scientific reports

OPEN

Co-speech gestures influence the magnitude and stability of articulatory movements: evidence for coupling-based enhancement

Karee Garvin✉, Eliana Spradling & Kathryn Franich

Humans rarely speak without producing co-speech gestures of the hands, head, and other parts of the body. Co-speech gestures are also highly restricted in how they are timed with speech, typically synchronizing with prosodically-prominent syllables. What functional principles underlie this relationship? Here, we examine how the production of co-speech manual gestures influences spatiotemporal patterns of the oral articulators during speech production. We provide novel evidence that words uttered with accompanying co-speech gestures are produced with more extreme tongue and jaw displacement, and that presence of a co-speech gesture contributes to greater temporal stability of oral articulatory movements. This effect–which we term *coupling enhancement*–differs from stress-based hyperarticulation in that differences in articulatory magnitude are not vowel-specific in their patterning. Speech and gesture synergies therefore constitute an independent variable to consider when modeling the effects of prosodic prominence on articulatory patterns. Our results are consistent with work in language acquisition and speech-motor control suggesting that synchronizing speech to gesture can entrain acoustic prominence.

Keywords Speech, Co-speech gestures, Articulation, Prosody, Speech-motor coupling

Co-speech gestures–the term for movements of the hands, arms, head, eyebrows, and other parts of the body that accompany speech–are ubiquitous in human language use. Although gestures have been shown to have a facilitative effect on speech perception and language processing[1–3], gestures occur even when communication is not face-to-face[4], suggesting their functional role is not limited to aiding the perceiver. Indeed, evidence indicates that even congenitally blind speakers utilize co-speech gestures, suggesting that visual input may not even be a precursor to gesture acquisition[5]. A body of recent research has suggested that co-speech gestures may play a more integral role in speech production. For example, co-speech gestures tend to occur synchronously with pitch-accented syllables in speech, suggesting that the planning of gesture and speech is closely linked[6–10,11,12]. Furthermore, speech tends to be more fluent when accompanied by co-speech gestures in individuals who stutter[13] as well as individuals with aphasia of speech[14,15]. Speaking while gesturing appears to bear some similarities to speaking with an external timekeeper (like a metronome), the key similarity being the act of synchronizing speech with another system[16,17].

Despite the intriguing links between gesture and stability of speech, little work has sought to directly investigate how synchronization between speech and co-speech gesture may influence speech production. Insights from motor control research provide some hints as to the nature of this relationship. A long line of research into movement dynamics in both humans and non-human animals has shown that movements of the upper limbs that are coordinated in-phase, or synchronously, with other movements are relatively larger and more stable than asynchronous or uncoupled movements[18–23]. For example, in non-speech motor tasks such as interlimb coordination, synchronous coordination has been shown to result in higher movement amplitude and greater timing stability of arm movements[24]. Increased amplitude and stability of synchronized movement is observed even when coordinating movement of the limbs to an external stimulus, such as a metronome[25]. In these studies, the positive relationship between movement amplitude and stability is striking given that larger movements are generally found to involve greater variability in timing[26,27]. These studies provide evidence that coordinated upper limb movements are subject to greater stability under a limited set of coordinative regimes: most notably, in-phase coupling[20,22].

Harvard University, Cambridge, MA, USA. ✉email: garvinkaree@gmail.com

In the speech domain, there is also evidence that synchronization leads to greater temporal stability of movements. For example, research from child language acquisition, speech errors, and articulatory timing suggests that the stability of syllable onsets is relatively greater than that of syllable codas for English and a variety of other languages[28–31]. Articulatory gestures associated with syllable onsets are also found to show extensive temporal overlap with those of a following vowel in several languages, while vowel and coda articulatory gestures tend to occur in sequence, rather than in synchrony[30]. There are a variety of models that have been developed to account for this relationship between gestural timing and stability, all of which incorporate notions of relative timing between articulatory landmarks, be they movement onsets[30], movement targets[32,33], or a combination of these landmarks[34]. Speech coordinated synchronously with external timekeepers or between individuals is also found to be more stable than uncoordinated speech. For example, in a study of metronome-timed speech[35], speakers of languages with different prosodic profiles demonstrated increased duration of metronome-synchronized syllables (as opposed to those produced on the offbeat of the metronome), as well as reduced variability in durations for synchronized syllables. Research has also shown that speech synchronized between speakers ('joint speech') is relatively less temporally variable than speech spoken alone[36]. Taken together, these findings suggest that speech movement is conditioned by the same principles of coordination-based enhancement shown in limb movement.

There has been comparably little investigation of the effects of co-speech gestures on speech itself. One study[37] had participants utter short sentences, manipulating where participants were to produce co-speech gestures within the sentence, and where the nuclear pitch accent was to be produced in the sentence. Thus, in a Dutch sentence like *Amanda gaat naar Malta* ('Amanda goes to Malta'), the relevant pitch accent would either be produced on the word *Amanda* or *Malta* (aligning with the underlined lexically-stressed syllable), and a beat gesture was to be produced either concurrently with the pitch accent ('congruent' conditions), or on the opposite word from the pitch accent ('incongruent' conditions). The study found that participants produced longer acoustic durations and lower second formant frequencies on syllables where a gesture was present, even in incongruent conditions, where there was no accompanying pitch accent. Although these results suggest a clear effect of gesture on speech, the authors note that the act of pairing a gesture with an unaccented syllable in the incongruent condition may have been unnatural, leading participants to produce a prominence where one was not intended. Thus, this study leaves open at least two possibilities for the source of gesture's effect on speech: (a) prosodic enhancement of unaccented syllables, or (b) a more direct effect of gestural-speech coupling on syllable duration.

