Gesture-speech coupling in L2 lexical stress production: A pre-registration of a speech acoustic and gesture kinematic study

Version: 1.2

Hans Rutger Bosker[1,3], Marieke Hoetjes[2], Wim Pouw[1,3], Lieke van Maastricht[2]

(shared first authorship among all authors; listed in alphabetical order)

[1]Max Planck Institute for Psycholinguistics

[2] Centre for Language Studies, Radboud University

[3]Donders Institute for Brain, Cognition and Behaviour

Author Note: The pilot data, scripts and notebook supporting this pre-registered report can be found on the Open Science Framework (https://osf.io/kqse2/) and Github (https://github.com/WimPouw/StressInMotion/).

Time stamp: *This pre-registration is not time-stamped yet. After a preliminary feedback phase, we will update the pre-registration and time stamp it on the OSF.*

Contact: hansrutger.bosker@mpi.nl; marieke.hoetjes@ru.nl; wim.pouw@donders.ru.nl; lieke.vanmaastricht@ru.nl

Abstract

The prosody of a second language (L2) is notoriously difficult to acquire. It requires the mastery of a range of nested multimodal systems, including articulatory but also gestural signals, as hand gestures are produced in close synchrony with spoken prosody. It remains unclear how easily the articulatory and gestural systems acquire new prosodic patterns in the L2 and how the two systems interact, especially when L1 patterns interfere. This interdisciplinary pre-registered study investigates how Dutch learners of Spanish produce multimodal lexical stress in Spanish-Dutch cognates (e.g., Spanish *profeSOR* vs. Dutch *proFESsor*). Acoustic analyses assess whether gesturing helps L2 speakers to place stress on the correct syllable; and whether gesturing boosts the acoustic correlates of stress through biomechanic coupling. Moreover, motion-tracking and time-series analyses test whether gesture-prosody synchrony is enhanced for stress-matching vs. stress-mismatching cognate pairs, perhaps revealing that gestural timing is biased in the L1 (or L2) direction (e.g., Spanish *profeSOR* with the gesture biased towards Dutch stressed syllable -*fes*). Thus, we will uncover how speakers deal with manual, articulatory, and cognitive constraints that need to be brought in harmony for efficient speech production, bearing implications for theories on gesture-speech interaction and multimodal L2 acquisition.

Keywords: gesture; prosody; gesture-speech coupling; second language acquisition; lexical stress.

Introduction

Spoken language is a complex aggregate of multiple oscillating systems working in concert, including respiratory cycles, vocal-fold vibrations, as well as tongue, body, jaw, and labial movements. As if the complexity of producing speech isn't challenging enough, spoken language incorporates manual communicative gestures that are seamlessly produced in close synchrony to the prosody of speech (Bolinger, 1983). This coupling of gesture and speech is so natural that when one is instructed to modulate production in one modality, for instance by placing an intended emphatic stress on a syllable or increasing the hand movement amplitude, this will unintentionally lead to a comparable increased amplitude/stress in the other modality — at least when speaking in one's native language (henceforth L1; Esteve-Gibert & Prieto, 2013; Krahmer & Swerts, 2007; Parrell et al., 2014; Rochet-Capellan et al., 2008). These challenges involved in speech-gesture coordination surface perhaps most pertinently in L2 prosody production. For instance, when a Dutch learner of Spanish wants to produce the cognate *profeSOR* in Spanish (cf. Dutch *proFESsor*, capitals reflect lexical stress), they not only need to know that the stress falls on *-sor* (not *-fes*), but also *how* to coordinate - in a timely fashion - the articulatory movements required to produce stress on the correct syllable, and align their manual gestural movements accordingly.

This study combines techniques and knowledge from the areas of cognitive psychology, human movement, gesture studies, phonetics, and L2 acquisition. Taking an interdisciplinary approach, we aim to understand a) how gesture and speech coordinate when L2 prosody is not yet mastered, b) how gesture and speech are organized when L2 and L1 prosody compete, thus testing c) whether and how manual gesture supports the speech system. As such, we will uncover how bodily skills and routines are negotiated with competing cognitive constraints during spoken language. This will inform theories on gesture-speech interaction and multimodal L2 acquisition. Before turning to the current experiment, the relevant literature on (L2) prosody, gesture-speech synchrony, gesture, L2 acquisition, and lexical stress production in Spanish and Dutch is reviewed.

Theoretical background

Prosody in communication

Prosody plays a vital role in spoken communication. Traditionally, prosody is viewed as implicating the pragmatics of conversations, most pertinently information structure (highlighting new information; Chen, 2012), speech acts (question vs. statement intonation; Xie et al., 2021), and conversation management (turn-taking; Bögels & Torreira, 2015). However, rather than only serving a pragmatic function, prosody directly shapes a wide range of cognitive processes involved in low-level speech perception, including segmental perception, word segmentation, and incremental lexical activation (McQueen & Dilley, 2021). For instance, prosodic prominence can change ‘pen’ into ‘pan’ (Steffman, 2021), sentence rhythm drives how words are segmented from a syllable stream (Dilley et al., 2010), talking fast or slowly influences word recognition (Bosker et al., 2020; Maslowski et al., 2019), and lexical stress guides the activation of lexical candidates in the mental lexicon (Cutler & Donselaar, 2001).

The present study focuses on the production of lexical stress, which is acoustically associated with positive peaks in F0 and amplitude contours, and longer vowel durations (Rietveld & Van Heuven, 2009). Acoustic cues to lexical stress may distinguish minimal word pairs in free-stress languages, such as English, Dutch, and Spanish (e.g., noun *OBject* vs. verb *obJECT*). Moreover, lexical stress even influences word recognition for words that do not form such, arguably rare, minimal word pairs. For instance, Reinisch, Jesse, and McQueen (2010) used eye-tracking to assess Dutch listeners' processing of suprasegmental cues to lexical stress. When presented with four words on a screen, including the partially segmentally overlapping word pair *OCtopus* and *okTOber*, and spoken instructions to '*Click once more on the OCtopus*', Dutch participants already preferentially fixated *OCtopus* well before hearing the segmentally disambiguating /p/ in the third syllable. This finding, replicated in English (Jesse et al., 2017) and Italian (Sulpizio & McQueen, 2012), emphasizes the critical role of lexical stress in incremental lexical activation.

Nevertheless, how talkers produce the prosody of their L1 is surprisingly variable. For instance, pause distributions, speech rate, and F0 patterns vary as a function of talker, dialect, gender, and register (Clopper & Smiljanic, 2011; Quené, 2008; Xie et al., 2021). L2 prosody production and perception are likely even more variable, being susceptible to influences of a speaker’s L1 prosody (Cutler, 2005). Talking too slowly, pausing in the wrong places, or

deviating from native-like turn-taking behavior can affect how an L2 speaker is perceived (Bosker et al., 2013; van Os et al., 2020). In fact, using prosody in an atypical way or context, as L2 learners sometimes do, can jeopardize effective communication with L1 speakers (Van Maastricht et al., 2016a). Prior work has shown that this is often due to the transfer of linguistic features from their L1 to their L2 (Ellis, 1994, p. 28), with prosodic features being especially difficult to acquire. For example, L2 learners have been shown to transfer intonation patterns (Van Maastricht et al., 2016b), and rhythm (Mattys et al., 2007; Van Maastricht, et al., 2019) from their L1 to their L2, even at an advanced proficiency level. L2 speakers also transfer the suprasegmental cues that signal lexical stress in their L1 when producing stress in their L2 (Chakraborty & Goffman, 2011). Similarly, in perception, they weigh L2 suprasegmental cues in line with their L1 (Tremblay et al., 2018). Hence, beyond acquiring the correct phonological segmental sequences, L2 learners also need to become sensitive to the prosodic features of their L2 if they want to approximate L1 speakers' speech production and/or communicate effectively with L1 speakers.

The degree to which the L1 and L2 clash in terms of their prosodic regimes predicts the learnability of the L2 prosody (Connell et al., 2018). These findings exemplify that prosodic L2 competence is not only a cognitive skill of *knowing* where stress needs to be placed, it is as much change of vocal-articulatory sensorimotor habits related to vowel duration, fundamental frequency (F0), and amplitude contours. Breaking out of L1 prosody habits is complicated even more because said prosodic modulations need to be timed with producing novel phonemes that might not yet be stably produced (Krivokapic, 2020). The atypical production of lexical stress is one of the key markers for distinguishing L1 from L2 speakers (Jilka, 2000), even slowing down the word recognition process in L1 listeners (Braun et al., 2011).

Finally, it is important to realize that lexical stress is not only an acoustic property of speech; stress is a multimodal phenomenon too. Although suprasegmental cues such as intonation and intensity are arguably less visually salient than some segmental features (e.g., consonantal place of articulation), humans are keenly sensitive to visual prosody (Bosker & Peeters, 2021). For instance, humans perform above chance on stress discrimination when presented with muted videos of a talking face (Jesse & McQueen, 2014; Scarborough et al., 2009). Moreover, visual stress is not restricted to articulatory cues alone. Recently, Bosker and Peeters (2021) demonstrated that the temporal alignment of relatively simple beat gestures to

speech influences lexical stress perception. That is, the same Dutch disyllabic auditory word could be perceived differently depending on whether a hand gesture was produced on the first or second syllable (e.g., distinguishing Dutch *PLAto* vs. *plaTEAU*). The authors interpreted this 'manual McGurk effect' as resulting from life-long exposure to close gesture-speech synchrony in speech production.

Gesture-speech synchrony

Gesture and speech are closely synchronized in time during speech production. Generally, the gesture apex falls on the stressed syllable of multisyllabic words, typically temporally aligned to the F0 peak (Leonard & Cummins, 2011a). While pitch accented speech may also be associated with negative dips in the F0 contour, gestures do not seem to align during those moments quite as much as compared to positive F0 excursions (Im & Baumann, 2020). Maintaining this synchronization between gesture and prosody requires continuous bidirectional feedback. It has been found, for example, that when speech is slurred due to a delayed auditory feedback of speech, the gesture-speech temporal synchrony is maintained due to equal slowing down of gesture (Chu & Hagoort, 2014; McNeill, 1992; Pouw & Dixon, 2019). In fact, Pouw and Dixon (2019b) found that when speech was hampered by delayed feedback, the kinematics of co-speech gestures were more tightly aligned with the peaks in the F0 of co-gesture speech. In children 4-5 years old, it has also been found that acoustic modulation of contrastive focus was boosted by the use of head gestures as opposed to speech productions without gesture (Esteve-Gibert et al., 2021). These findings have been understood as a form of multimodal entrainment, where recruiting one system for (quasi)rhythmic behavior can stabilize the rhythm of the speech system that is either perturbed or under development. Thus, according to this view, gesture can help the speech system by offering a temporal anchor that can stabilize aspects of speech (Dohen & Roustan, 2017; Iverson & Thelen, 1999; Treffner & Peter, 2002; Vilà-Giménez et al., 2019).

Previous research has generally assumed a cognitive basis for the close temporal relationship between speech and gesture (McClave, 1998; Wagner et al., 2014). And indeed, for example in the case of beat gestures, it is not such a strange idea that a shared cognitive source determines the exact timing of prosody both in speech and gesture (Krauss et al., 2000; Ruiter, 2000). However, a recently proposed mechanism for gesture-speech coupling (Pouw, Harrison, et al., 2019) also suggests a direct biophysical coupling of upper limb movements with the

respiratory-vocal system (Aruin & Latash, 1995; Cordo & Nashner, 1982; Hodges & Richardson, 1997; Levin, 2006; Turvey & Fonseca, 2014) which can change subglottal pressures, which is an important parameter for intonational control (Finnegan et al., 2000). This physical gesture-speech link is supported by findings showing that intensity and secondarily F0 of speech reach positive peaks at moments when there is a (higher) physical impulse of rhythmic upper limb movements. This has now been observed in steady-state vocalizations (Pouw, Paxton, et al., 2019; Pouw et al., 2020; Pouw et al., 2019), mono-syllabic utterances (Pouw, et al., 2019), and fluent L1 speech (Pouw et al., 2020). In this way, gesture-speech synchrony is part of a wider phenomenon of respiratory-limb biomechanical interactions that have been observed across a range of mammals (see Pouw et al., 2021 for a review). To what extent this mechanism plays a role in L2 speech control is yet unknown.

