowler & Brown, 1997). Note that the vowel 'æ', not a Maltese vowel, has

Table 2. Count of vowels in the accented syllables of the examples analysed, grouped by degree of openness

Openness	Vowel	Total
Close	iː	8
	ɪː	4
	ɪ	5
	uː	2
Subtotal		19
Mid-to-close	ɛɪ	3
	ɛʊ	1
Subtotal		4
Mid	æ	4
	ɛ	8
	ɛː	3
	ɔ	4
Subtotal		19
Open-to-close	ɐʊ	1
	ɐɪ	3
Subtotal		4
Open	ɐː	8
	ɐ	6
Subtotal		14
Grand total		60

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

also been included in the list, since a vowel having this quality is found in these data, mostly in loanwords from English, such as *clan* and *jazz*.

Previous studies have observed effects of co-speech gesture on some of the acoustic properties of the associated stressed syllable, particularly duration (Krahmer & Swerts, 2007), one of the higher formants (F_2) (Bernardis & Gentilucci, 2006; Krahmer & Swerts, 2007) and fundamental frequency (F_0) (Ambrazaitis & House, 2022). A biomechanical explanation for such effects, linking hand movements to movements of the respiratory system, has been suggested by Pouw et al. (2020a) and Pouw et al. (2020b).

Contrary to the studies that found an effect of F_2, our target words contained a wide variety of vowels (see again Table 2). Therefore, measuring F_1, F_2 and F_3 would not have been useful because the dataset is too small to differentiate the effects of vowel quality and prominence on the formant measures. As a consequence, in our acoustic analysis, we focus on duration, pitch and intensity.

In order to conduct the analysis, the orthographic transcription had to be word-aligned. Therefore, the Munich Automatic Segmentation System (Kisler et al., 2017) was run on all the sound files, and the output was manually corrected. An example of the resulting word-by-word segmentation can be seen in the tier below the oscillogram in Figure 3. The stressed vowel of the target word was segmented by hand in a separate tier below the tier containing the word-by-word segmentation of the text.

Praat (Boersma & Weenink, 2022) was then used to manually extract average pitch and intensity values of the stressed vowel in the target word of each example sentence. Due to the fact that slightly divergent values may be computed by the tool in different runs, values were averaged across three different runs. Average sentence pitch and intensity values were then obtained automatically by means of a Praat script that extracts pitch and intensity values at steps of 0.01 s. The script used a pitch floor of 75 Hz and a pitch ceiling of 400 Hz, values that were deemed adequate for the Maltese speakers in the dataset used. Values below or above this range were ignored.

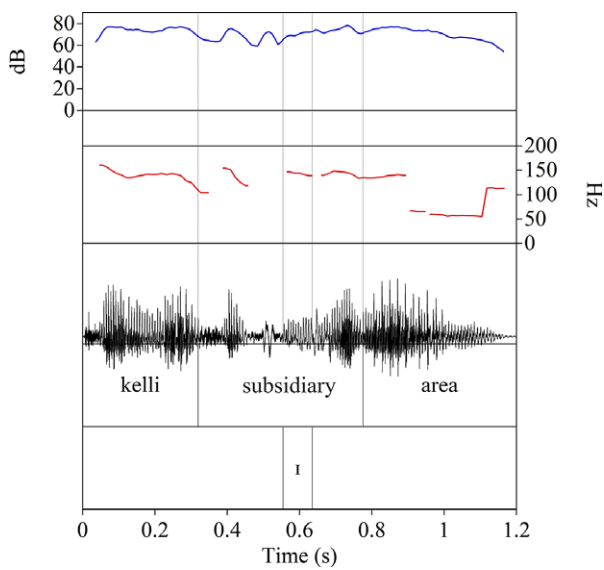


Figure 3. Example stimulus ('withGesture') with, starting in the topmost part of the figure, intensity (in blue), pitch (in red) and the oscillogram.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

Since a number of different speakers were involved, a comparison of the absolute values would not have been meaningful. The measures were normalised by subtracting from them the average pitch and intensity values of the entire sentence in which the word occurs.

For instance, in the example quoted earlier, where the speaker says *kelli subSIdiary area* 'I had a subSIdiary area' and makes a hand gesture that is aligned with the accented syllable *SI*, the pitch and intensity of the vowel are 142.93 Hz and 73.86 dB, whereas the mean pitch and intensity for the entire sentence are 133.23 Hz and 68.01 dB, respectively. In the case of both pitch and intensity, therefore, average values in this example are higher for the stressed vowel compared with those for the sentence as a whole. The acoustic properties of the example are visualised in Figure 3.

Table 3 displays the normalised mean (SD) pitch and intensity values of the stressed vowel in the target words in the 30 'noGesture' examples and the 30 'withGesture' examples. Note that a positive value means that the value of the accented vowel is higher than the average value of the relevant sentence, and vice versa for negative values. The difference between the pitch of the target vowel and the average sentence pitch is expressed in semitones (ST). The table also reports the average duration of the accented vowels in both conditions.