Research questions and hypotheses

In this paper, we investigate how the presence vs. absence of a temporally-synchronized co-speech gesture (CSG) can influence the articulation of speech. We utilize electromagnetic articulography (EMA) to measure the position and velocities of the oral articulators of speakers of English to examine how the presence of a co-speech gesture influences both the magnitude and timing of articulatory movements. We likewise explore how increased magnitude and temporal stability manifest in the context of synchronization to tease apart the role of gesture from speech prosody.

Literature on enhancement effects in speech production has proposed two main mechanisms through which increased articulatory magnitude could be realized: (i) *hyperarticulation,* whereby articulatory movements are made more extreme in vowel-specific ways: front vowels become fronter, back vowels become backer, low vowels become lower, and high vowels become higher[38–40]; and (ii) *sonority expansion,* whereby articulators like the tongue and jaw will be realized with more extreme downward positions in the presence of a gesture[38]. In short, hyperarticulation serves to increase the vowel space (in all directions) through the enhancement of articulatory movements, whereas sonority expansion serves to increase the overall openness of the oral cavity, and thus does not predict any vowel-specific effects of enhancement on articulatory movement direction. We investigate how coupling impacts speech magnitude by comparing these two dimensions of increased gestural magnitude. There is overwhelming evidence from both speech production and perception that accentual prominence in English is associated with at least some level of hyperarticulation of vowels, particularly in tongue position[39–42]. Thus, if gestures were found to induce articulatory enhancement without hyperarticulation, this would suggest that gesture-based enhancement effects were independent from those related to prosodic prominence more generally.

To provide a baseline for our hypotheses, we first investigate two primary aspects of our data to confirm that our results replicate the essential findings from key prior studies: (i) co-speech gestures are timed to stressed/pitch-accented syllables (H1), and (ii) stressed syllables exhibit effects of hyperarticulation when compared to unstressed syllables (H2). We formalize these and our main hypotheses (H3-H4) as follows:

H1: Co-speech gestures (CSGs) will be timed to the pitch peak of the stressed syllable.

H2: Tongue gestures will have more extreme target achievements in stressed syllables compared to unstressed syllables, with high vowels realized with higher tongue positions, low vowels with lower tongue positions, back vowels with more back tongue positions, and front vowels with more front tongue positions (consistent with hyperarticulation).

H3: Coupling with CSG will lead to increased magnitude of movement of oral articulators.

H3a hyperarticulation: The direction of the effect will differ between vowels, with /i/ realized as fronter and/or higher and /o, a/ realized as lower and/or backer.

H3b sonority-expansion: The direction of the effect is consistent across vowel types, where the oral articulators have a more open posture for /i, o, a/, regardless of vowel quality.

H4: Coupling with CSG leads to increased temporal stability of movement between oral articulators.

To test these predictions, we analyze productions of CVCV tokens with alternating stress on either the first syllable (initial stress) or the second syllable (final stress), where each stress condition was produced with (CSG condition) and without (no-CSG condition) an accompanying CSG. To investigate the coordinative and temporal properties of these utterances, we analyze the movement of the tongue tip (TT), tongue body (TB), and jaw (JW) for speech articulation and the CSG apex in co-speech gesture articulation, where target achievement for both oral articulations and CSGs was defined as the point of minimum speed/maximum displacement of the articulator (details on this method are provided in §4.4.2 and §4.4.3, respectively). All methods were performed in accordance with relevant guidelines and regulations.

Results

Replication of previous findings

Prior to presenting results of our study, we establish that overall patterns for co-speech gesture timing and stress-based enhancement are similar in our study to those that have been found in previous work. First, co-speech gesture apexes have been shown to correlate in time with pitch peak of stressed/pitch accented syllables[8,43,44]. In our own study, we find that the apex is timed to the target word (e.g. 'speebee') in 100% of utterances, and timed to the stressed syllable within this word in 80% of tokens. Furthermore, we conducted a Pearson correlation test between the timing of gesture apex and pitch peak, which shows a strong correlation between stressed syllable peak f0 and apex timing, as shown in Fig. 1 ($r(1513) = 0.88$, $p < 0.001$). This result confirms H1, replicating the finding of existing literature that co-speech gestures are generally timed to pitch extrema in stressed syllables.

Second, in English, stressed syllables have been shown to have more extreme tongue displacement compared to unstressed syllables. To test for these effects, we compared the position of the TB during target achievement of stressed and unstressed syllables in both the vertical (y) and horizontal (x) dimensions. A stepwise ANOVA model comparison demonstrated that both vowel quality and an interaction with stress, as well as stress as a random slope, improved the fit of the model for the horizontal dimensions in predicting target achievement values, as expected (vowel quality: $\Delta\text{AIC} = 6720$, $X^2(2, \text{N} = 6414) = 6608.1$, $p < 0.001$; vowel quality*stress: $\Delta\text{AIC} = 116$, $X^2(3, \text{N} = 6414) = 109.95$, $p < 0.001$; stress as random slope: $\Delta\text{AIC} = 12$, $X^2(2, \text{N} = 6414) = 16.324$, $p < 0.001$) and vertical (vowel quality: $\Delta\text{AIC} = 9021$, $X^2(2, \text{N} = 6414) = 8374.2$, $p < 0.001$; vowel quality*stress: $\Delta\text{AIC} = 651$, $X^2(3, \text{N} = 6414) = 633.31$, $p < 0.001$; stress as random slope: $\Delta\text{AIC} = 23$, $X^2(2, \text{N} = 6414) = 26.982$ $p < 0.001$). These results are consistent with the idea that stress-based enhancement involves vowel-specific hyperarticulation. As seen in Fig. 2, stressed vowels were realized with both lower and backer tongue positions for /a/ and /o/, but higher and fronter tongue positions for /i/, consistent with hyperarticulation, per H2a.