While biomechanics may provide a basic explanation for why gesture and speech prosody couple in the way they do, as well as account for the relative ease with which humans accomplish this feat of coordination, the biomechanics is just one level of explanation (Pouw et al., 2021), which is not sufficient to explain gesture-speech synchrony.[1] Indeed, not gesturing does not seem to hamper speech prosody (Cravotta et al., 2019; Hoetjes et al., 2014), and speakers will have to negotiate their gesturing with the information structure, as well as the language-specific prosodic-syntactic conventions when speaking and gesturing at the same time.

Gesture and L2 acquisition

Although there are differences between individuals as well as languages in how much, when and how they gesture (Kita, 2009), most, if not all, languages employ co-speech manual or head gestures. In the current study, the focus lies on hand gestures, which are generally defined as symbolic hand movements that are produced during speaking and that are semantically and temporally closely related to speech (Gullberg et al., 2008; Kendon, 2004; McNeill, 1992). Especially relevant for the current study is the subcategory of beat gestures (or the beat-quality of gesture), which have been defined as (superimposed) biphasic (e.g., up-and-down) movements, made with one or two hands, without an immediately apparent semantic meaning

[1] For instance, stress in speech is primarily performed without upper limb movements through, for example, respiratory or laryngeal actions. Furthermore, stress may be performed via vowel-lengthening and can still be tightly coordinated with gestures (Krivokapić, 2014; Krivokapić et al., 2017). Thus, there are other constraints that determine how stress is realized, many of which do not have such a clear biomechanical basis.

(Leonard & Cummins, 2011b). Still, beat gestures serve a functional role in spoken language comprehension as multimodal prominence cues (Shattuck-Hufnagel & Ren, 2018), influencing pragmatic inferences (Krahmer & Swerts, 2007) and even low-level spoken word recognition (Bosker & Peeters, 2021). One area in which the close relationship between gesture and speech is particularly apparent is L1 development (see Gullberg et al., 2008 for an overview), with gesture paving the way for, or 'bootstrapping', L1 development (Iverson & Goldin-Meadow, 2005). With this potential facilitative role of gesture in language acquisition in mind, recent research has started to study whether beat gestures also play a role in L2 acquisition.

There is evidence suggesting that especially producing gestures (rather than only perceiving them) facilitates L2 acquisition (Li et al., 2020; Morett, 2018; Tellier, 2008). Yet most previous work on gesture and L2 acquisition focused on the use of representational (iconic and metaphoric) gestures on vocabulary acquisition (Huang et al., 2019) and relatively little work has been done in this context on beat gestures and their role in L2 phonology acquisition. Depending on the phoneme and gesture in question, perceiving gestures during L2 phoneme training can help in acquiring target-like L2 phoneme pronunciation (Hoetjes & Van Maastricht, 2020). Likewise, representational gestures can facilitate the L2 acquisition of vowel length production (Li et al., 2020), intonational contrasts (Kelly et al., 2017), and lexical tone perception (Morett & Chang, 2015). However, for most of these studies, although gestures facilitated L2 acquisition to some extent, the picture is generally less clear cut than for studies focusing on the use of iconic gesture in L2 vocabulary acquisition.

Although initial findings suggested that beat and metaphoric gestures may facilitate L2 lexical stress production, no convincing effect of beat or metaphoric gesture perception or production during lexical stress training on L2 production was found (Van der Heijden, 2021; Van der Heijden et al., 2021; Van Maastricht, Hoetjes, et al., 2019). Despite the temporal synchrony between beat gestures and prosodic emphasis, the question thus remains to what extent these gestures affect L2 lexical stress acquisition. And more importantly, studies have yet to explain why gestures appear to benefit L2 phonology learning in some contexts, but not in others. One often proposed factor is the cognitive demand associated with the task. Several studies report that when cognitive demands are high, for instance because the L2 (supra)segment is challenging for learners or because they are less proficient in general, the benefit of gestures appears to decrease (Hoetjes & Van Maastricht, 2020; Kelly et al., 2014).

Recent work on gesture-speech synchrony in bilinguals supports the assumption that beat gestures are not only closely temporally related to prosodic emphasis in the L1 but also in an L2. Hence, they may play a facilitative role in L2 lexical stress acquisition. For instance, the temporal relation between referential gestures and the co-occurring speech is the same for monolingual and bilingual speakers (Graziano et al., 2020). Presumably this finding also holds for beat gestures, which can be argued to be even more closely related to speech temporally (Wagner et al., 2014). Testing lexical stress production in Dutch learners of Spanish provides a unique testing ground for how beat gestures influence L2 acquisition.

Lexical stress in L1 and L2 Spanish

Dutch and Spanish are similar in that they are both free (lexical) stress languages: A given multisyllabic word must be characterized by one of the syllables receiving primary stress.[2] In both languages, lexical stress is phonologically contrastive: For example, in Dutch, *VOORkomen* means 'to occur', while *voorKOmen* means 'to prevent', while in Spanish *HAblo* means 'I speak', whereas *haBLÓ* means '(s)he spoke'. Hence, stress placement can determine the meaning of a word while its segments remain the same.

In Spanish, stress typically falls on one of the last three syllables of a word (Henriksen, 2013; Hualde, 2005). A distinction is made between oxytones, in which the last syllable is stressed (e.g., *numeRÓ*, '(s)he numbered'); paroxytones, in which the penultimate syllable is stressed (e.g., *nuMEro*, 'I number'); and proparoxytones, in which the antepenultimate syllable is stressed (e.g., *NÚmero*, '(the) number'). L2 learners can predict the stress placement in Spanish words following three rules (Aragonés Fernández & Palencia del Burgo, 2010; Kattán-Ibarra & Pountain, 2003):

1. If a word has a written stress mark (´), primary stress will be placed on the syllable containing the vowel with the written stress mark. For example: *soFÁ* ('sofa'), *LÁpiz* ('pencil') *teLÉfono* ('telephone').

[2] Aside from primary stress, it has been argued that both Spanish and Dutch also have secondary stress. For example, in the Dutch word **cho**coLA ('chocolate') the final syllable receives primary stress and is the most prominent but the first syllable receives secondary stress (in **bold**) and is more prominent than the second one (Kooij & Van Oostendorp, 2003). Similarly, in the Spanish word sen**ci**llaMENte ('simply') the penultimate syllable receives primary stress and the second syllable receives secondary stress (Hualde, 2005). In the current research, we only address the acquisition of primary stress.

For words without a written stress mark, it holds that:

2. If a word ends in a vowel or -n / -s, the penultimate syllable will be stressed. For example, *CLIma* (‘climate’), *amBIENte* (‘ambiance’), *arTIStas* (‘artists’), *eXAmen* (‘exam’).
3. If a word ends in a consonant that is not -n or -s, the ultimate syllable will be stressed. For example, *feLIZ* (‘happy’), *universiDAD* (‘university’).

The majority of Spanish words, irrespective of the presence of a written stress mark, are oxytonic or paroxytonic. Proparoxytonic words always require a written stress mark and are far less frequent (Hualde, 2005).

In Dutch, stress placement also generally occurs within the final three-syllable window (Kager, 1989, but cf. Köhnlein & Oostendorp, 2018; Oostendorp, 2012), as in, for example, *voorNAAM* (‘respectable’), *VOORnaam* (‘first name’), and *VOORnamen* (‘first names’). Lexical stress is further governed by phonological regularities, such as syllable weight and, in Dutch but not Spanish, vowel reduction (Trommelen et al., 1999). In sum, Spanish and Dutch are quite comparable; they both have free stress, both have language-specific stress placement rules and mostly suprasegmental lexical stress cues, which makes a comparison between the two languages both feasible and relevant.

Given these similarities between Dutch and Spanish concerning lexical stress production and perception, how do we explain that lexical stress production by L2 learners often is not target-like? That is, L2 speakers tend to transfer their L1 lexical stress patterns to the L2 (Archibald, 1992, 1993; Guion et al., 2003, 2004), especially when producing cognates. Since cognates largely share their segmental phonology with the L1, they appear to function as ‘false friends’ (Da Silveira et al., 2014; Edmunds, 2009). For instance, *kiLÓmetro* in Spanish and *KIlometer* (‘kilometer’) in Dutch are almost identical segmentally and would therefore seem easy to pronounce for Dutch learners of Spanish, however, the antepenultimate syllable is stressed in Spanish, whereas the first syllable is stressed in Dutch. This may lead to productions in Spanish that have a Dutch or mixed stress pattern, e.g., *KIlómetro* or even *KI**LÓ**metro*, in which the L2 learner may start to produce the word with a stress pattern that is typical of Dutch by realizing

primary stress on the first syllable, then sees the written stress mark and understands that the second syllable should receive primary stress and produces an even stronger emphasis on this syllable than they did on the first (marked in **bold capitals**).

L1 transfer of stress patterns in Dutch learners of Spanish is particularly pertinent because Spanish actually has a relatively large number of minimal stress pairs, emphasizing the importance of correct stress placement. That is, while Dutch only has a few minimal stress pairs that hardly share any semantic features (e.g., *CAnon*, 'musical canon' vs. *kaNON*, 'cannon'), Spanish minimal stress pairs are often common verb conjugations (e.g., *coMENto*, 'I comment' vs. *comenTÓ*, 'he commented'). Hence, in L2 Spanish, the semantic similarity between minimal stress pairs, in combination with the frequent absence of subject pronouns, can complicate communication when lexical stress is produced incorrectly as it becomes unclear which actor performed which action (Saalfeld, 2012).

Current study

Previous research has shown that gesture and prosodic aspects of speech naturally align in the L1. The current study investigates how L2 learners negotiate and potentially exploit this natural alignment of gesture with stress in L2 speech. Dutch learners of Spanish are an appropriate test population in this respect as the stress systems in the two languages are very similar, using the same suprasegmental prosodic cues, while sharing a good number of stress-matching and stress-mismatching cognates. As a result, they may not only provide a unique window into whether gesturing facilitates correct stress placement in the L2, but perhaps even more importantly how the two action systems (articulation and manual gesture) are brought in harmony in L2 prosody production. For instance, do prosody and gesture always go 'hand in mouth', or do speakers demonstrate correct L2 stress placement first in their speech and only later in their hands, or *vice versa*?

In this proposed within-subject study, Dutch learners of Spanish are video-recorded while pronouncing Spanish lexical items that have a cognate counterpart in Dutch, either with or without producing a beat-like gesture. Acoustic analysis of the audio recordings and motion-tracking in the video recordings will be combined to assess stress placement (which syllable carries stress), acoustic stress peaks (where in the syllable does the stress fall on a millisecond timescale), and gesture-speech synchrony (temporal distance in ms. between acoustic stress peak

and maximum hand extension). Critically, some Spanish words share the lexical stress pattern with the Dutch cognate (stress-match; e.g., Spanish *MANgo* vs. Dutch *MANgo*), while others have a different lexical stress pattern (stress-mismatch; e.g., Spanish *eRROR* vs. Dutch *Error*). To add a further level of difficulty, some Spanish items have orthographic marking of stress by means of a written stress mark (e.g., Spanish *soFÁ* vs. Dutch *SOfa)* while others don't (e.g., Spanish *profeSOR* vs. Dutch *proFESsor*). Manipulating L1-L2 stress (mis)match and the presence of a written stress mark in this way allows us to create conditions that vary in how cognitively demanding they are for L2 learners, as well as shed some light on the factors that may further explain the relationship between gesture and speech. Hence, this study aims to answer two main research questions:

1. How does gesturing influence the acoustic production of stress by L2 learners?
2. How is gesture-prosody coupling in L2 influenced by competition from the speaker's L1?