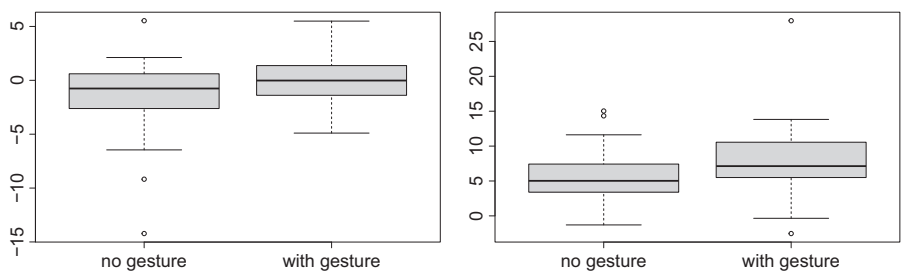
As can be seen, there is no notable difference in vowel duration in the two conditions. This is not very surprising since our dataset included target words containing phonemically short and long vowels as well as diphthongs, so differences in duration due to these sources of variation may well have hidden any possible effect of co-speech gesture. The differences in pitch and intensity between the target vowel and the sentence, on the contrary, are larger if a gesture is present. The distribution of the differences for both types of measures is shown in the boxplots in Figure 4, with the distribution for the 'noGesture' examples displayed on the left in both plots.

Two mixed-effects models were tested to predict (i) the difference between the target vowel pitch and the average pitch of the corresponding sentence and (ii) the difference between the target vowel intensity and the average intensity of

Table 3. Mean (SD) values for pitch, intensity and duration of the stressed vowels in the target words

Condition	Pitch (ST)	Intensity (dB)	Duration (s)
Without gesture	−1.43 (3.77)	5.4 (4.14)	0.10 (0.03)
With gesture	0.05 (2.05)	7.9 (5.39)	0.11 (0.05)

Note: Pitch and intensity values were normalised by subtracting the mean sentence value for each example. The pitch value is expressed in semitones.



(a) Difference in pitch (ST) (b) Difference in intensity (dB)

Figure 4. Acoustic differences between the target vowel and the sentence.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

Table 4. Predicting vowel pitch from gesture and vowel openness: model coefficients; significant p values are shown in boldface

	Coef. β	SE (β)	*t*	*df*	*p*
Intercept	−1.48	0.89	−2.39	22.23	**<0.05**
Gesture	1.53	0.76	2.02	48.39	**<0.05**
Openness	−0.58	1.04	−0.56	56.45	0.58

Table 5. Predicting vowel intensity from gesture and vowel openness: model coefficients; significant p values are shown in boldface

	Coef. β	SE (β)	*t*	*df*	*p*
Intercept	5.54	0.89	6.20	29.65	**<0.0001**
Gesture	2.34	1.25	1.87	48.19	0.067
Openness	1.60	1.65	0.97	56.68	0.33

the corresponding sentence. In both cases, the difference in acoustic property, either in pitch or in intensity, is the dependent variable, and the condition ('noGesture' *vs.* 'withGesture') is used as a fixed factor. We attempted to use both speaker and vowel as factors in both cases. However, the effect structure was too complex to be supported by the data. Therefore, vowel openness rather than the full range of vowels was used as an additional fixed effect, while speaker was added as a random effect.

The coefficients of these models can be inspected in Table 4 for pitch and Table 5 for intensity. We see that the effect of gesture in the case of intensity only approaches significance, whereas it is significant in the prediction of pitch. If the gesture is present, the difference in pitch between the target vowel and the sentence average increases. This preliminary exploration of some of the acoustic properties of the target words in the presence of a co-occurring gesture suggests some interesting avenues for more in-depth study. The variability in the naturally occurring stimuli does not, however, lend itself well to further analysis at this point.

But in general, the finding that the target words that were accompanied by gestures had stronger cues to prosodic prominence than target words without accompanying gestures makes it reasonable to ask whether these stronger cues have an influence on prominence perception.

3.3. Procedure

The stimuli described in Section 3.2 were used to build the experiment used in this study. This was run online using JSPsych (de Leeuw, 2015). All materials used, including the explanations, were presented to participants in Maltese. The experiment began with a short questionnaire asking participants about their age, gender, language background, self-rated proficiency and patterns of use. This was followed by a training phase, in which they were exposed to six examples based on two sentences.

First, they heard three versions of the sentence *Jien qed nistudja l-mediċina* 'I am studying medicine': one with the pronoun 'I' carrying the sentence accent; another with the sentence accent on 'medicine', however with broad focus signalling a 'neutral' rendering; and a third with a sentence accent again on 'medicine', but this

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

time with narrow focus. In these utterances, *l-mediċina* became successively longer, louder and produced with a higher maximum pitch as we moved from the version with an accent on *Jien*, to the unmarked utterance, and finally to the version with an accent on *l-mediċina*. Moreover, in the utterances with a narrow focus on either *Jien* or *l-mediċina*, there was a clear pitch accent on that element. Participants were told that in the first case, the word *l-mediċina* might be rated as *one* or *two* on a scale of importance/prominence from one to seven, as either *three* or *four* in the second case, and *five* or *six* in the third. Note that we are using a wider scale than is usually employed in linguistic coding of prosodic prominence, for example, in Krahmer and Swerts (2007), since a scale with more levels may make it easier to reach a decision (as argued for segmental perception by Apfelbaum et al., 2022).