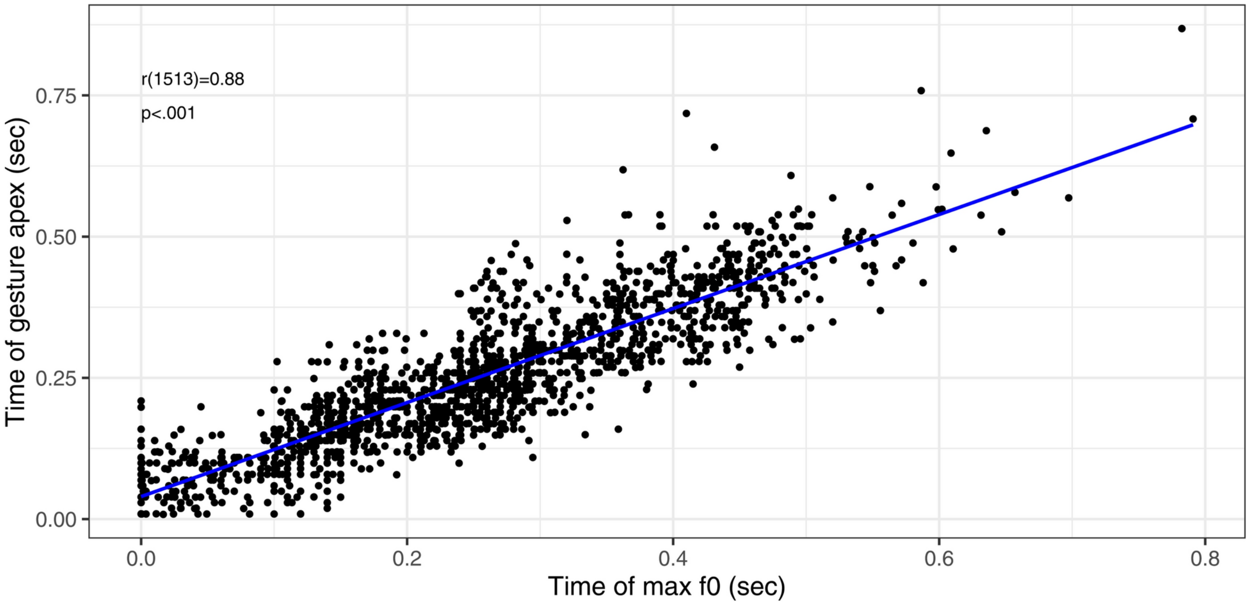


Fig. 1. Correlation of timing between F0 maximum and co-speech gesture apex, where both the time of gesture apex and time of max f0 are relative to the target word onset.

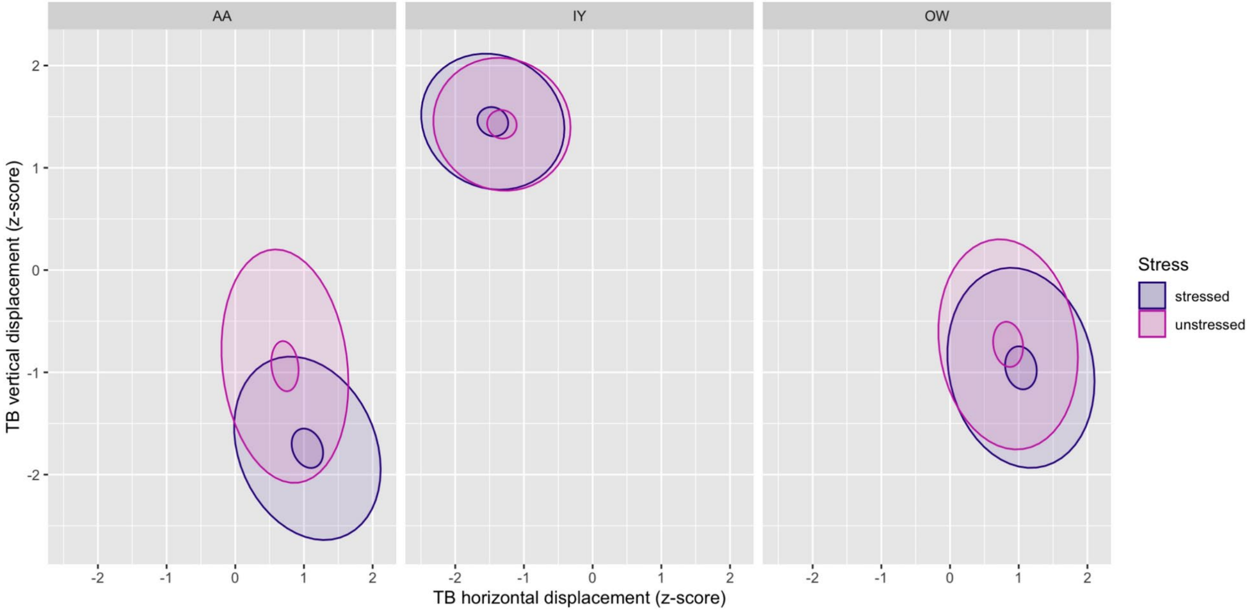


Fig. 2. TB horizontal (x axis) and vertical (y axis) displacement during target achievement for /i, o, a/. More negative values on the y-axis correspond to lower positions of the TB and more negative values on the x-axis correspond to fronter positions of the TB. The plot ellipses illustrate 5% of the mean and one standard deviation of the mean.