The first research question involves comparisons of acoustic stress production when L2 speakers gesture or not (i.e., disregarding gestural timing). Specifically, (1A) we will assess whether gesturing makes learners more accurate in L2 stress placement (i.e., whether there is a greater likelihood of correctly placing stress on *-sor* in *profeSOR* when L2 speakers gesture vs. do not gesture). Such a beneficial effect may be driven by the fact that another effector system (i.e., gesture) with its own timing regime may raise awareness of where the L2 stress should be placed, thereby increasing performance as an attentional anchor. Additionally, it is possible that by recruiting a manual system that has its own flexibilities in timing, it allows for the more habituated speech system to cognitively anchor to gesture so as to maintain correct L2 timing (Esteve-Gibert et al., 2021; Iverson & Thelen, 1999; Pouw & Dixon, 2019). The general idea that a manual motor system can cognitively support timing processes aligns with research on basic sensorimotor timing where it was found that overt tapping movement improved temporal auditory predictions (Morillon & Baillet, 2017). Gesture can be a physical anchor for the respiratory-vocal system as well, such that physical impulses are generated onto the respiratory-vocal system during beats (Pouw, Harrison, et al., 2020), which supports the production of a correctly timed stressed syllable which may help compete against the habit of incorrect stress placement. Thus, gestures can function as attentional, cognitive, and biomechanical anchors for

L2 prosody production, none of which are mutually exclusive, and they may even function in concert. Furthermore, (1B) we will assess whether gesturing boosts the acoustic markers of stress. That is, irrespective of stress placement accuracy, does gesturing on a given syllable make that syllable louder, and higher pitched, as predicted by biomechanical accounts?

The second research question involves assessment of gesture-prosody coupling using time-series of the motion-tracking data. First, we will test (2A) whether learners demonstrate closer or 'sloppier' gesture-prosody synchrony in stress-matching vs. stress-mismatching cognates. This would indicate an effect of L1 competition that could either enhance gesture-prosody coupling under cognitively more challenging conditions (much like how delayed auditory feedback enhances gesture-speech coupling; Pouw & Dixon, 2019) or hinder gesture-prosody coupling through L1 transfer. We will also assess the (possibly moderating) role of stress placement accuracy (see below). Second, we will test whether the two systems (vocalization and gesture) demonstrate L1 influences in their timing on a millisecond timescale. Specifically, (2B) focusing on stress-mismatching cognates, do 'correct L2 productions' still demonstrate evidence of L1 temporal attraction in gestural timing in the direction of the L1 stressed syllable (i.e., correctly saying *profeSOR* but gesture apex, as the L1 cognate is *proFESsor*)? And the reverse: (2C) do 'incorrect L1-like productions' still demonstrate evidence of L2 temporal attraction in the form of (visual) gestural timing in the direction of the L2 syllable (i.e., incorrectly saying *proFESor* but with a relatively late gesture apex, as the correct L2 pronunciation is *profeSOR*). Answering 2B and 2C will demonstrate whether gesture and prosody always go 'hand in mouth' or whether the gesture system is still susceptible to attract to competing targets.

Note that in our analyses the orthographic marking of stress will be incorporated as a potential modulator of the above tested gesture-prosody synchrony. On the one hand, this orthographic marking may intuitively be taken to facilitate stress placement accuracy, as it marks the syllable that carries stress. As a result, gesture and speech may be less closely coupled than in more challenging conditions. However, recent research (Gutiérrez-Palma et al., 2020) has shown that words with orthographic stress marking are *more taxing* to process than words without. This would predict that gesture and speech may be more closely coupled in words with orthographic stress marking than without.

Method

Participant and design current pilot study

We have pilot data (N = 2) from two of the authors who performed the entire experiment. One of them (male) had low proficiency in L2 Spanish, while the other (female) had high proficiency, teaching undergraduate courses at university level in Spanish. Note that this data is only used for the demonstration of our planned statistical modeling, as well as to provide an initial power analysis designed for power estimations based on pilot data.

Participants and design confirmatory study

Our recruitment population will consist of L1 Dutch first-year students of Spanish at Radboud University. We aim to recruit at least 30 participants, but we will aim to recruit as much of the convenience population as possible with a lower limit of 20 participants. This is a fully within-subject study with 3 factors with each 2 levels: Gesture condition (no gesture vs. gesture), L1-L2 stress mismatch (same vs. different), and stress mark presence (absent vs. present). Please see our power analysis reported below for further sample size justification.

Materials and equipment

Camera. We use a Canon XF105 camera for audiovisual recording. We record at a high frame rate of 50 frames per seconds to maximize the temporal resolution for motion tracking.

Audio Recording. We use a direct cardioid beam Sennheiser microphone (model K6/ME 64) sampling at 41.1 kHz. The microphone is connected to the camera, thereby synchronizing high quality audio streams with the video (and motion tracking) data.

Experiment Presentation. We use `R` (R Development Core Team, 2012) for presenting the stimuli to the participants and for item randomization (see the script on OSF). Although `R` is commonly used for statistical analysis, it can also be purposed for the presentation of stimuli via its base graphic presentation functions.

Stimuli. The stimuli consist of 96 Dutch/Spanish cognates that are equally divided across four conditions:

1) L1-L2 stress match without written stress mark
   (e.g., Spanish: *inSECtos*, Dutch: *inSECten*)

2) L1-L2 stress match with written stress mark
   (e.g., Spanish: *BRÓcoli*, Dutch: *BROccoli*)
3) L1-L2 stress mismatch without written stress mark
   (e.g., Spanish: *farMAcia*, Dutch: *farmaCIE*)
4) L1-L2 stress mismatch with written stress mark
   (e.g., Spanish: *FÓRmula*, Dutch: *forMUle*)

Due to the fact that all stimuli are cognates to maximize the probability of L1 to L2 transfer, they are not perfectly balanced for their number of syllables or the extent and direction of the mismatch between the stress position in the L1 and L2; see Appendix 1 for the full stimuli list and their characteristics. The order of the stimuli is randomly shuffled for each participant. Then, a gesture condition (no gesture vs. gesture) is randomly assigned to a trial in such a way that conditions occur in blocks of 6 consecutive trials. This way participants do not need to change between gesture vs. no gesture mode on each and every trial. When all stimuli are shown once, the trials are repeated in the same order but with the opposite gesture condition assigned to the second round of trials.

Procedure

Participants are welcomed into the lab, a largely empty room, with a camera positioned such that it records the participants' upper body, and a computer screen visible to the participant which will show the orthographic stimuli one by one. Participants are given written instructions in Dutch, the opportunity to ask questions, and sign an informed consent form. The first instruction is about how to perform the stereotypical gesture, which will be performed in half of the trials. Participants are instructed to produce a single gesture while reading out loud a Spanish word. They are shown a muted video example of such a gesture, without any audio or facial articulatory information by masking the face, and are then asked to practice the articulation of the gesture with a stress-matching practice word (*virus*). In this way the experimenter can watch the performance of these practice gestures to see whether any of the instructions are misunderstood. The experimenter will only intervene if the instructions are misunderstood, and under no circumstances will the participant be instructed to align their gesture in a particular way to acoustically produced stress. The example video and accompanying instructions in Dutch can

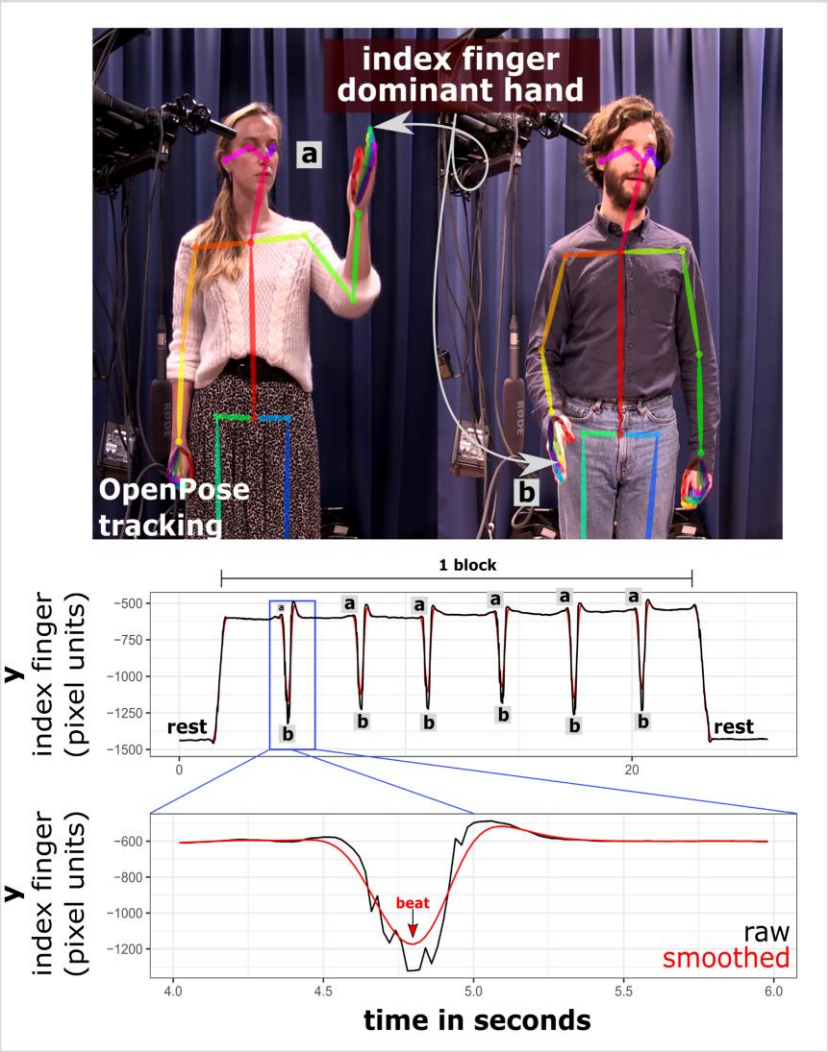
be seen following this link. Hence, participants are prompted to move one hand in an up (flexion) and down (extension) motion of the lower arm around the elbow joint, along the sagittal axis (i.e., in front of their body, also see Figure online). Further instructions explain that the study consists of reading out loud Spanish words while standing, which for half of the trials involves keeping the arms and hands still and relaxed alongside the body. Participants are instructed that a word presented in a blue font requires a gesture to be produced during the pronunciation of that word. If the color of the word is red, then participants should keep their hands still. Gesture vs. no gesture conditions were always presented in blocks of 6 trials, with 3-second pauses in between blocks; this way participants experience minimal cognitive load due to having to switching procedures, and when they switch they have some time to prepare. After participation, participants are asked to fill out the Spanish Lextale test (Izura et al., 2014) as a measure of their L2 Spanish proficiency, and answer demographic questions. The entire experiment is estimated to take about 30 minutes.

Post-Processing and Key Measures

Motion tracking. Motion tracking is performed in the post-processing phase using video-based motion tracking software OpenPose (Cao et al., 2017). OpenPose utilizes a deep neural net that is trained to recognize human poses from video data. Deep-learning based motion tracking has been shown to be suitable for gesture timing analysis as compared to high performance wired motion tracking devices (Pouw, Trujillo, et al., 2019). Indeed, video-based motion tracking methods are increasingly used by researchers as they allow quantifying bodily movements in a non-invasive and reproducible way (Alviar et al., 2020; Ripperda et al., 2020; Trettenbrein & Zaccarella, 2021).

Using OpenPose, we extract the human pose information from the video frames at a temporal resolution of 50 Hz and spatial resolution of 1080 by 1920 pixels. From this raw pose-tracking data, we construct a time series with information about our key kinematic variable of interest: the change in vertical position of the dominant index finger. We apply a 3rd order Kolmogorov-Zurbenko filter (span = 5) to smooth out high-frequency jitters; this filter is comparable to a moving average filter without phase distortion. We use a custom-made script to construct time series from the raw OpenPose data (for a tutorial and scripts see Pouw & Trujillo, 2019). See Figure 1 for an overview of this motion tracking procedure.

Figure 1. Example OpenPose motion tracking



Note Figure 1. The upper panel shows OpenPose tracking for 1 frame for pilot participant 1 (left) and 2 (right). For each frame, the human pose data was computed and then a time series was constructed collecting information of the vertical position of the dominant index finger through time (50 Hz sampling rate). The middle panel shows such a time series for a single block, where the position of the index finger starts from rest, is raised to a start position (a) followed by 6 trials of gesture-speech utterances where a beat (b) is timed with a speech unit. The lower panel shows a time series that takes up about a single trial in length. The black line shows the original estimate of OpenPose, and in red is the smoothed version of this motion trace. The smoothed time series are the motion variables we submit for further analysis. For example, if we determine the moment of a gestural beat (i.e., maximum extension), we determine the negative peak of the through of the smoothed red line.