Next, participants were asked to rate three versions of the sentence *Iz-ziju għandu kelb tal-kaċċa* 'My uncle has a hunting dog' (literally 'The uncle has a dog for hunting'). They were asked to rate the prominence of *kelb tal-kaċċa* 'hunting dog' when it was in narrow focus, when the sentence was pronounced signalling broad focus, that is, with a 'neutral' rendering, and when *iz-ziju* 'the uncle' was in narrow focus. Participants saw the whole sentence written on the screen, with the element they had to judge in terms of prominence highlighted in colour. After listening to the sentence, they could rate the prominence of the highlighted element on a scale from 1 to 7, or press the 'R' key to listen to the whole sentence again.

After this practice phase, the main experiment started. Participants were asked to listen to a number of Maltese sentences and rate the prominence of a target word. All participants were exposed to both audio-only and audiovisual conditions. Each trial started with a visual display of the whole sentence in written form, together with a prompt requiring them to indicate how important they considered the target element in that sentence. An example is shown below:

Il-fraži hija: 'għandi xi ħbieb minn Architecture'.
'The sentence is: "I have some friends who study Architecture"'.
Kemm hi importanti: 'Architecture'?
'How important is: "Architecture"?'

The display also mentioned that the audio or video clip could be replayed by pressing 'R' if participants were unsure of their rating. We gave participants the option to repeatedly hear the stimuli, because speech stimuli from a spontaneous speech corpus taken out of context are often difficult to process (Brouwer et al., 2013). This differs from what has been done in other studies (e.g., Krahmer & Swerts, 2007) that use a fixed utterance, usually recorded carefully as read speech. Participants saw the sentence for 2.5 s before the audio or video clip started playing. The experiment only proceeded to the next trial when participants had rated the importance of the highlighted word, with an inter-trial interval of 50 ms.

The experiment was blocked by modality condition, so that participants first rated 45 sentences (15 stimuli 'withGesture', 15 'noGesture' stimuli and 15 fillers) presented either as audio-only or audiovisual stimuli, and then 45 stimuli (again 15 × 3) in the other condition. Participants were given the opportunity to take a short break after one block of 45 trials. Most participants finished the study in 20–30 min. Block order was counterbalanced across participants, and stimuli used in each condition were counterbalanced across participants, so that each stimulus appeared as often in both the audiovisual and visual conditions in the complete dataset.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

3.4. Statistical analysis

The data were analysed using a linear mixed-effects model to accommodate the fact that we have a sample of participants and a sample of stimuli (Westfall et al., 2014). We used the lmerTest package (v3.1-3) in R (v4.1.3). The data and analysis code are available on the Open Science Framework.

The dependent variable is prominence rating. The analysis used stimuli and participants as random factors and started with a maximal random-effect structure (but note that the factor Gesture is between stimuli, rendering a random slope of Gesture over stimuli nonsensical). Due to convergence issues, we removed the correlations between random effects and the random slope of the interaction over participants. The fixed factors were Modality (audio-only *vs.* audiovisual) and Gesture ('withGesture' *vs.* 'noGesture') and their interaction. The fixed factors were contrast coded (Modality: audio-only = −0.5, audiovisual = 0.5; Gesture: 'noGesture' = −0.5, 'withGesture' = 0.5). With this contrast coding, the regression weights show the mean differences between the conditions, and interactions and main effects are linearly independent (i.e., the interaction does not influence the main effects).

4. Results

We first investigated whether the task, asking participants to rate the importance of a given word in a sentence, managed to get them to rate the prosodic prominence of the utterances, by analysing the responses obtained in the training phase. As indicated in the description of the procedure, participants rated three versions of *Iz-ziju għandu kelb tal-kaċċa.* 'My uncle has a hunting dog' (literally 'The uncle has a dog for hunting'), one with a broad focus and two with a narrow focus on either *Iz-ziju* or *kelb tal-kaċċa.* The participants rated the 'importance' of *kelb tal-kaċċa* as 1.16 if there was a narrow focus on *Iz-ziju*, as 3.57 if there was a broad focus and as 4.98 if there was a narrow focus on *kelb tal-kaċċa.* With a standard deviation of around 1.3 across these items, this represents a large effect. Unsurprisingly, this effect was significant in a repeated-measure ANOVA, $p < 0.001$. This indicates that the task was sensitive to eliciting responses that reflect the cueing of prosodic prominence.

Figure 5 shows the mean prominence ratings for all six conditions (including the fillers). Target words in 'noGesture' stimuli were generally rated as less prominent than those in the 'withGesture' stimuli. There is no overall effect of Modality, as, in the 'noGesture' condition, audio-only stimuli receive higher ratings than audiovisual stimuli, while the reverse is observed in the 'withGesture' condition. This suggests an interaction between the factors based on the descriptive data.

The linear mixed-effects model (see Table 6) only fully confirms the first two observations from the descriptive data. These observations were that there is a clear effect of the Gesture condition and no effect of the Modality condition. The third observation was that the effect of having a gesture is larger in the audio-visual modality, but the corresponding interaction is only marginally significant.