Effect of co-speech gesture on oral articulator magnitude

Hyperarticulation vs sonority expansion

We now examine how the presence of a co-speech gesture may influence speech articulations, considering first the magnitude of speech gestures from the standpoint of both hyperarticulation and sonority expansion. To evaluate these effects, we used a series of GAMM analyses of TB vertical and horizontal displacement across each of the vowel contrasts included in the study, /i, o, a/, z-scored by subject and time normalized to the target word. The use of GAMMs allows us to examine not only the overall displacement of the tongue across conditions, but also how these effects may vary over time in the articulation of the target word.

Turning first to vertical displacement, Fig. 3a,b demonstrates that for all vowels in the final stress condition and vowels /i/ and /o/ in the initial stress condition, the TB is significantly lower across vowel types in the presence of a CSG. Furthermore, differences across conditions appear to be roughly localized to the site of the CSG across conditions. Although we focus on the TB results here because TB is most closely associated with vowel production, we likewise tested TT and TD (tongue dorsum) to ensure that the hyperarticulation effect wasn't localized to the front or back of the tongue. Our results for the TT and TD are similar to those provided here for the TB. In other words, the direction of the effect was the same for all vowels, consistent with an effect of sonority expansion across the board (hypothesis H3b).

We next analyzed TB horizontal displacement across vowels in both stress conditions as a function of gesture presence. As demonstrated in Fig. 3c,d, there was no significant difference in horizontal displacement for any vowel type for either stress condition, though displacement values were numerically in the same direction across vowels. Results were comparable for TT and TD. Together, these results support H2b, consistent with sonority expansion, and we find no support for H2a, the hyperarticulation hypothesis.

Effects of co-speech gesture on gesture magnitude across oral articulators

In line with prior work on coupling-based enhancement, results from our own data presented in §2.2.1 suggest that the vertical movements of the tongue are enhanced for target words when produced in time with a co-speech gesture, regardless of vowel quality. To further delve into the effect of CSG coupling on articulatory gesture enhancement, we analyzed both the acoustic duration of stressed vowels (pooled across all vowel qualities) in target words, as well as the magnitude of movement of various oral articulators in the presence vs. absence of a CSG.

Our results reveal increased acoustic duration of stressed syllables when coordinated with a co-speech gesture, as illustrated in Fig. 4. An ANOVA model comparison illustrates that the predictor stress improved model fit (stress: ΔAIC = 948,420, X^2(1, N = 156,361) = 44,633, $p < 0.001$) as did the interaction between stress and gesture, with final-stress tokens showing a larger effect of gesture presence (stress*gesture: ΔAIC: = 20,981, X^2(2, N = 156,361) = 4836.5, $p < 0.001$). Stress as a random slope likewise improved model fit (ΔAIC = 16,149,X^2(2, N = 156,361) = 16,153, $p < 0.001$). These results demonstrate acoustic enhancement of vowels, where stressed vowels are longer when coordinated with a CSG in both stress conditions, consistent with[37].

We also analyzed the kinematic enhancement of tongue sensor trajectories using a series of GAMM analyses of vertical displacement and velocity, where values in the y axis are z-scored by subject and the x axis is time