Syllable boundary detection. We use R-package `speakR` (Coretta, 2021) to execute a Praat script implementing forced-alignment in Spanish using EasyAlign (Goldman, 2011). As

input, it takes the speech recordings and accompanying orthographic transcripts, producing word- and syllable-level annotations in TextGrid format as output. The number of syllables that EasyAlign detects will be checked manually to avoid missegmentation due to L2 atypicalities that EasyAlign is unfamiliar with. For instance, L2 speakers sometimes fail to produce a diphthong as one syllable (*far-ma-CI-a* instead of *far-MA-cia*).

F0. We apply a Schaefer-Vincent periodicity detection algorithm to extract F0 traces using R-package `wrassp` (Winkelmann et al., 2018). To avoid irregularities due to period doubling or noise, we further apply a 40 Hz Hanning window to smooth F0 traces. Female preset range is 80.0 - 640.0 Hz and male range is 50.0 - 400.0 Hz. If sex at birth is not provided by the participant we will use the default range 0.0 - 600.0 Hz. We sample F0 at 200 Hz.

Amplitude envelope. As a measure of intensity, we compute a smoothed amplitude envelope by applying a Hilbert transform to the waveform, taking the complex modulus, and then smoothing with a 10 Hz Hanning window. This follows a procedure proposed by He and Dellwo (2017). We downsample the amplitude envelope to 200 Hz.

Duration. Having identified the syllable boundaries using EasyAlign, we determine for each syllable its vocal duration (phonation duration in ms. as detected by the F0 detection algorithm).

Acoustically stressed syllable identification. Once we have temporal estimates for syllable boundaries within each trial, we must determine which of the syllables is acoustically most prominent and can hence be classified as the stressed syllable (link to script). We identify three acoustic markers of stress: F0, amplitude (envelope), and duration. Considering variability in suprasegmental cue weighting in Spanish vs. Dutch, and presumably substantial between-talker variability in the use of these cues in non-native speakers, we compute a weighted lexical stress score *S*:

$$S_i = W_F(F_i^z) + W_I(I_i^z) + W_D(D_i^z)$$

Where the stress score S_i for syllable *i* in an utterance is determined by the weights ($W = \{W_F, W_I, W_D\}$) multiplied by the z-normalized F_i^z (peak F0), I_i^z (peak amplitude), D_i^z (Duration) measures and then summed. Each acoustic measure extracted for each syllable is z-normalized for the whole utterance, so that they are comparable (i.e., can be added). We set all weights (W_D,

W_F , W_I) to 0.33 in order to equally weigh their contributions to the stress score. The syllable with the highest stress score relative to all other syllables in the utterance is then selected as the acoustically stressed syllable. Note that this procedure allows flexibility for exploratory analysis, applying an asymmetric weighting set *W*'; for instance, weighing duration with 0.50 while amplitude and F0 with 0.25. This way we can determine whether our findings are reproducible across different phonetic operationalizations of stress.

Acoustic stress peak. Once we have nominated the acoustically stressed syllable, we also derive a time point estimate that we can relate, for example, to a point in time of the gesture movement. This stress time point was determined as the (local) maximum of the amplitude envelope for the interval of the nominated stressed syllable. Peaks in the amplitude envelope are used as they seem to be most directly tied to the gesture's physical impulse (as compared to F0; e.g., Pouw, Esteve-Gibert et al., 2020). Note that this maximum of the amplitude envelope is not necessarily the global maximum (i.e., the syllable with the highest amplitude within the word), as we determine the stressed syllable based on three acoustic markers, only one of which was the amplitude. Figure 1 provides an example.

L2 Target peak. The L2 target peak is a time point in the syllable that *should* (but not necessarily does) carry stress in correct L2 pronunciation. We compute the *L2 target peak* dynamically, that is, based on the actual pronunciation of the participant for that trial. Namely, we select the syllable that should be stressed in L2, and then we determine the peak in amplitude for that syllable, in a similar way as the stress time point discussed previously. This L2 target peak can then be related to other time points we estimated (e.g., acoustic stress peak).

L1 stress competitor syllable. In the case of a stress mismatch between the L1 and L2, there is another syllable that is a potentially competing stress target. This L1 stress competitor syllable is also dynamically computed, by selecting the syllable that would be stressed in the L1.

Directional stress timing. Our main measure of L2 stress placement performance is the degree of temporal offset in milliseconds between the L2 target peak and the acoustic stress peak. Hence, for some stressed syllable, we have a stress peak, and we have a peak for the L2 target syllable. Then we simply compute the difference in timing between the stress and the L2 peak. Critically, however, we adjust the sign of timing offset in the case of L1/L2 stress (mis)match so that our stress timing offset becomes *directional*, that is, to make the timing offset interpretable such that we know whether the mistiming was such that the stress peak became

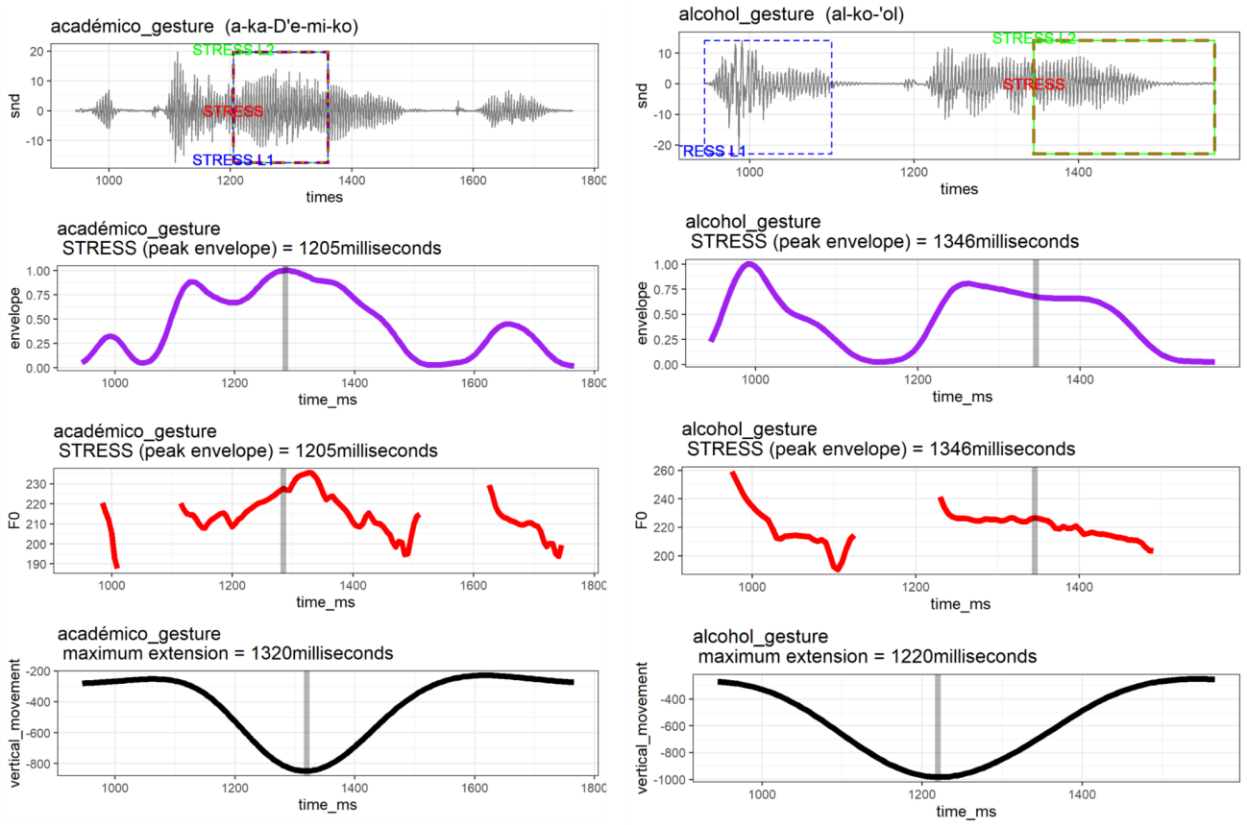
attracted or not towards L1. Note, that when the nominated stress placement is the same as the L2 stress target (i.e., when lexical stress is produced correctly), then the stress timing offset equals zero.

Gesture apex. For the gesture condition, we determine the apex of a beat gesture as the point of maximum extension of the index finger of the dominant hand (see Figure 2). This point in time will be used so as to compute gesture-speech (a)synchrony.

Gesture-speech (a)synchrony. The asynchrony between gesture and speech (*D*) is the offset between gesture beat time point minus the stress time point. Note that the sign of *D* indicates that gesture apex followed (positive sign) or led (negative sign) the acoustic stress time point; that is, positive values indicate that gesture followed prosody and negative values indicate that gesture preceded prosody.

Directional gesture-speech (a)synchrony. To make the gesture-speech asynchrony offsets interpretable relative to L1/L2 stress (mis)match, we transform the asynchrony in a similar way as we did for the directional stress timing. Namely, when the L1 stress competitor occurs after L2, then we reverse the sign of D to give D'. This means for example, that when there is an offset such that the stress peak occurred -100 ms. earlier in time than the gesture apex and L1 competitor target also occurred earlier in time than the L2 target, we would transform D into D' to indicate that the gesture was misaligned with speech *in the direction of the L1 distractor* (i.e., D' = D *-1 = 100 ms.).

Figure 2. Example trials with and without L1/L2 stress match



Note. Our script produces plots (examples) with information about acoustic stress placement as determined by our weighted stress score computed for each syllable (STRESS in red), L2 stress target (target STRESS L2 in green), and stress according to L1 competitor (STRESS L1 in blue). The stress L2 target and L1 competitor may either overlap (left plot) or not (right plot), depending on our stress (mis)match condition. In purple, the smoothed amplitude envelope is shown, which traces and smooths the maxima of the amplitude of the waveform shown in the top panel. F0 is given in red. In black, the vertical displacement of the index finger of the dominant hand is given. The beat time point of the gesture is given with a vertical line, indicating the maximum extension (here detected at 1320 and 1220 milliseconds in the trial, respectively). These plots will be generated as part of the supplemental data for all trials in the final experiment. Note that in the no gesture condition, the movement time series is not informative and variables that derive from these time series are absent.

Confirmatory Analyses

All analyses are tailored to provide statistically informed conclusions about the research questions formulated under Section 2, titled ‘Current Study’. Below, we report statistical models as confirmatory analysis of those research questions. We report graphical and statistical results of the pilot data to exemplify the type of inferences we can draw and how we will report our

models, but the outcomes should not be taken as representative of our predictions or expectations. Our results section derives from a fully reproducible R Markdown notebook.