Our first hypothesis was verified, in that the model found a clear effect of the Gesture condition: the target words accompanied by a gesture tend to be rated as more prominent by listeners. However, the interaction of Modality by Gesture, which would indicate that seeing gestures enhances the perceived prominence, only approaches significance; thus, contrary to what our second hypothesis predicted, the small interaction effect provided by the audiovisual modality does not reach significance.

Based on the request of a reviewer, we performed an exploratory analysis of whether the order of the blocks (audiovisual then audio-only *vs.* audio-only then

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

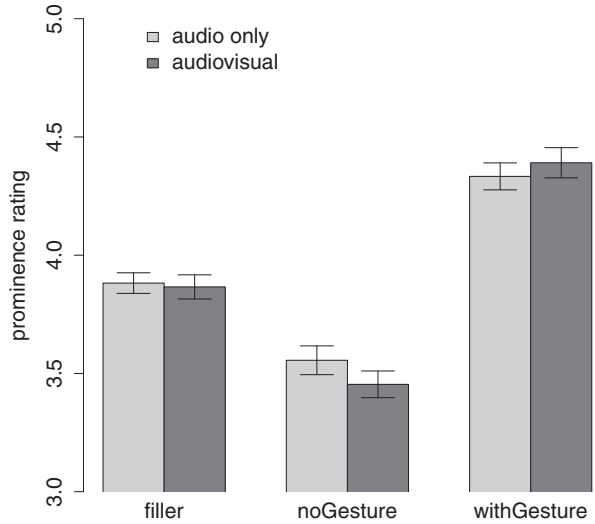


Figure 5. Mean prominence ratings for all six conditions (including the fillers) with error bars based on the method of Morey (2008).

Table 6. Results from the linear mixed-effects model of the prominence ratings; significant p values are shown in boldface

	Coef. β	SE (β)	*t*	*df*	*p*
Intercept	3.93	0.11	36.7	105.1	**<0.0001**
Modality	−0.02	0.06	−0.4	68.6	0.683
Gesture	0.86	0.19	4.6	66.2	**<0.0001**
Gesture × Modality	0.15	0.09	1.7	55.3	0.088

audiovisual) influenced the ratings. To that end, we added to the model a contrast-coded predictor for block order interacting with the other two factors. Comparing this model with the initial model that did not take into account block order indicated that block order did not affect the pattern of results ($\chi 2(4) = 5.504, p = 0.239$). Moreover, none of the individual beta weights for the order factor were close to significance (min(p) = 0.12).

5. Discussion

As we have seen, in our experiment, the presence of an accompanying hand gesture can increase the strength of some acoustic cues to the prominence of target words, and consequently their perceived prominence for a listener. In spite of this, no effect of modality (audio-only *vs.* audiovisual) nor any interaction between gesture condition and modality was found. While words accompanied by a gesture were rated by listeners as more prominent than those without a co-occurring gesture, the prominence difference between these two sets of stimuli was not enhanced in the audiovisual condition. This suggests that words that are accompanied by gestures are also more prominent in the auditory domain and that the visibility of the gestures does not further increase the perceived prominence, at least not in a significant way. The

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

failure to find a significant effect of the audiovisual modality contrasts with the results by Bosker and Peeters (2021), who found clear evidence of an effect due to beat gestures with a smaller sample of participants and items (though items were used repeatedly in their experiments). However, our results do not necessarily contradict Bosker and Peeters's. The stimuli in the Bosker and Peeters study were acoustically ambiguous, and even with a matching gesture, they were never perceived as unambiguously prominent on the first or second syllable. Our stimuli, in contrast, were selected because they are produced in a prominent fashion, and for such stimuli, there might be diminishing effects of the visibility of gestures, given that the auditory modality is relatively strong in and by itself. Therefore, our results are probably more in line with a small additional effect of gestures on prominence perception (given acoustic prominence to begin with) than with a null effect.

The acoustic analysis of the stimuli shows that the presence of gestures has some effect on the realisation of prominence, as also observed by others (Bernardis & Gentilucci, 2006; Krahmer & Swerts, 2007). The difference in intensity between target words and the rest of the sentence is larger, on average, when words are accompanied by gestures than when they are not. However, the effect of gesture presence on intensity only approaches significance. In contrast, we found a small effect of gesture co-occurrence on pitch: on average, the difference between the pitch of the vowel and the mean pitch of the sentence is larger when a gesture is present than when it is not. However, the normalisation method we applied is a simple difference of pitch values between the target vowel and the sentence average. A more fine-grained measure that captures F_0 movement in the vicinity of the stressed vowel might yield a different result. Finally, no effect was found for duration.

Thus, our results diverge somewhat from the findings reported in Krahmer and Swerts (2007), who found significant effects on target words due to duration and F_2 values. Their analysis, however, contrasts two target words in the same sentence rather than one target word and the associated sentence as we do. Furthermore, only one vowel is represented in the materials on which that study is based. Since, in contrast, a range of different vowels occur in our examples, it made little sense for us to even look at formants.