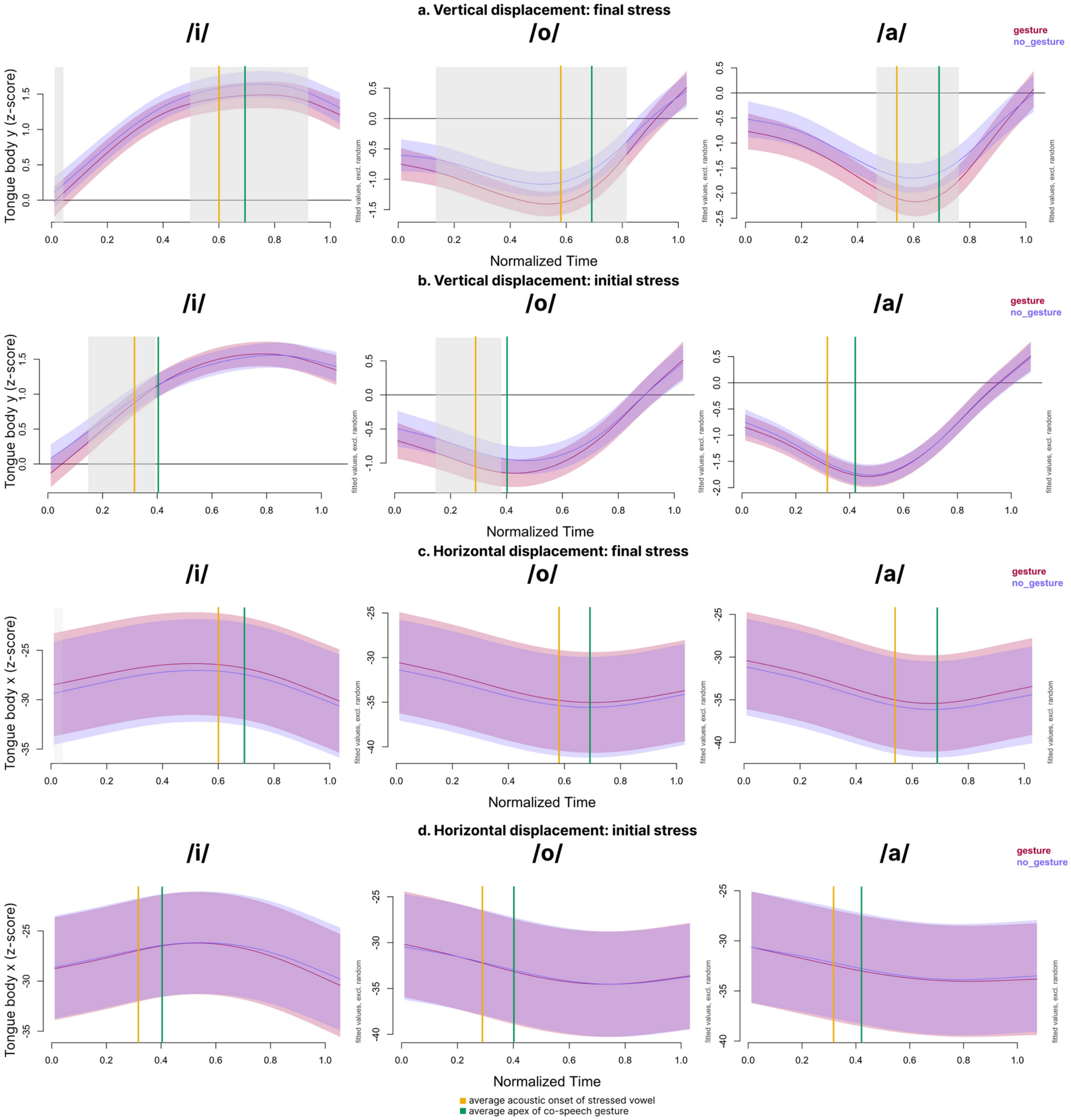


Fig. 3. GAMMs of TB vertical and horizontal displacement as a function of CSG presence for /i, o, a/ across stress conditions, z-scored by subject. In (**a**) and (**b**) the top of the plot corresponds with higher tongue posture and the bottom corresponds with lower tongue posture. In (**c**) and (**d**) the top of the plot corresponds with a fronter tongue posture and the bottom corresponds with a backer posture. Shading indicates portions of trajectory that are significantly different. x-axis time is normalized by target word.

normalized to the target word. For final stress tokens (Fig. 5a,b), the results reveal greater displacement of both the TT and TB, with the significant difference in gesture magnitude between CSG and no-CSG conditions extending throughout the target word for the TT, and timed at the locus of the stressed syllable and the point of maximum displacement in the TB. In other words, maximum displacement of both the TT and TB during the stressed vowel was greater when the target word was coordinated with a CSG.

As expected based on prior work[45], increases in gesture magnitude for stressed syllables were accompanied by increases in articulatory velocity in the CSG condition. For the TB, downward movement in preparation for the stressed vowel target and the closure movement during the stressed syllable both occur at higher velocities when accompanied by a CSG. In the TT gesture, closure movement of the tongue following maximum displacement of the stressed vowel occurs later in the CSG condition.

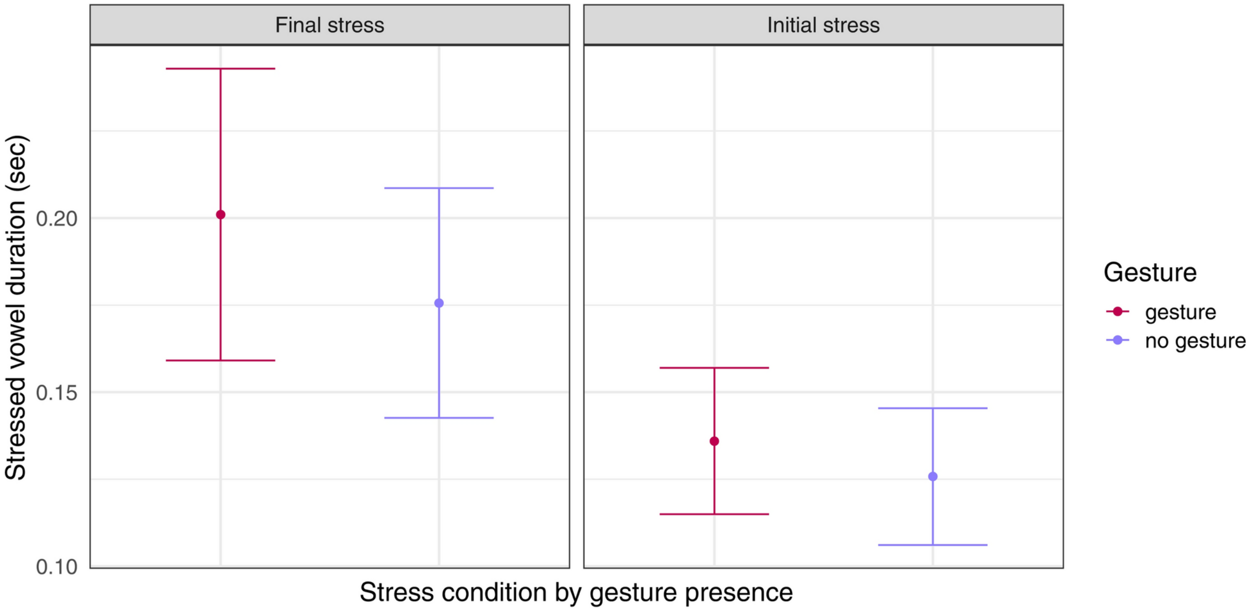


Fig. 4. Acoustic vowel duration in stressed syllables in final stress (left) and initial stress (right) tokens.