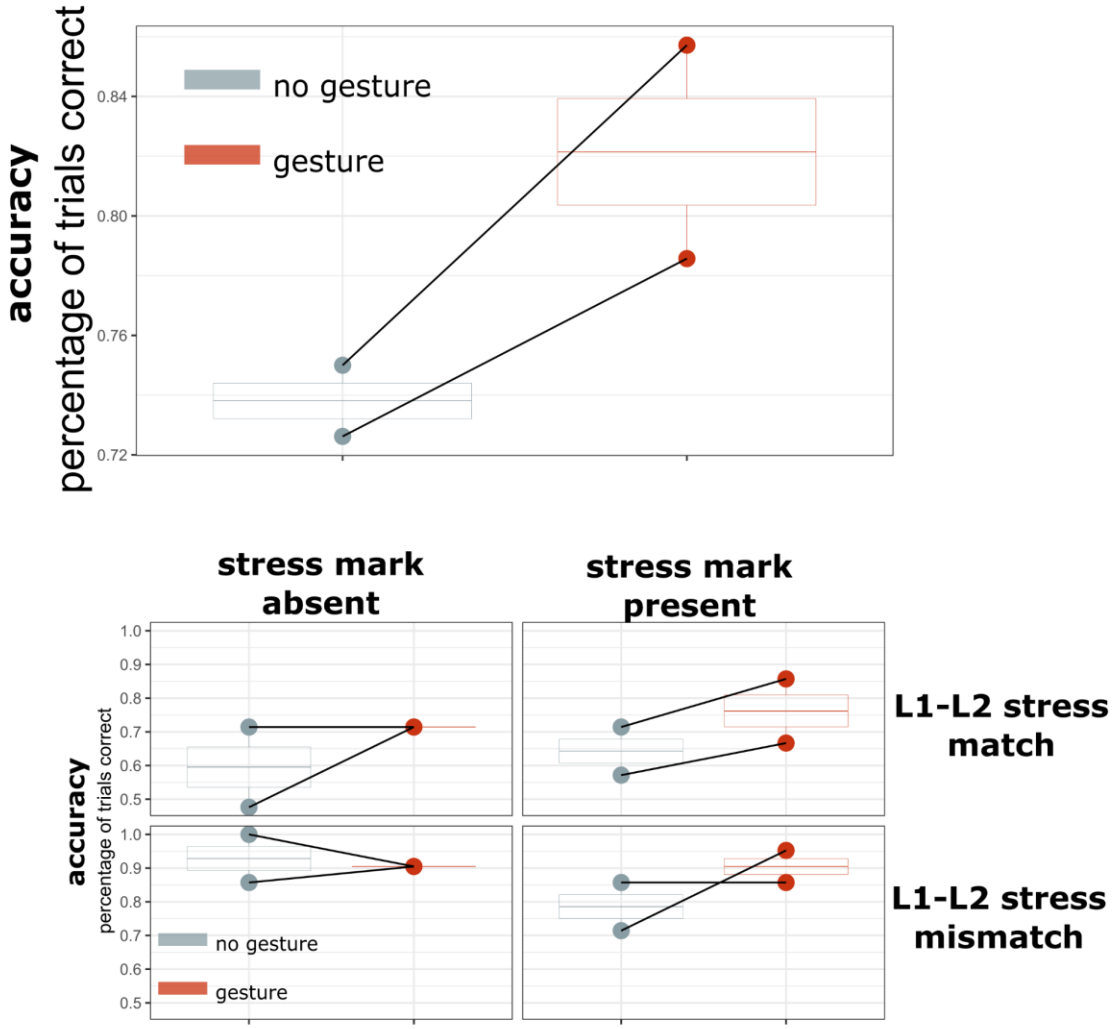
For all analyses, we will use mixed linear regressions with maximum likelihood estimation using R-package `nlme` (Pinheiro et al., 2019), or `lme4` (Bates et al., 2021) for binomial data. Our models will always have participant and trial ID as random variables. We will always try to fit random slopes, next to random intercepts. With the current pilot data however, adding random slopes resulted in non-converging models. Thus, for all models reported we have participant and trial ID as random intercepts. We further report a Cohen's *D* for our model predictors using R-package `EMAtools` (Kleiman, 2017). For interaction effects we will follow up with a post-hoc contrast analysis using R-package `lsmeans` (Lenth & Lenth, 2017) and apply a Bonferroni correction for such multiple comparison tests.

Confirmatory analysis 1A: Effect of gesture on stress placement accuracy

Here, we assess whether the accuracy in stress placement is higher in the Gesture versus No Gesture condition, with a primary interest in stress-mismatching cognates. We use a binomial mixed linear regression, with stress correct (=1) or incorrect (=0) as the dependent variable. We first construct a base model predicting the average performance, with participant and trial ID as random intercepts. In our first model, we assessed whether adding gesture condition as a variable improved predictions relative to the base model, which it did, χ^2 (1) = 4.643, p = .0312. In our pilot data, trials in the gesture condition were more likely to have a correct stress as compared to trials in the no gesture condition, b = -0.671, z = 2.117, p = .034, Cohen's D = 0.276.

A more complex model will be assessed as well, where, next to gesture condition, we add L1-L2 stress match, written stress mark presence, and their interactions as predictors. This more complex model increased accuracy predictions relative to the model with only gesture condition as independent variable, change in χ^2 (6) = 16.133, p = .013. For brevity, we will not report on the pilot results here, but we do report the model outputs and post-hoc tests in the R Markdown notebook (code chunk 1). Note that only if one of the interactions is statistically reliable we will report the post-hoc comparison with R-package `lsmeans` with a bonferroni correction in the confirmatory study.

Figure 3. Effects of gesture on stress timing



Note. Percentage correctly stressed trials are reported for gesture condition (upper panel), as well as stress mark presence and L1-L2 stress match (lower panel). Points indicate percentage for a single participant, and lines indicate relative change within participants. Percentages are expressed for each (sub)condition, such that the number of correct trials are expressed relative to the total number of trials for that (sub)condition. In the pilot data, there were more correctly stressed trials for the gesture condition for both participants. Note that for the left lower corner, in the gesture condition, both participant 1 and 2 performed equally well, therefore yielding overlapping scores.

Confirmatory analysis 1B: Effect of gesture on acoustic markers of stress

Does gesturing enhance the acoustic realization of stress? We perform a mixed linear regression with normalized acoustic output as dependent variable, and acoustic marker (peak F0,

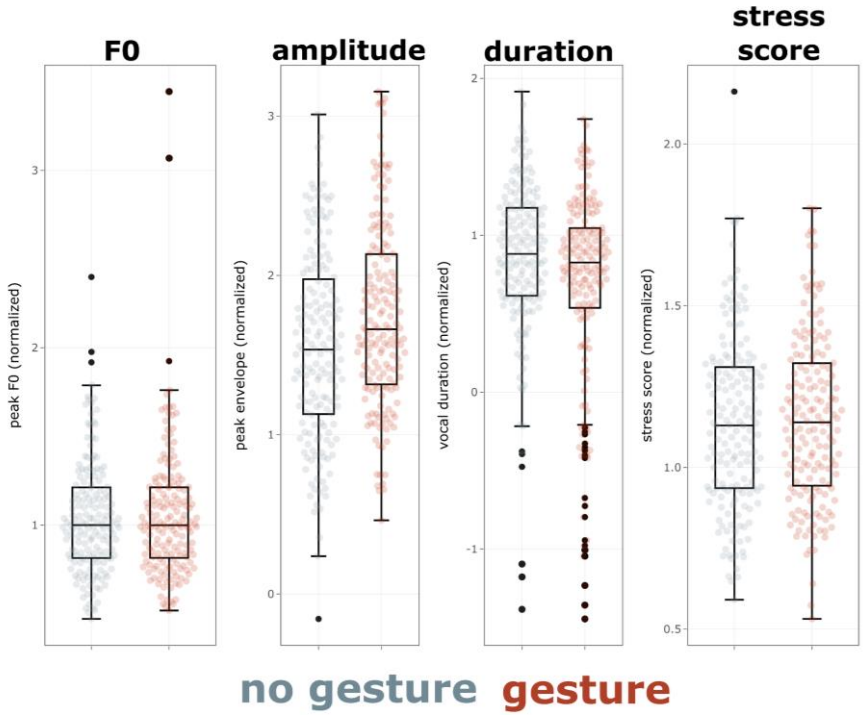
peak envelope, and duration) x gesture condition as independent variable. We again test this model against a base model predicting the overall mean.

The model with acoustic markers x condition was a more reliable model than the base model predicting the overall mean of the acoustic output, χ^2 (5) = 426.593, $p < .001$. Table 3 provides an overview of the model predictors. Figure 4 provides a graphical overview of the results from the pilot data.

Table 3. Model predictors condition x acoustic marker

	b	*SE*	*t* (835)	*p*	*Cohen's D*
Intercept	1.55	0.04	39.00	< .001	
F0 vs. Amplitude	-0.51	0.06	-9.15	< .001	-0.63
Duration vs. Amplitude	-0.70	0.06	-12.69	< .001	-0.88
Gesture vs. no gesture	0.19	0.06	3.34	< .001	0.23
F0 x gesture	-0.17	0.08	-2.18	.03	-0.15
Duration x gesture	-0.35	0.08	-4.41	< .001	-0.31

Figure 4. Gesture vs. no gesture and acoustic markers of stress



Note. For each acoustic marker of the stressed syllable (peak F0, peak amplitude envelope, and duration; all z-scaled) we compare gesture vs. no gesture condition. The right-most panel shows the stress score, which is the weighted sum of the acoustic markers.

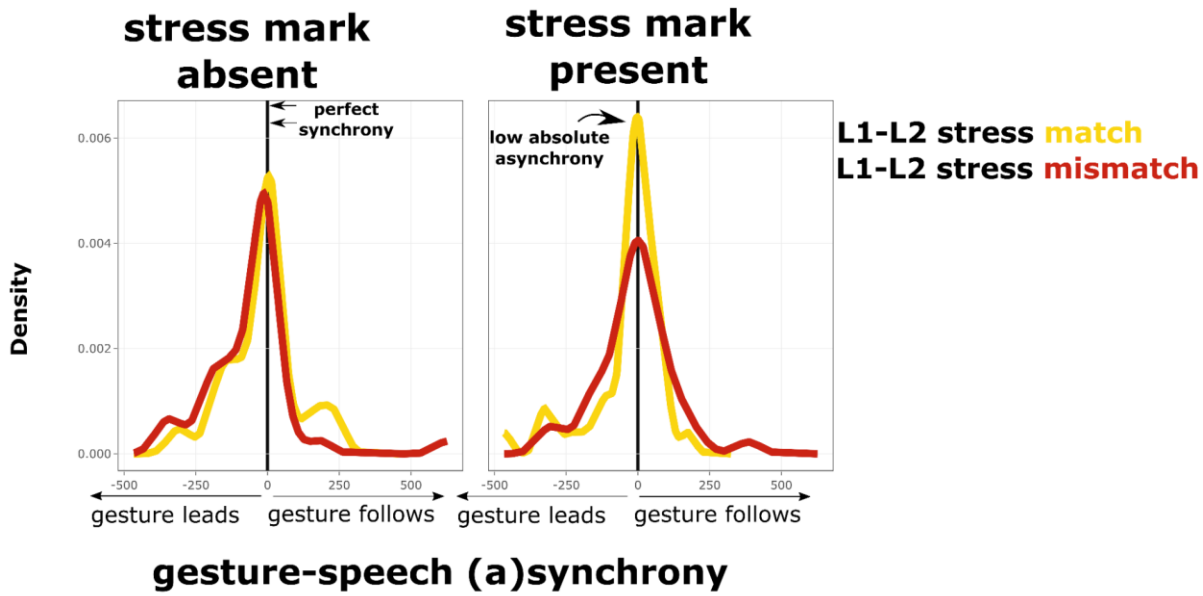
We will perform a post-hoc analysis of disentangling interaction effects when they are found, in which we assess which acoustic marker was most affected by gesturing (see R Markdown notebook, code chunk 3 for full details on the post-hoc model code). Further, if there is an effect of gesture on acoustic stress realization, we will assess in a more complex model whether stress mark presence and L1-L2 stress mismatch interact with the gesture condition effect on stress realization.

Confirmatory analysis 2A: Comparing gesture-speech (a)synchrony in stress-matching vs. stress-mismatching cognates

Here, we test whether the synchrony between gesture and speech is affected by L1-L2 stress (mis)match and written stress mark presence, which would signal that gesture does not always synchronize with speech to a similar extent, but that gesture-speech coordination is reduced or in fact enhanced due to cognitively challenging conditions, such as having to reach L2 targets without orthographic cues or with an L1 stress competitor.

Using a similar linear mixed modeling approach as in the previous analyses, we compared a base model with models with L1-L2 stress (mis)match and written stress mark presence (and their possible interactions) as predictors for the absolutized gesture-speech asynchrony (see figure 5 for a graphical overview). For our pilot data, including L1-L2 stress (mis)match and written stress mark presence as predictors in an alternative model was not more reliable than the base model predicting the overall mean of the absolutized gesture-speech (a)synchrony, χ^2 (2) = 1.245, p = 0.537, and adding interactions between L1-L2 stress (mis)match and written stress mark presence also did not further improve predictions of gesture-speech asynchrony, χ^2 (3) = 1.290, p = 0.732. Table 4 provides an overview of the model predictors for the model without interactions.