One could discuss whether the small effect we do see on the pitch produced by speakers when they also produce a gesture on the target word is actually caused by the gesture. A causal relationship has been suggested in connection with arm and wrist movements, for example, by Pouw et al. (2021), who found that participants asked to retell a cartoon in three different gestural conditions (no gesture, wrist movement and larger arm movement) showed increased F_0 values in the two gestural conditions compared with the passive position. This difference, however, was only significant for arm movements. Generally, however, a correlation between the presence of gestures and acoustic properties could have other sources, such as a common motor system, as suggested, for example, by Parrell et al. (2014).

While this study has managed to show, in naturally occurring multimodal data in Maltese, that the perceived auditory prominence of a target word is greater if a hand gesture is produced by the speaker while uttering the word, we must also note that extracting stimuli from naturally occurring data may also have created a number of confounds, making the audiovisual condition not totally 'clean', and thus diminishing the likelihood of observing a potential effect.

A potential source of complexity, for example, could be due to different types of pitch accents. Ambrazaitis and House (2022) discussed how different types of lexical

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

accents in Swedish are affected differently by the presence of visual beats, with larger effects seen on accentual rises than on falls. We cannot exclude that such differences may also hold for our data, although care was taken to ensure that target words in the stimuli used in the experiment were restricted to the part of the utterance carrying the nuclear pitch accent rather than the phrase accents typical on post-nuclear stretches of speech in Maltese (see Vella, 1995, 2003).

Another source of variation we did not control for relates to different properties of the hand gestures. All the gestures in the examples have a beat-like quality, but about 40% of them also have a deictic component or iconic properties, and both one-handed and two-handed gestures are present. Although no detailed analysis of the kinematic characteristics of the gestures was conducted to investigate the possible impact of gesture size and velocity, there is probably variation in all these dimensions. The overall complexity may have diluted the general effect on prominence perception.

6. Conclusion and future work

Our study has shown that, on the one hand, the perceived auditory prominence of a target word is greater if a hand gesture is produced by the speaker while uttering the word. Importantly, our study uses examples from natural conversational data to confirm effects established previously on the basis of highly controlled multimodal stimuli (Bolinger, 1986; Krahmer & Swerts, 2004) or read text (Ambrazaitis et al., 2020a; Ambrazaitis & House, 2022). On the other hand, however, whether the participants see the gesture or not does not make a significant difference in our study. An interpretation of these two results leads us to suggest that gestures appear to provide an additional cue to prominence, but are not in themselves a necessary prominence cue for the listener.

A number of follow-up studies could be carried out to further corroborate and understand the results presented here. One possibility is certainly that of looking in more detail at the individual stimuli used in the experiment, to see whether other factors may be at play in the perception of prominence in the different conditions. One such factor that certainly merits some probing is constituent order, in particular the fact that the use of certain syntactic constructions in some of the examples may have an additive effect on perception prominence.

Another direction would be to study in more depth in what way prominence perception is affected by different types of gestures, both in terms of their different functions and their different kinematic characteristics. Yet another would be to investigate the effect of gestures on the perception of prominence where sentence accent is not placed towards the end of a sentence, or where it is combined with a phrase accent following the principles explained in Section 2.2. However, it is likely to be a challenge to find enough relevant examples of these more complex phenomena in naturally occurring data, so it may be necessary to revert to the use of elicited experimental data.

It would also be interesting to extend our investigation to multimodal stimuli involving facial beats. It must be noted, however, that in the design phase of our study, we searched for relevant examples of head movements co-occurring with sentence accents, but it proved impossible to isolate a sufficient number of short sentences in which only one head movement was present. To solve this problem, one possible avenue could be to have artificial agents perform our original stimuli with the

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

addition of head movements following a methodology employed in several other studies to test the perception of multimodal signals (Heylen et al., 2011; House et al., 2001; Prieto et al., 2015).

Acknowledgements. We acknowledge the work of our research assistant, Amanda Muscat. We also thank the reviewers for their useful comments.

Funding statement. This work was supported by the University of Malta's Research Seed Fund.

References

Alahverdzhieva, K., & Lascarides, A. (2010). Analysing speech and co-speech gesture in constraint-based grammars. In S. Müller (Ed.), *Proceedings of the HPSG10 conference* (pp. 6–26). CSLI Publications.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In J.-C. Martin, P. Paggio, P. Kuehnlein, R. Stiefelhagen, & F. Pianesi (Eds.), Multimodal corpora for modelling human multimodal behaviour. Special Issue of the *International Journal of Language Resources and Evaluation*, 41, pp. 273–287.

Ambrazaitis, G., Frid, J., & House, D. (2020a). Word prominence ratings in Swedish television news readings: Effects of pitch accents and head movements. In *10th international conference on speech prosody 2020* (pp. 314–318). The International Speech Communication Association (ISCA).

Ambrazaitis, G., Frid, J., & House, D. (2022). Auditory vs. audiovisual prominence ratings of speech involving spontaneously produced head movements. In *Speech prosody 2022, Lisbon, Portugal, May 23–26, 2022*, International Speech Communication Association (pp. 352–356).