Results in the initial stress condition were more modest, yet still exhibit several effects of coupling enhancement. Namely, Fig. 5c,d demonstrates that TB vertical displacement at the onset of the stressed syllable was significantly greater and the TT gesture was significantly longer when speech was accompanied with a CSG. Nevertheless, there was no significant difference in TT vertical displacement and the TB velocity patterned opposite from the predicted direction of effect; we return to a discussion of these effects and differences between stress conditions in §4. Together, these results demonstrate that coupling of speech with CSG leads to enhancement effects on oral gestures, consistent with studies on synchronized limb movement[24,25].

Despite significant differences in the magnitude of the TT and TB gestures across subjects, GAMMs analyses of JW gestures across subjects revealed no significant difference in JW horizontal or vertical displacement or velocity in either stress condition. We return to this finding in light of individual variability in JW displacement in §2.2.3.

Individual differences in jaw magnitude enhancement

The asymmetry in results between the tongue and the jaw was surprising, as the jaw has been shown to be an important effector in the articulation of prosody and sonority expansion[40]. However, prior work has demonstrated individual differences in the extent to which speakers move their jaw while speaking due to anatomical differences between speakers[46]. Accordingly, we analyzed individual differences in JW vertical displacement. The analyses revealed that some subjects do indeed show a significant difference in vertical JW displacement, with greater displacement in the CSG condition (Table 1). Importantly, there was an implicational relationship between vertical tongue movement and JW movement: if a participant had greater JW displacement in the CSG condition, they also had greater tongue displacement in this condition; yet the reverse was not true. These results are consistent with individual differences found in Johnson [46] suggesting that some speakers make greater use of the tongue than the jaw in vowel articulations.

Effect of co-speech gesture on temporal stability

In addition to greater movement magnitude, studies on coupling enhancement have shown greater temporal stability of synchronized vs. unsynchronized movement[18,24,25,35]. We analyzed temporal stability in synchronization between articulators in producing the stressed vowel. We focus here on the stability between the TB and other articulators because of the clear effect of CSG on TB magnitude. In addition, we are able to analyze the relative synchronization of multiple articulators, i.e., TB, JW, and CSG, which all have a maximum displacement landmark that occurs during the stressed vowel. Specifically, both the tongue and the jaw are implicated during the production of a stressed vowel and in addition, the CSG apex also tends to co-occur with the stressed vowel; thus, the measure of variability in the timing of maximum displacement in the TB and JW can be understood as how consistently the maximum displacement of the TB and JW are timed during stressed vowel production. Figure 6 demonstrates that there is less variability in the timing of TB and JW maximum displacement in the presence of a CSG compared to tokens produced without a CSG. An ANOVA model comparison demonstrates that the predictor stress significantly improves model fit (stress: ΔAIC = 3576, X^2(1, N = 156,361) = 7.1077, $p < 0.01$), as does the interaction with gesture (stress*gesture: ΔAIC = 3571, X^2(2, N = 156,361) = 281.47, $p < 0.001$), and stress as a random slope (ΔAIC = 3294, X^2(2, N = 156,361) = 3297.9, $p < 0.001$),