Figure 5. Gesture-speech (a)synchrony depending on L1-L2 stress (mis)match and written stress mark presence



Note. Smoothed density distributions are shown for gesture-speech (a)synchrony, split for trials where the word had a written stress mark (right panel) or not (left panel), and separated by L1/L2 stress match (yellow colored) or mismatch (red colored). When gesture apex and stress peak were perfectly synchronous, we yield a value of 0 milliseconds. If a gesture followed or led the stress peak, we yield positive versus negative values, respectively. Note that a more peaked distribution indicates higher gesture-speech synchrony, while a more spread out distribution entails lower gesture-speech synchrony.

Table 4. Gesture-speech (a)synchrony fitted predictions by L1-L2 stress (mis)match and written stress mark presence

	b	*SE*	*t* (164)	*p*	*Cohen's D*
Intercept	88.90	13.95	6.37	<.001	
L1-L2 stress mismatch yes vs. no	14.71	16.11	0.91	.362	0.14
Written stress mark presence yes. vs. no	-10.10	16.11	-0.63	.533	-0.10

Confirmatory analysis 2B (and 2C): L1 temporal attraction in gesture

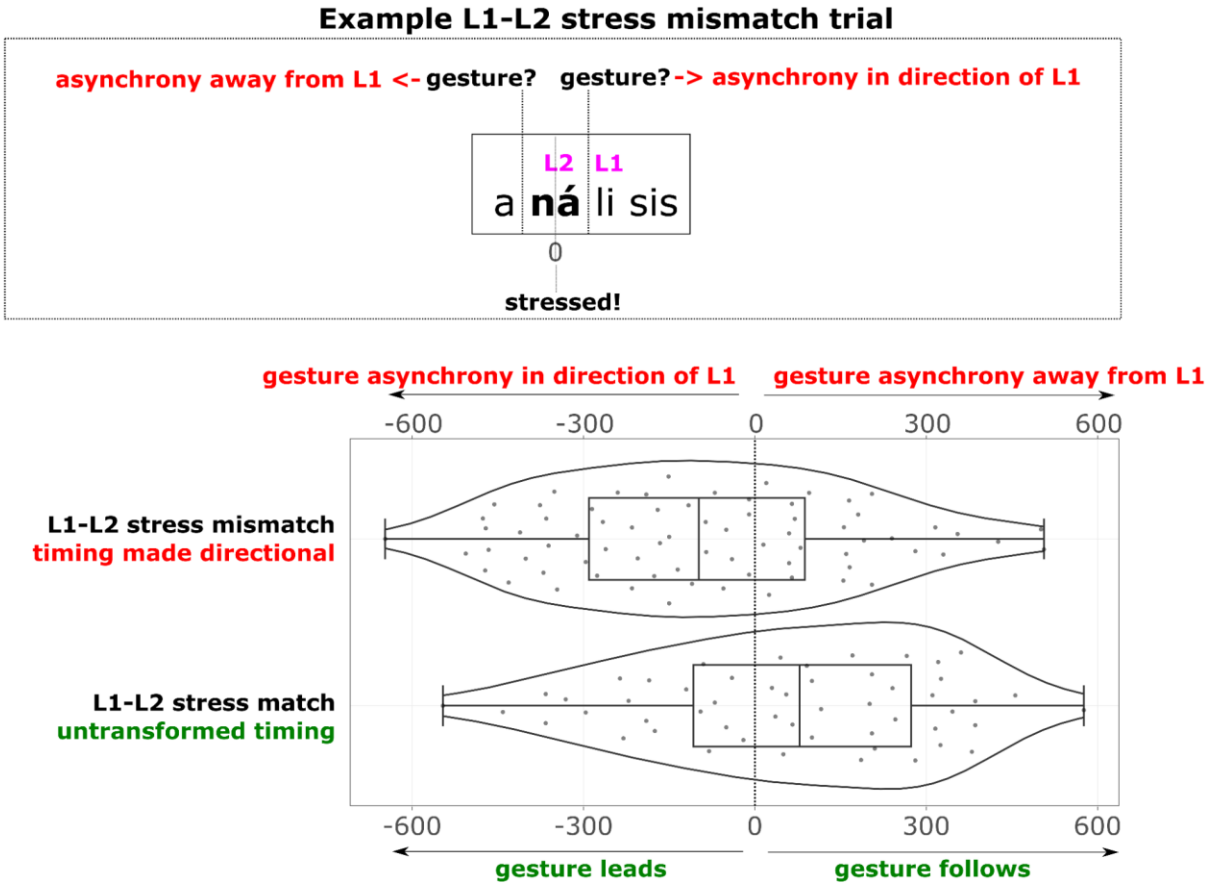
Here we ask *how* gesture, prosody, or perhaps even both systems are influenced by L1 temporal attraction (or L2 attraction). We will assess this here for 2B, by looking at the gesture-speech asynchrony when the acoustic stress peak is correctly placed on the L2 target. Figure 6 provides an example of our pilot data results where we report directional gesture-speech (a)synchrony when acoustic stress is correctly placed on the L2 target. We will assess whether the gesture is attracted to be asynchronous with speech in the direction of the L1 stress competitor as compared to the no stress difference condition (which acts as a baseline). We compare this directional (a)synchrony when there is a L1-L2 stress mismatch, versus when there is no L1-L2 stress mismatch. If there is an attraction of gesture towards L1, we predict a more negative directional gesture-speech (a)synchrony in stress-mismatching vs. stress-matching words (see Figure 6).

The current analysis is conditional upon having at least 33% of the total responses for a particular response type (at least 33% of the trials of L2 correct, L2 incorrect-L1 match, and L1-L2 incorrect). In the current pilot data, for example, we have primarily L2 correct responses[3] (~78%), so we could only analyze this response type. For each conditional analysis we perform a linear mixed regression analysis with participant and trial ID as random intercepts, L1-L2 stress (mis)match as IV, and directional gesture-speech (a)synchrony as DV.

Assessing the differences in gesture-speech asynchrony as shown in Figure 6, including L1-L2 stress (mis)match as predictors in an alternative model was not more reliable than the base model predicting the overall mean of the (directional) gesture-speech (a)synchrony, χ^2 (2) = 0.014, p = .0974.

[3] This is not necessarily to be expected for the data that is to be collected, as in the pilot data, half of the trials were produced by an experienced L2 speaker of Spanish, which the target participant group will not be.

Figure 6. Directional gesture-speech (a)synchrony for trials where the L2 target was correctly stressed



Note. The directional gesture-speech asynchrony as observed in the pilot study is shown for trials where the L2 target syllable was correctly acoustically stressed. The x-axis indicates a perfect synchrony between gesture and speech at 0 milliseconds, and the gesture timing in the direction of L1 competitor in negative values, and gesture timing in the opposite direction of L1 in positive values. In the lower row, the L1 and L2 stress targets were on the same syllable, thus the direction of the gesture-asynchrony is informative about general gesture-speech timing tendencies, indicating that gesture leads [negative values] or follows [positive values] acoustically stressed peak.

Exploratory Hypothesis

We retain some flexibility to further analyze variables of interest in an exploratory fashion. All these exploratory analyses will be labeled as such to distinguish them from the confirmatory analysis reported above. Possible analyses will be shortly discussed below.

L2 proficiency. We collect information about the L2 proficiency of our participants using the Spanish Lextale test (Izura, Cuetos & Brysbaert, 2014). This allows us to assess

possible differences in L2 stress placement and the role of gesture therein as a factor of general L2 proficiency.

Weightings. If we deem it necessary for the interpretability of the results we can change the weighting set W, so as to see whether the choice of the importance of acoustic markers matters for our measurements and results.

Continuous kinematic analysis. We can further probe whether the kinematic gesture trajectory is biased depending on stress difference competition; for example, by using generalized additive modeling of the trajectories.

Alpha restriction

To ensure control of false positive error due to multiple hypothesis testing we will deem any result statistically reliable if, and only if, *p*-values are lower than the restricted alpha = 0.05/3 = .0166.

Power analysis

To provide some indication of the amount of data needed to get meaningful results, we performed a power analysis for the first confirmatory research question based on the pilot data. Note that the pilot data is not an ideal dataset to base our power calculations on, given that two of the current authors were the pilot participants and thus diverge from the eventual sample population. However, we can use the pilot data as an initial basis for understanding how many participants we need given a certain effect size (independently of whether we believe the effect to be realistic for the sample data or not). We first assessed the power of a model with one main effect (gesture condition), which in the pilot data had a small effect size of $D \sim .2$, on stress accuracy. We further assume a restricted alpha of .05/3, so as to determine how many subjects we need to detect the main effect at a power of 80%. We used R-package `mixedpower`, which is designed to simulate data and power of linear mixed effects models from pilot data (see Kumle et al., 2021) for a tutorial). Table 5 shows the power estimates for the effects for N = {20,30,40,50,60} participants. It can be seen that for the effect of gesture on stress accuracy we already have enough power to detect an actual effect at $N = 20$ and higher. Thus, our design with 20 participants or more is sensitive enough to detect a gesture effect of a small effect size, raising confidence that our study has sufficient power to be scientifically informative.

Table 5. Power analysis for confirmatory research question 1, main effect only

N	**20**	**30**	**40**	**50**	**60**
Gesture vs. no gesture	**1**	**1**	**1**	**1**	**1**

In Table 6, the power calculations are given for a more complex model with a three-way interaction, and the lower-order interaction effects between L1-L2 stress (mis)match, written stress mark presence, and gesture condition. It can be seen that especially for main effects and two-way interactions involving stress mark presence, we need substantially more participants to reach comparable power as compared to the other variables. Further note that even in this complex model, the gesture condition and L1-L2 stress mark main effects and their interactions reach > 80% for 20 participants. The optimal and planned number of participants is therefore set at $N = 30$, with a lower bound of 20 participants should we experience substantial difficulty recruiting suitable participants.

Table 6. Power analysis for confirmatory research question 1, main effects and interactions

N	**20**	**30**	**40**	**50**	**60**
Gesture vs. no gesture	**0.96**	**0.99**	**1.00**	**1.00**	**1.00**
stress mark presence (yes vs. no)	**0.20**	**0.33**	**0.38**	**0.47**	**0.51**
L1-L2 stress mismatch (yes vs. no)	**1**	**1**	**1**	**1**	**1**
Written stress mark presence * Gesture	**0.02**	**0.02**	**0.04**	**0.05**	**0.04**
L1-L2 stress mismatch * Gesture	**0.80**	**0.94**	**0.99**	**1**	**1**
Written stress mark presence * L1-L2 stress mismatch	**1**	**1**	**1**	**1**	**1**

Written stress mark presence * L1-L2 stress mismatch * Gesture	**0.85**	**0.97**	**0.99**	**1.00**	**1.00**

Outliers and exclusions

Any unforeseen data exclusions will be reported in the eventual research report. We do not expect exclusions based on outliers, but should we decide to do so, we will report our results with and without such exclusions. Whenever a participant produces a segmental speech error, we will allow the participant to retry until successful, only including the last successful retry in our analyses. Should a participant not understand or not follow the instructions, their data will be excluded from analysis and a new participant will be tested.

Open Data

All raw and processed data (including audio recordings) supporting the study will be uploaded to our OSF/Github project page, with only one exception. Due to privacy issues, the videos from the participants will not be publicly available. However, these are available to researchers upon request.

References

Alviar, C., Dale, R., Dewitt, A., & Kello, C. (2020). Multimodal Coordination of Sound and Movement in Music and Speech. *Discourse Processes*, *57*(8), 682–702. https://doi.org/10.1080/0163853X.2020.1768500

Aragonés Fernández, L., & Palencia del Burgo, R. (2010). *Gramática de uso del español: Teoría y práctica A1-B2*. Ediciones SM.

Archibald, J. (1992). Transfer of L1 Parameter Settings: Some Empirical Evidence from Polish Metrics. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *37*(3), 301–340. https://doi.org/10.1017/S0008413100019903

Archibald, J. (1993). The Learnability of English Metrical Parameters by Adult Spanish Speakers. *IRAL : International Review of Applied Linguistics in Language Teaching*, *31*(2), 129–142.