Ambrazaitis, G., & House, D. (2022). Probing effects of lexical prosody on speech–gesture integration in prominence production by Swedish news presenters. *Laboratory Phonology*, 24(1), 1–35. https://doi.org/10.16995/labphon.6430

Ambrazaitis, G., & House, D. (2023). The multimodal nature of prominence: Some directions for the study of the relation between gestures and pitch accents. In *Proceedings of the 13th international conference of Nordic prosody* (pp. 262–273). Sciendo.

Ambrazaitis, G., Zellers, M., & House, D. (2020b). Compounds in interaction: Patterns of synchronization between manual gestures and lexically stressed syllables in spontaneous Swedish. In *Gesture and speech in interaction (GESPIN2020), 7–9 September 2020, Stockholm, Sweden*. KTH Royal Institute of Technology.

Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, 152(6), Article 3728.

Berger, S., & Zellers, M. (2022). Multimodal prominence marking in semi-spontaneous YouTube monologs: The interaction of intonation and eyebrow movements. *Frontiers in Communication*, 7, Article 903015.

Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, 44(2), 178–190.

Boersma, P., & Weenink, D. (1992–2022). *Praat: Doing phonetics by computer*. Version 6.1.08. http://www.praat.org/

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford.

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, 288(1943), Article 20202419.

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, 34, 519–539.

Brucks, M. S., & Levav, J. (2022). Virtual communication curbs creative idea generation. *Nature*, 605(7908), 108–112.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F_o variations. In *Proceeding of fourth international conference on spoken language processing (ICSLP'96)* (Vol. 4, pp. 2175–2178). IEEE.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6), 871–887.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12.

de Ruiter, J. P. (2000). The production of gesture and speech. *Language and Gesture*, 2, 284–311.

Debreslioska, S., & Gullberg, M. (2020). The semantic content of gestures varies with definiteness, information status and clause structure. *Journal of Pragmatics*, 168, 36–52.

Dohen, M., & Loevenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, 52(2–3), 177–206.

Drijvers, L., & Holler, J. (2022). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, 30, 792–801.

Ebert, C., Evert, S., & Wilmes, K. (2011). Focus marking via gestures. In *Proceedings of Sinn und Bedeutung 15*. Universaar – Saarland University Press.

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

Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., & d'Imperio, M. (2022). Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Developmental Science*, 25(1), Article e13154.

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

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Language resources and evaluation* (Vol. 6, pp. 86–91). Workshop on Multimodal Corpora.

Fowler, C. A., & Brown, J. M. (1997). Intrinsic *f*0 differences in spoken and sung vowels and their perception by listeners. *Perception & Psychophysics*, 59(5), 729–738.

Gentilucci, M., & Dalla Volta, R. (2007). The motor system and the relationships between speech and gesture. *Gesture*, 7(2), 159–177.

Glave, R., & Rietveld, A. (1979). Bimodal cues for speech loudness. *The Journal of the Acoustical Society of America*, 66(4), 1018–1022.

Grice, M., & Kügler, F. (2021). Prosodic prominence – a cross-linguistic perspective. *Language and Speech*, 64(2), 253–260.

Grice, M., Ladd, D. R., & Arvaniti, A. (2000). On the place of phrase accents in intonational phonology. *Phonology*, 17(2), 143–185.

Gussenhoven, C. (1983). Focus, mode and the nucleus. *Journal of Linguistics*, 19(2), 377–417.

Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., & Schröder, M. (2011). Generating listening behaviour. In *Emotion-oriented systems* (pp. 321–347). Springer.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 89–97.

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, 25(5), 1900–1908.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of the Seventh European conference on speech communication and technology*, 387–390.

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

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

Kendon, A. (1980). Gesture and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *Nonverbal communication and language* (pp. 207–227). Mouton.

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

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347.

Krahmer, E. & Swerts, M. (2004). More about brows. In *From brows to trust* (pp. 191–216). Springer.

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.

Krauss, R. M., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. In *Gesture, speech, and sign* (pp. 93–116). Oxford.