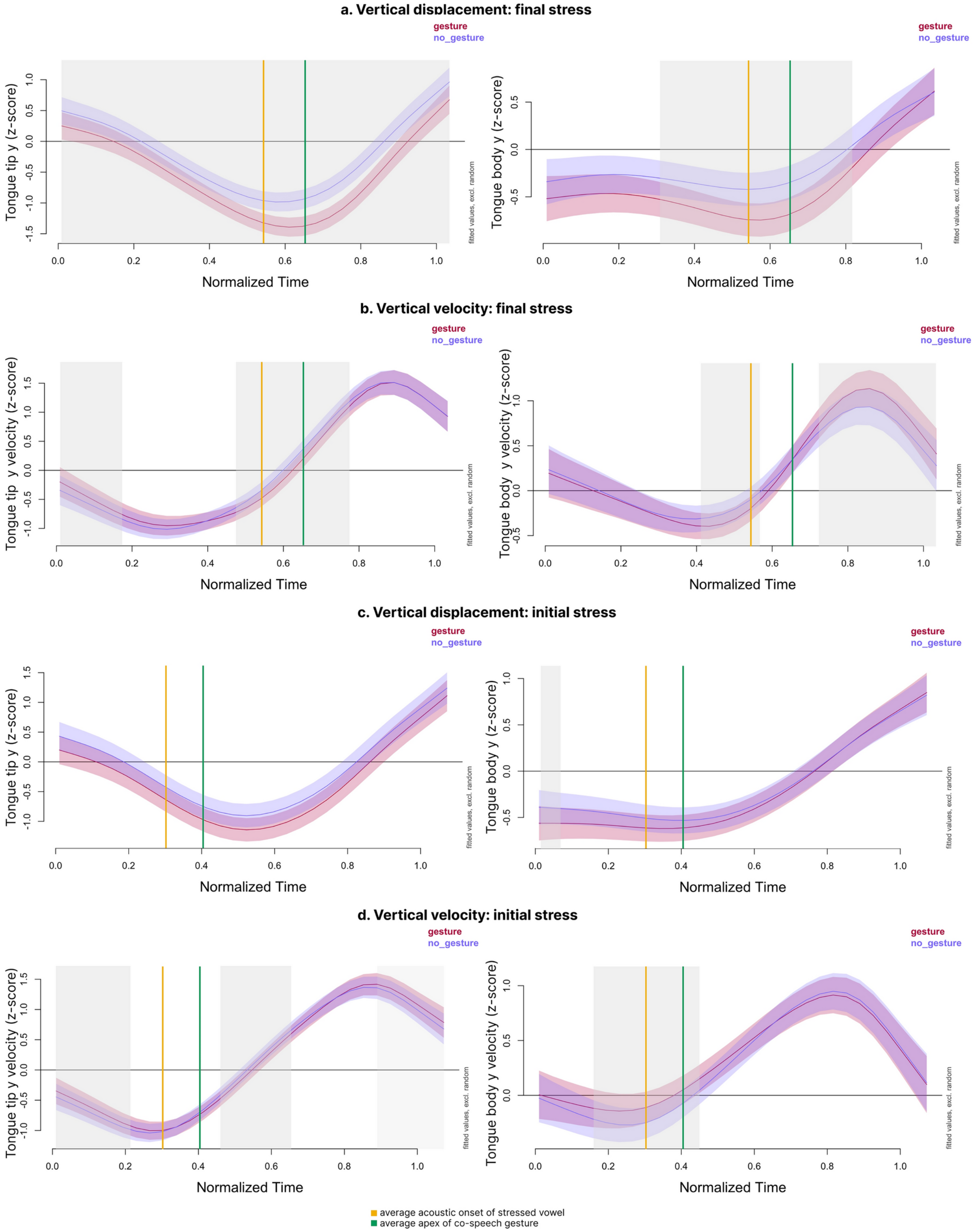


Fig. 5. GAMMs of vertical displacement and velocity of the TT (left) and TB (right) across stress conditions, z-scored by subject. Shading indicates portions of trajectory that are significantly different. x-axis time is normalized by target word. In (**a**) and (**c**) the top of the plot corresponds to higher positions of the articulator and the bottom of the plot corresponds to lower positions of the articulator. In the velocity plots (**b**) and (**d**), values above the zero line indicate upward vertical movement and values below the zero line indicate downward vertical movement. Values near zero correspond with low velocity. Where values cross the zero line, this indicates a change in the direction of the articulator.

Individual differences in jaw magnitude										
Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Lower tongue position with CSG	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lower jaw position with CSG	✓	✓	n.s	✓	X	n.s	X	✓	✓	n.s

Table 1. By-subject comparison of magnitude effect of CSG on tongue (TT and/or TB) and jaw displacement in analyses. Checkmark indicates a significantly lower position in articulator vertical displacement during the stressed syllable during the gesture condition compared to the no gesture condition; n.s. indicates no significant difference; X indicates a significant difference in the opposite direction, where the articulator was significantly higher during the stressed syllable in the gesture condition versus the no gesture condition.