Aruin, A. S., & Latash, M. L. (1995). Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements. *Experimental Brain Research*, *103*(2), 323–332. https://doi.org/10.1007/BF00231718

Bates, D., Maechler, M., Bolker [aut, B., cre, Walker, S., Christensen, R. H. B., Singmann, H.,

Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & simulate.formula), P. N. K. (shared copyright on. (2021). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (1.1-27.1) [Computer software]. https://CRAN.R-project.org/package=lme4

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004

Bolinger, D. (1983). Intonation and Gesture. *American Speech*, *58*(2), 156–174. https://doi.org/10.2307/455326

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B*, *288*(1943), 1–9. https://doi.org/10.1098/rspb.2020.2419

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 157–175. https://doi.org/10.1177/0265532212455394

Bosker, H. R., Sjerps, M. J., & Reinisch, E. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Scientific Reports*, *10*(5607), 1–11. https://doi.org/10.1038/s41598-020-62613-8

Braun, B., Lemhöfer, K., & Mani, N. (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*(1), 376–387. https://doi.org/10.1121/1.3500688

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1302–1310. https://doi.org/10.1109/CVPR.2017.143

Chakraborty, R., & Goffman, L. (2011). Production of Lexical Stress in Non-Native Speakers of American English: Kinematic Correlates of Stress and Transfer. *Journal of Speech, Language, and Hearing Research*, *54*(3), 821–835. https://doi.org/10.1044/1092-4388(2010/09-0018)

Chen, A. (2012). The prosodic investigation of information structure. In *The expression of information structure* (pp. 249–286). De Gruyter.

Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, *143*(3).

https://doi.org/10.1037/a0036281

Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, *39*(2), 237–245. https://doi.org/10.1016/j.wocn.2011.02.006

Connell, K., Hüls, S., Martínez-García, M. T., Qin, Z., Shin, S., Yan, H., & Tremblay, A. (2018). English Learners' Use of Segmental and Suprasegmental Cues to Stress in Lexical Access: An Eye-Tracking Study. *Language Learning*, *68*(3), 635–668. https://doi.org/10.1111/lang.12288

Cordo, P. J., & Nashner, L. M. (1982). Properties of postural adjustments associated with rapid arm movements. *Journal of Neurophysiology*, *47*(2), 287–302. https://doi.org/10.1152/jn.1982.47.2.287

Coretta, S. (2021). *A Wrapper for the Phonetic Software "Praat" [R package speakr version 3.1.1]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=speakr

Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of Encouraging the Use of Gestures on Speech. *Journal of Speech, Language, and Hearing Research*, *62*(9), 3204–3219. https://doi.org/10.21437/SpeechProsody.2018-42

Cutler, A. (2005). Lexical Stress. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception*. Blackwell Publishing Ltd.

Cutler, A., & Donselaar, W. V. (2001). Voornaam is not (really) a Homophone: Lexical Prosody and Lexical Access in Dutch: *Language and Speech*, *44*(2), 171–195. https://doi.org/10.1177/00238309010440020301

Da Silveira, A. P., van Heuven, V. J., Caspers, J., & Schiller, N. O. (2014). Dual activation of word stress from orthography. *Dutch Journal of Applied Linguistics*, *3*(2), 170–196. https://doi.org/10.1075/dujal.3.2.05sil

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

Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*, 274–294.

Dohen, M., & Roustan, B. (2017, August). Co-production of speech and pointing gestures in

clear and perturbed interactive tasks: Multimodal designation strategies. *Interspeech 2017*. https://hal.archives-ouvertes.fr/hal-02367749

Edmunds, P. (2009). *ESL speakers' production of English lexical stress: The effect of variation in acoustic correlates on perceived intelligibility and nativeness*. https://digitalrepository.unm.edu/ling_etds/10

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press.

Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., & D'Imperio, M. (2021). Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Developmental Science*, e13154. https://doi.org/10.1111/desc.13154

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

Finnegan, E. M., Luschei, E. S., & Hoffman, H. T. (2000). Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech. *Journal of Speech, Language, and Hearing Research: JSLHR*, *43*(4), 934–950. https://doi.org/10.1044/jslhr.4304.934

Goldman, J.-P. (2011). EasyAlign: An automatic phonetic alignment tool under Praat. *Proceedings of InterSpeech*, 1–4.

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. https://doi.org/10.1111/lang.12376

Guion, S. G., Clark, J. J., Harada, T., & Wayland, R. P. (2003). Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech*, *46*(Pt 4), 403–427. https://doi.org/10.1177/00238309030460040301

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish–English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*(3), 207–226. https://doi.org/10.1017/S1366728904001592

Gullberg, M., Bot, K. de, & Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gesture*, *8*(2), 149–179. https://doi.org/10.1075/gest.8.2.03gul

Gutiérrez-Palma, N., Suárez-Coalla, P., & Cuetos, F. (2020). Stress assignment in reading aloud

in Spanish. *Applied Psycholinguistics*, *41*(4), 753–769. https://doi.org/10.1017/S014271642000020X

He, L., & Dellwo, V. (2017). Amplitude envelope kinematics of speech: Parameter extraction and applications. *The Journal of the Acoustical Society of America*, *141*(5), 3582–3582. https://doi.org/10.1121/1.4987638

Henriksen, N. (2013). Suprasegmental Phenomena in Second Language Spanish. In *The Handbook of Spanish Second Language Acquisition* (pp. 166–182). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118584347.ch10

Hodges, P. W., & Richardson, C. A. (1997). Feedforward contraction of transversus abdominis is not influenced by the direction of arm movement. *Experimental Brain Research*, *114*(2), 362–370. https://doi.org/10.1007/pl00005644

Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication*, *57*, 257–267. https://doi.org/10.1016/j.specom.2013.06.007

Hoetjes, M., & Van Maastricht, L. (2020). Using Gesture to Facilitate L2 Phoneme Acquisition: The Importance of Gesture and Phoneme Complexity. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.575032

Hualde, J. I. (2005). *The Sounds of Spanish*. Cambridge University Press.

Huang, X., Kim, N., & Christianson, K. (2019). Gesture and Vocabulary Learning in a Second Language. *Language Learning*, *69*(1), 177–197. https://doi.org/10.1111/lang.12326

Im, S., & Baumann, S. (2020). Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America*, *5*(1), 685–697. https://doi.org/10.3765/plsa.v5i1.4755

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

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.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*(1), 49–66.

Jesse, A., & McQueen, J. M. (2014). Suprasegmental Lexical Stress Cues in Visual Speech can

Guide Spoken-Word Recognition. *Quarterly Journal of Experimental Psychology*, *67*(4), 793–808. https://doi.org/10.1080/17470218.2013.834371

Jesse, A., Poellmann, & Kong, Y. (2017). English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition. *Journal of Speech, Language, and Hearing Research*, *60*(1), 190–198. https://doi.org/10.1044/2016_JSLHR-H-15-0340

Jilka, M. (2000). Testing the contribution of prosody to the perception of foreign accent. In A. James & J. Leather (Eds.), *Proceedings of New Sounds 2000* (pp. 199–207).

Kager, R. W. J. (1989). *A Metrical Theory of Stress and Destressing in English and Dutch* [Utrecht University]. https://dspace.library.uu.nl/handle/1874/8680

Kattán-Ibarra, J., & Pountain, C. J. (2003). *Modern Spanish Grammar: A Practical Guide*. Psychology Press.

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*(7). https://doi.org/10.1525/collabra.76

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00673

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kleiman, E. (2017). *EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data* (0.1.3) [Computer software]. https://CRAN.R-project.org/package=EMAtools

Köhnlein, B., & Oostendorp, M. van. (2018). Where Is the Dutch Stress System?: Some New Data. In H. van der Hulst, J. Heinz, & R. Goedemans (Eds.), *The Study of Word Stress and Accent: Theories, Methods and Data* (pp. 346–360). Cambridge University Press. https://doi.org/10.1017/9781316683101.012

Kooij, J., & Van Oostendorp, M. (2003). *Fonologie: Uitnodiging tot de klankleer van het Nederlands*. Amsterdam University Press.

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

Krauss, R. M., Chen, Y., Gottesman, R. F., & McNeill, D. (2000). Lexical gestures and lexical access: A Process Model. In *Language and Gesture* (pp. 261–283). Cambridge University Press.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 1–44. https://doi.org/10.1098/rstb.2013.0397

Krivokapić, J. (2020). Prosody in articulatory phonology. In S. Shattuck-Hufnagel & J. Barnes (Eds.), *Prosodic Theory and Practice*. MIT Press.

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

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01546-0

Lenth, R., & Lenth, M. R. (2017). Package 'lsmeans'. *The American Statistician*, *34*(4), 216–221.

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

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

Levin, S. M. (2006). Tensegrity: The new biomechanics. In M. Hutson & R. Ellis (Eds.), *Textbook of muscularskeletal medicine* (pp. 69–80). Oxford University Press.

Li, P., Baills, F., & Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition*, *42*(5), 1015–1039.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). Listeners normalize speech for contextual speech rate even without an explicit recognition task. *The Journal of the Acoustical*

Society of America, *146*(1), 179–188. https://doi.org/10.1121/1.5116004

Mattys, S. L., Melhorn, J. F., & White, L. (2007). Effects of syntactic expectations on speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 960–977. https://doi.org/10.1037/0096-1523.33.4.960

McClave, E. (1998). Pitch and Manual Gestures. *Journal of Psycholinguistic Research*, *27*(2), 69–89. https://doi.org/10.1023/A:1023274823974

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

McQueen, J. M., & Dilley, L. C. (2021). Prosody and spoken-word recognition. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (p. 13). Oxford University Press.

Morett, L. M. (2018). In hand and in mind: Effects of gesture production and viewing on second language word learning. *Applied Psycholinguistics*, *39*(2), 355–381. https://doi.org/10.1017/S0142716417000388

Morett, L. M., & Chang, L.-Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, *114*(42), E8913–E8921. https://doi.org/10.1073/pnas.1705373114

Oostendorp, M. van. (2012). Quantity and the Three-Syllable Window in Dutch Word Stress. *Language and Linguistics Compass*, *6*(6), 343–358. https://doi.org/10.1002/lnc3.339

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Team, R. C. (2019). *nlme: Linear and nonlinear mixed effects models* [R].

Pouw, W., De Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2020). Gesture-speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.14532

Pouw, W., & Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony

under delayed auditory feedback. *Cognitive Science*, *43*(3), e12721. https://doi.org/10.1111/cogs.12721

Pouw, W., Harrison, S. J., & Dixon, J. A. (2019). Gesture-speech physics: The biomechanical basis of the emergence of gesture-speech synchrony. *Journal of Experimental Psychology: General*, *149*(2), 391–404. https://doi.org/10.1037/xge0000646

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

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences*, *117*(12), 11364–11367. https://doi.org/10.1073/pnas.2004163117

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2019). Acoustic specification of upper limb movement in voicing. *Proceedings of the 6th Meeting of Gesture and Speech in Interaction*, 75–80. https://doi.org/10.17619/UNIPB/1-812

Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., Schaefer, R., & Wiggins, G. (2021). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2020.0334

Pouw, W., Trujillo, J., & Dixon, J. A. (2019). The quantification of gesture-speech synchrony: A tutorial and validation of multi-modal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01271-9

Pouw, W., & Trujillo, J. P. (2019). *Materials Tutorial Gespin2019—Using video-based motion tracking to quantify speech-gesture synchrony*. 10.17605/OSF.IO/RXB8J

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, *123*, 1104–1113.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing [computer program]*.

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *The Quarterly*

Journal of Experimental Psychology, *63*, 772–783.

Rietveld, A. C. M., & Van Heuven, V. J. J. P. (2009). *Algemene Fonetiek*. Uitgeverij Coutinho.