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*, 8(1), Article 3.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Volume 71 of Cambridge Studies in Linguistics. Cambridge University Press.
Leonard, T., & Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471.
Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2), 179–214.
Loehr, D. P. (2004). *Gesture and Intonation* [Doctoral dissertation]. Georgetown University.
McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic research*, 27(1), 69–89.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
McNeill, D. (2008). *Gesture and thought*. University of Chicago Press.
Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau. *Tutorials in Quantitative Methods for Psychology*, 4, 61–64.
Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., & Navarretta, C. (2010). The NOMCO multimodal Nordic resource – goals and characteristics. In *Proceedings of the seventh conference on international language resources and evaluation (LREC'10)* (pp. 2968–2973). European Language Resources Association (ELRA).
Paggio, P., Galea, L., & Vella, A. (2018). Prosodic and gestural marking of complement fronting in Maltese. In A. Gatt & P. Paggio (Eds.), *The languages of Malta* (Vol. 18, pp. 81–116). Language Science Press.
Paggio, P. & Vella, A. (2014). Overlaps in Maltese conversational and task-oriented dialogues. In *Proceedings from the 1st European symposium on multimodal communication* (pp. 55–64). Linköping Electronic Conference Proceedings.
Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, 42, 1–11.
Peirce, C. S. (2009). *Writings of Charles S. Peirce: A chronological edition, volume 8: 1890–1892*. Indiana University Press.
Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2021). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, 1491(1), 89–105.
Pouw, W., Harrison, S. J., & Dixon, J. A. (2020a). Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, 149(2), 391–404.
Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020b). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, 52, 723–740.
Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49, 41–54.
Vella, A. (1995). *Prosodic structure and intonation in Maltese and its influence on Maltese English* [Doctoral dissertation]. The University of Edinburgh.
Vella, A. (2003). Phrase accents in Maltese: Distribution and realisation. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 1775–1778). International Phonetic Association.
Vella, A. (2009a). Maltese intonation and focus structure. In R. Fabri (Ed.), *Maltese linguistics: A snapshot. In memory of Joseph A. Cremona* (Vol. 1, pp. 63–92). Lingwa Tagħna.
Vella, A. (2009b). On Maltese prosody. In B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz, & M. Vanhove (Eds.), *Introducing Maltese linguistics* (pp. 47–68). John Benjamins.
Vella, A. (2013). Languages and language varieties in Malta. *International Journal of Bilingual Education and Bilingualism*, 16(5), 532–552.
Vella, A., & Grice, M. (2024). Association of tones in Maltese. In *Prosodic typology III: The phonology of intonation and phrasing* (pp. 1–24). Oxford University Press.
Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.
Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology*, 143(5), 2020–2045.
Willems, R. M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain and Language*, 101(3), 278–289.

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

A. Appendix

In all the examples below, the syllable containing the main sentence accent is capitalised. In the examples with gestures, the hand gesture stroke is aligned with the accented syllable. The order of constituents in the original Maltese examples is retained in the translations provided below, although this may make them sound unidiomatic in English. In the examples involving pro-drop, the pronoun is placed in square brackets in the translation; the same holds for the verb in sentences that are verbless in Maltese. In a few of the translations, elements that may help increase the idiomaticity are also included in square brackets.

A.1. Examples with gestures

1. *ee jiena min-naħa tas-SOUTH*
   'yeah I'm from the south side'
2. *aħna nibdew mis-SEcond year*
   'we start in the second year'
3. *sixth year insiru avuKAti*
   '[in the] sixth year [we] become lawyers'
4. *t'id tkun avuKAT*
   '[you] have to be a lawyer'
5. *jien kelli em IngLIŻ*
   '[I] had um English'
6. *kelli interMEdiate*
   '[I] had intermediate'
7. *kelli subSIdiary area*
   '[I] had a subsidiary area'
8. *għandi naq'a probLEma*
   '[I] have a bit of a problem'
9. *nħobb ħafna nSAjar ukoll*
   '[I] love a lot cooking as well'
10. *għamilna naq'a fiGOLli*
    '[we] made some figolli'
11. *l'aqwa li m'hemmx xagħar ABjad u*
    'the most important thing is that there is no white hair and'
12. *il-mużika BAXxa*
    'the music [is] low'
13. *sal-aħħar tax-XAHAR għandhom*
    'till the end of the month [they] have'
14. *imma kelli IngLIŻ*
    'but [I] had English'
15. *di' waħda minnhom is-celeBRAtions allura jew?*
    'this [was] one of them the celebrations so or?'
16. *ħafna nies it-teQUIla jdejjaqhom ħafna*
    'many people the tequila [it] bothers them a lot'
17. *taraha bilQIEGĦda*
    '[you] see her sitting'
18. *le quantu tal-kwalità tal-MUżika*
    'no depending on the quality of the music'
19. *diffiċli social LIFE għalina*
    '[it is] difficult social life for us'
20. *inħobb immur NIEkol*
    '[I] love going to eat'
21. *inħobb immur nara FILM*
    '[I] love going to watch a film'
22. *inħobb noħroġ JIEna 'iġifieri*
    '[I] love to go out me so'

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

23. *infatti jkollna CLAN*
'In fact [we] have a clan'
24. *minn Ħal Tarxien, xagħarha NOKkli*
'from Ħal Tarxien, her hair [is] curly'
25. *jiena iktar volontarJAT*
'I [do] more voluntary work'
26. *jien KAren*
'I [am] Karen'
27. *qiegħda l-universiTA'*
'[I] am at the university'
28. *il-ballu li GĦADda kien*
'the ball that passed it was'
29. *kien ikun hemm il-BASketball*
'[There] used to be the basketball'
30. *tista' ma' tagħmel XEJN*
'[You] can do nothing'