Ripperda, J., Drijvers, L., & Holler, J. (2020). Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): A toolkit for the automatic detection of hand movements and gestures in video data. *Behavior Research Methods*, *52*(4), 1783–1794. https://doi.org/10.3758/s13428-020-01350-2

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J. (2008). The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1507–1521. https://doi.org/10.1044/1092-4388(2008/07-0173)

Saalfeld, A. K. (2012). Teaching L2 Spanish Stress. *Foreign Language Annals*, *45*(2), 283–303. https://doi.org/10.1111/j.1944-9720.2012.01191.x

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Language and Speech*, *52*(2–3), 135–175. https://doi.org/10.1177/0023830909103165

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

Steffman, J. (2021). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, *0*(0), 1–26. https://doi.org/10.1080/23273798.2020.1862259

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*(1), 177–193. https://doi.org/10.1016/j.jml.2011.08.001

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235.

Treffner, P. J., & Peter, M. (2002). Intentional and attentional dynamics of speech–hand coordination. *Human Movement Science*, *21*(5–6), 641–697. https://doi.org/10.1016/S0167-9457(02)00178-1

Tremblay, A., Broersma, M., & Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language.

Bilingualism: Language and Cognition, *21*(3), 640–652. https://doi.org/10.1017/S136672891700030X

Trettenbrein, P. C., & Zaccarella, E. (2021). Controlling Video Stimuli in Sign Language and Gesture Research: The OpenPoseR Package for Analyzing OpenPose Motion-Tracking Data in R. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.628728

Trommelen, M., Zonneveld, W., & Jessen, M. (1999). Word stress in West Germanic languages. In G. Bossong, B. Comrie, & H. V. D. Hulst (Eds.), *Word prosodic systems in the languages of Europe* (pp. 478–515). Mouton de Gruyter.

Turvey, M. T., & Fonseca, S. T. (2014). The medium of haptic perception: A tensegrity hypothesis. *Journal of Motor Behavior*, *46*(3), 143–187. https://doi.org/10.1080/00222895.2013.798252

Van der Heijden, L. (2021). *Gestures in Learning Foreign Language Prosody: The Importance of Considering Task and Learner Characteristics* [MA Thesis]. Radboud University.

Van der Heijden, L., Van Maastricht, L., & Hoetjes, M. (2021). Using gestures in L2 prosody acquisition training: The role of individual differences. *Proceedings of the 4th Phonetics and Phonology in Europe Conference*.

Van Maastricht, L., Hoetjes, M., & Van Drie, E. (2019). *Do gestures during training facilitate L2 lexical stress acquisition by Dutch learners of Spanish?* Internation Conference on Auditory-Visual Speech Processing (AVSP2019), Melbourne.

Van Maastricht, L., Krahmer, E., & Swerts, M. (2016a). Native speaker perceptions of (non-) native prominence patterns: Effects of deviance in pitch accent distributions on accentedness, comprehensibility, intelligibility, and nativeness. *Speech Communication*, *83*, 21–33. https://doi.org/10.1016/j.specom.2016.07.008

Van Maastricht, L., Krahmer, E., Swerts, M., & Prieto, P. (2019). Learning direction matters: A study on L2 rhythm acquisition by Dutch learners of Spanish and Spanish learners of Dutch. *Studies in Second Language Acquisition*, *41*(1), 87–121.

Van Maastricht, L. van, Krahmer, E., & Swerts, M. (2016b). Prominence Patterns in a Second Language: Intonational Transfer From Dutch to Spanish and Vice Versa. *Language Learning*, *66*(1), 124–158. https://doi.org/10.1111/lang.12141

Van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Fluency in dialogue: Turn-taking behavior shapes perceived fluency in native and non-native speech. *Language Learning*, *70*(4),

1183–1217. https://doi.org/10.1111/lang.12416

Vilà-Giménez, I., Igualada, A., & Prieto, P. (2019). Observing storytellers who use rhythmic beat gestures improves children's narrative discourse performance. *Developmental Psychology*, *55*(2), 250–262. https://doi.org/10.1037/dev0000604

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

Winkelmann, R., Bombien, L., & Scheffers, M. (2018). *wrassp: Interface to the "ASSP" Library* (0.1.8) [Computer software]. https://CRAN.R-project.org/package=wrassp

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619. https://doi.org/10.1016/j.cognition.2021.104619

Study Context

This interdisciplinary study arose out of a shared interest in gesture-prosody coupling and with an additional aim to perform fully open team science. Each of the present authors is used to approaching the topic of gesture-prosody coupling from their own perspective (L2 acquisition [LvM, MH], gesture studies [MH], human movement [WP], prosody production and perception [LvM, HRB]). The present team study is a unique opportunity to combine these approaches and accompanying toolkits in one team, and present our approach under a fully shared authorship. We believe that the study of communication movement and spoken language requires an understanding of the language spoken, the body that speaks it, and the cognitive processes that regulate actions within these constraints. Since this kind of understanding has not been reached by any of us individually, we hope a combined effort helps us attain it together.

Contributions

	HRB	MH	WP	LvM
Conceptualization	■	■	■	■
Methodology	■	■	■	■
Analysis Plan	■		■	■
Analysis & post-processing coding			■	
Experiment coding			■	
Data collection			■	■
Writing - Original Draft	■	■	■	■
Writing - Review & Editing	■	■	■	■

Appendix

trial	target	written stress mark presence	L1/L2 stress (mis) match	stressed syllable L1	stressed syllable L2	stressed syllable L1-L2
C1.I.1	**análisis**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.2	**centímetro**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.3	**pirámide**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.4	**cardiólogo**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.5	**máquina**	**yes**	**mismatch**	**2**	**1**	**1**
C1.I.6	**automóvil**	**yes**	**mismatch**	**4**	**3**	**1**
C1.I.7	**teléfono**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.8	**fórmula**	**yes**	**mismatch**	**2**	**1**	**1**
C1.I.9	**acróbata**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.10	**satélite**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.11	**gráfico**	**yes**	**mismatch**	**2**	**1**	**1**
C1.I.12	**océano**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.13	**carnívora**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.14	**católico**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.15	**demócrata**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.16	**gramófono**	**yes**	**mismatch**	**3**	**2**	**1**

C1.I.17	**micrófono**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.18	**metáfora**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.19	**catástrofe**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.20	**cafetería**	**yes**	**mismatch**	**3**	**4**	**-1**
C1.I.21	**sofá**	**yes**	**mismatch**	**1**	**2**	**-1**
C1.I.22	**simpático**	**yes**	**mismatch**	**3**	**2**	**1**
C1.I.23	**símbolo**	**yes**	**mismatch**	**2**	**1**	**1**
C1.I.24	**kilómetro**	**yes**	**mismatch**	**3**	**2**	**1**
C2.I.1	**clima**	**no**	**mismatch**	**2**	**1**	**1**
C2.I.2	**vacaciones**	**no**	**mismatch**	**2**	**3**	**-1**
C2.I.3	**profesor**	**no**	**mismatch**	**2**	**3**	**-1**
C2.I.4	**ventilador**	**no**	**mismatch**	**3**	**4**	**-1**
C2.I.5	**internet**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.6	**mamut**	**no**	**mismatch**	**1**	**2**	**-1**
C2.I.7	**horizonte**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.8	**doctor**	**no**	**mismatch**	**1**	**2**	**-1**
C2.I.9	**radiador**	**no**	**mismatch**	**2**	**3**	**-1**
C2.I.10	**ilustrador**	**no**	**mismatch**	**3**	**4**	**-1**

C2.I.11	**color**	**no**	**mismatch**	**1**	**2**	**-1**
C2.I.12	**error**	**no**	**mismatch**	**1**	**2**	**-1**
C2.I.13	**factor**	**no**	**difference**	**1**	**2**	**-1**
C2.I.14	**festival**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.15	**uniforme**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.16	**alcohol**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.17	**actor**	**no**	**mismatch**	**2**	**2**	**0**
C2.I.18	**alergia**	**no**	**mismatch**	**3**	**2**	**1**
C2.I.19	**farmacia**	**no**	**mismatch**	**3**	**2**	**1**
C2.I.20	**aristocracia**	**no**	**mismatch**	**5**	**4**	**1**
C2.I.21	**elefantes**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.22	**carnaval**	**no**	**mismatch**	**1**	**3**	**-2**
C2.I.23	**democracia**	**no**	**mismatch**	**4**	**3**	**1**
C2.I.24	**voleibol**	**no**	**mismatch**	**1**	**3**	**-2**
C3.I.1	**brócoli**	**yes**	**match**	**1**	**1**	**0**
C3.I.2	**político**	**yes**	**match**	**2**	**2**	**0**
C3.I.3	**cámara**	**yes**	**match**	**1**	**1**	**0**
C3.I.4	**número**	**yes**	**match**	**1**	**1**	**0**

C3.I.5	**fútbol**	**yes**	**match**	**1**	**1**	**0**
C3.I.6	**saxofón**	**yes**	**match**	**3**	**3**	**0**
C3.I.7	**carácter**	**yes**	**match**	**2**	**2**	**0**
C3.I.8	**analítico**	**yes**	**match**	**3**	**3**	**0**
C3.I.9	**círculo**	**yes**	**match**	**1**	**1**	**0**
C3.I.10	**álbum**	**yes**	**match**	**1**	**1**	**0**
C3.I.11	**ángel**	**yes**	**match**	**1**	**1**	**0**
C3.I.12	**capitán**	**yes**	**match**	**3**	**3**	**0**
C3.I.13	**pragmática**	**yes**	**match**	**2**	**2**	**0**
C3.I.14	**volcán**	**yes**	**match**	**2**	**2**	**0**
C3.I.15	**teoría**	**yes**	**match**	**3**	**3**	**0**
C3.I.16	**geográfico**	**yes**	**match**	**3**	**3**	**0**
C3.I.17	**académico**	**yes**	**match**	**3**	**3**	**0**
C3.I.18	**romántico**	**yes**	**match**	**2**	**2**	**0**
C3.I.19	**espárrago**	**yes**	**match**	**2**	**2**	**0**
C3.I.20	**helicóptero**	**yes**	**match**	**3**	**3**	**0**
C3.I.21	**histórico**	**yes**	**match**	**2**	**2**	**0**
C3.I.22	**económico**	**yes**	**match**	**3**	**3**	**0**

C3.I.23	**energía**	**yes**	**match**	**3**	**3**	**0**
C3.I.24	**ecología**	**yes**	**match**	**4**	**4**	**0**
C4.I.1	**parasol**	**no**	**same**	**3**	**3**	**0**
C4.I.2	**universidad**	**no**	**match**	**5**	**5**	**0**
C4.I.3	**formulario**	**no**	**match**	**3**	**3**	**0**
C4.I.4	**mango**	**no**	**match**	**1**	**1**	**0**
C4.I.5	**tomate**	**no**	**match**	**2**	**2**	**0**
C4.I.6	**princesa**	**no**	**match**	**2**	**2**	**0**
C4.I.7	**cables**	**no**	**match**	**1**	**1**	**0**
C4.I.8	**proceso**	**no**	**match**	**2**	**2**	**0**
C4.I.9	**papel**	**no**	**match**	**2**	**2**	**0**
C4.I.10	**programa**	**no**	**match**	**2**	**2**	**0**
C4.I.11	**ensalada**	**no**	**match**	**2**	**3**	**-1**
C4.I.12	**chocolate**	**no**	**match**	**3**	**3**	**0**
C4.I.13	**limonada**	**no**	**match**	**3**	**3**	**0**
C4.I.14	**restaurante**	**no**	**match**	**3**	**3**	**0**
C4.I.15	**blusa**	**no**	**match**	**1**	**1**	**0**
C4.I.16	**cabina**	**no**	**match**	**2**	**2**	**0**

C4.I.17	**familia**	**no**	**match**	**2**	**2**	**0**
C4.I.18	**canal**	**no**	**match**	**2**	**2**	**0**
C4.I.19	**centro**	**no**	**match**	**1**	**1**	**0**
C4.I.20	**insectos**	**no**	**match**	**2**	**2**	**0**
C4.I.21	**planeta**	**no**	**match**	**2**	**2**	**0**
C4.I.22	**mandarina**	**no**	**match**	**3**	**3**	**0**
C4.I.23	**sistema**	**no**	**match**	**2**	**2**	**0**
C4.I.24	**bailarina**	**no**	**match**	**3**	**3**	**0**