A.2. Examples without gestures

1. *fis-sajf ikun l-aħjar Wied il-GĦAJN*
'in the summer [it] is best [at] Wied il-Għajn'
2. *emozzjoniJIET qiegħda nagħmel*
'emotions [I] am working on'
3. *minn Ħ'AtTARD jien*
'from Ħ'Attard I [am]'
4. *għandi xi ħbieb minn ArchiTECture*
'[I] have some friends from Architecture'
5. *fih ukoll ee ĦAmes snin*
'[it] has as well eh five years'
6. *jien qed nagħmel ArchiTECture*
'I am doing Architecture'
7. *mort xi tlett SNIN sentej'*
'[I] went about three years two years'
8. *qed issa t-tiel't SEna u*
'[I] am now [in] the third year and'
9. *ijja kelli BŻONN jien*
'yes [I] needed I'
10. *jiena gradwajt sa third year LIġi*
'I graduated until third year law'
11. *qed nagħmel kriminoloĠIja*
'[I] am doing criminology'
12. *jien għandi dsatax-il SEna*
'I have nineteen years'
13. *jiena qalli sieħbi NIġi imma*
'I [he] told me my friend to come but'
14. *jie' ħa nagħlaq twenty three dis-SEna*
'I will be twenty three this year'
15. *haw' sabiħ IMma haw'*
'here [is] nice but here'
16. *qed niftehmu fil-PARty*
'[we] are agreeing in the party'
17. *vera tajba kienu FL-AĦħar*
'really good [they] were in the end'
18. *tagħmilha l-LUmi jiġ'ieri*
'make it the lemon so'

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

19. *llum il-ġurnata ma tantx ikonna kunċerti tal-JAZZ*
'today the day [we] do not have many jazz concerts'
20. *rari ssib xi ħadd BĦAlek*
'[it is] rare to find someone like you'
21. *kont għamilt ħin nisTENna bilwiefqa*
'[I] had done [some] time waiting standing'
22. *le hekk sabiħ uKOLL*
'no like that [it is] nice too'
23. *qed nistudja accounts u BANking*
'[I] am studying accounts and banking'
24. *qiegħda second YEAR*
'[I] am [in] second year'
25. *ħabibti tmu' masqueRADE*
'my friend goes [to] Masquerade'
26. *tri' tkun deĊIża*
'[you] must be definite'
27. *fil-verita' fil-BIdu jidħlu ħafna*
'in truth in the beginning [they] enter a lot'
28. *em għandi TWENty*
'em [I] have twenty'
29. *intom kontu tagħmlu l-ĠLIED*
'you used to make fights'
30. *konna għamilna żmien ħbi'b SEW*
'[we] had made [some] time good friends'

A.3. Fillers

1. ee minn Ħaż-ŻABbar
'eh from Ħaż-Żabbar'
2. *forsi għandna xi ħbieb in KOmuni mela*
'maybe [we] have some friends in common then'
3. *noqGĦOD Ħaż-Żabbar*
'[I] live [in] Ħaż-Żabbar'
4. *hemm l-istatWA eħe*
'there [is] the statue yes'
5. *toqgħod targuMENta qisek avukat*
'[you] stay arguing like [a] lawyer'
6. *ĊEMpluli mill-fakultà*
'[they] called me from the faculty'
7. *jkun haw' wisq SĦANA*
'[it] is too much heat'
8. *meta ikolli ċans għax iFHIMni*
'when [I] have chance because [you] understand me'
9. *użgur għax, m'hemmx GĦALfejn BEd*
'of course because, there is no need [for a] bed'
10. *qed jBATtu n-nies*
'[they] are decreasing the people'
11. *JIEna Karen*
'I [am] Karen'
12. *aħjar milli għaLAQT*
'better than [it] closing'
13. *la lesta TAJtha*
'since [it is] ready [I] gave it'
14. *għaMILT tlett ijiem*
'[I] did three days'

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press

15. *disserTAtion ukoll*
'dissertation as well'
16. *le PARty bil-*
'no party with'
17. *morna naraw OPra*
'[we] went to see [an] opera'
18. *meta tpoġġi għandek LUSsu*
'when [you] sit [you] have luxury'
19. *għax bħalissa BROKE*
'because right now [I am] broke'
20. *daw' id-diski tal-PARties jew hekk*
'these songs of [the] parties or that'
21. *twenty TWO ħa nagħlaq. F'June propjament.*
'twenty two [I] will be. In June actually'
22. *jiddependi ĦAFna mi… miċ-ċirkustanzi*
'[it] depends a lot o… on the circumstances'
23. *eżatt allura tiġini awtoMAtika il-ħaġa*
'exactly so [it] comes automatic[ally] the thing'
24. *kull m'għandi ERbgħa*
'all [I] have is four'
25. *t'id tifFOkka*
'[you] have to focus'
26. *kont immur MasqueRADE*
'[I] used to go to Masquerade'
27. *F'GĦAxar snin mhux tilħaq*
'in ten years not [you] don't reach'
28. *għamilna reUNion milux*
'[we] did a reunion not long ago'
29. *għax propj'ment għandu BAby*
'because actually [he] has a baby'
30. *m'għadx hemm dik il-ĦEĠġa ta' futbol*
'[there] isn't there still that enthusiasm of football'

Cite this article: Paggio, P., Mitterer, H., Attard, G., & Vella, A. (2025). Do hand gestures increase perceived prominence in naturally produced utterances?, *Language and Cognition*, 17, e54, 1–26. https://doi.org/10.1017/langcog.2025.20

https://doi.org/10.1017/langcog.2025.20 Published online by Cambridge University Press