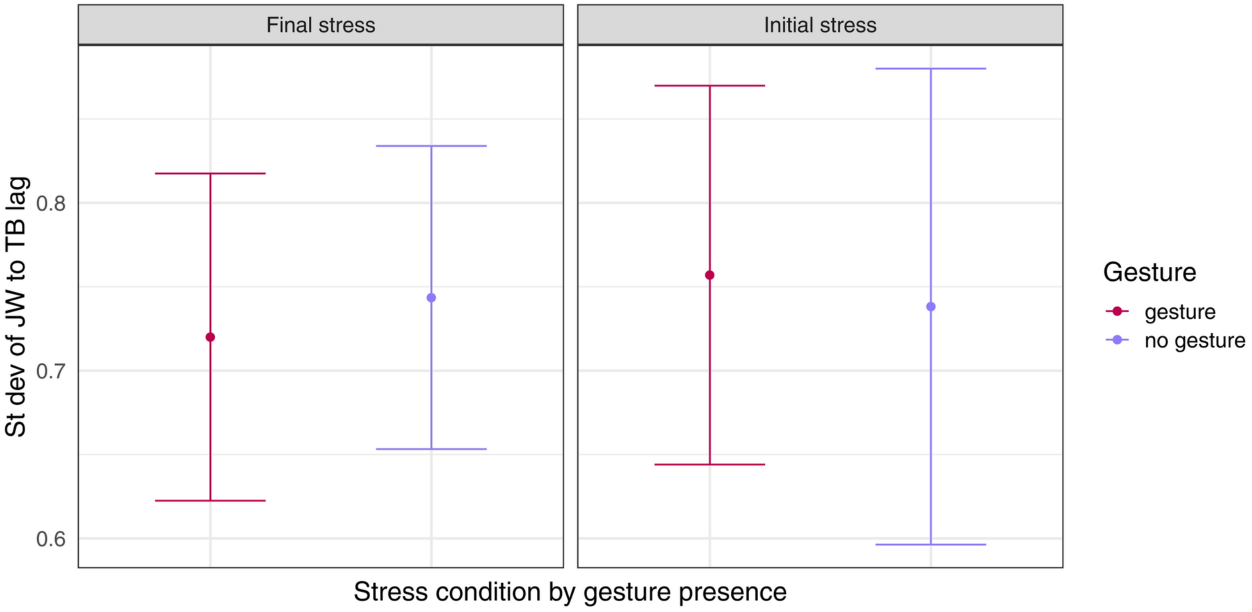


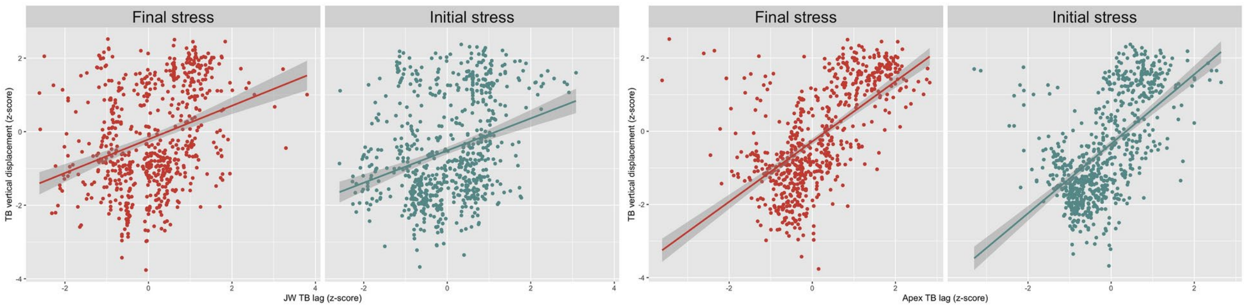
Fig. 6. Variability in synchronization (st dev) between maximum displacement of TB and JW in stressed syllables.

As mentioned in §1, studies suggest a positive relationship between the strength of temporal coupling of movements and the magnitude of movement[18,24,25,35]. We test this hypothesis in our data by analyzing the correlation between the temporal lag, i.e., synchronization, between TB and JW movements, and the magnitude of TB vertical displacement. We also examined the correlation between the lag in timing between gesture apex and TB and the magnitude of TB vertical displacement. A Pearson correlation demonstrates a significant positive correlation between TB-JW lag and TB displacement (Fig. 7a) in both stress conditions (Final Stress: $r(650) = 0.32$, $p < 0.001$; Initial Stress: $r(693) = 0.30$, $p < 0.001$). A stronger correlation was observed between Apex-TB lag and TB displacement (Fig. 7b) in the final stress condition (Final Stress: $r(650) = 0.61$, $p < 0.001$; Initial Stress: $r(693) = 0.58$, $p > 0.001$). In the case of Apex-TB lag, values clustered around zero in the raw data, with some subjects showing a consistent positive lag. We z-scored the lag by subject to normalize differences between subjects. Overall, these results demonstrate that greater synchronization is correlated with higher magnitude in jaw movement. Specifically, in Fig. 7, higher magnitude corresponds with more negative values, as more negative values are associated with a more open posture of the tongue, resulting in a positive correlation. We return to differences across stress conditions and explore possible sources of these differences in §4.

Discussion

Co-speech gestures are ubiquitous in naturalistic communication. Although they are known to aid in perception of speech and in conveying semantic and pragmatic information, our study provides evidence that the facilitative role of gestures extends to speech production. Specifically, our results suggest that speech movements timed to co-speech gestures show enhanced magnitude and temporal stability. Consistent with H3, across all subjects in our sample, vowel durations were longer and tongue displacement was lower for speech produced synchronously with a manual co-speech gesture than speech produced without a gesture. Most subjects likewise achieved lower JW positions for target words in utterances containing a gesture.

Complementing the increase in magnitude found in the gesture condition, we also find evidence that articulatory movements were less variable when produced with co-speech gestures, consistent with H4. Within stressed vowels (the syllables to which gestures were most closely timed in our data), variability in lag between



a. TB/JW synchronization vs. magnitude

b. TB/apex synchronization vs. magnitude

Fig. 7. Positive correlation between articulatory magnitude and stability. Lower (more negative) values in vertical displacement indicate a lower posture of the tongue, i.e., increased movement magnitude. The x-axis shows TB vertical displacement. The y-axis shows the lag between TB and JW maximum displacement (**a**) and TB maximum displacement and CSG apex timing (**b**) during the stressed vowel. All values are z-scored by subject.

target achievement of TB and JW movements (measured in terms of standard deviation) was found to be lower overall in the CSG condition. Furthermore, we find a direct link between stability of timing to the CSG and magnitude of oral articulatory gestures, as shorter lag times between TB and CSG are correlated with increased movement amplitude of the TB across both stress conditions. Taken together, our results display the hallmarks of coupling-based movement enhancement and stabilization that have been described within several domains of motor control[18,21,24,25,28,30,35].

One interpretation of our results is that by producing speech with gestures, speakers are, in effect, stabilizing their own articulatory movements and enhancing coordination between those movements. The idea that coupled movements should be more stable than uncoupled movements is familiar from research on the well-known ‘bimanual advantage’ in tasks such as finger tapping, in which variability in tapping within the hands is reduced when the two hands are tapping together, as opposed to when participants tap with just one hand[47]. The effect has been found to persist even when non-homologous digits are coordinated together across the two limbs. Under this account, speech and gesture could be seen as forming an articulatory synergy akin to that formed between the two hands or between oral articulators during speech[48]. There are various ways in which co-speech gesture could be seen to contribute to articulatory stability. One explanation, offered by Helmuth and Ivy[47] for the bimanual advantage, is that stability arises from the integration of the timekeepers across speech and gesture. Integration could occur either at the level of the timekeeping mechanism itself (e.g. through stronger coupling within a coupled oscillator model of speech and gestural timing) or at the level of timekeeper outputs in a non-oscillatory model. Still another explanation (not mutually-exclusive with the above) could be that stability of movements is related to the presence of additional sources of somatosensory feedback associated with effectors within the articulatory synergy, be they tactile[49], or a combination of tactile, auditory, and visual feedback[50].

An alternative interpretation of our results is that presence of a gesture is associated with (but not necessarily causing) changes in overall *exertion* during speech, perhaps in the form of increased attention paid to articulatory movements. This state of increased exertion could also be linked with increased gestural activation, or more extreme gestural targets[51], explaining why the tongue (and jaw, for some participants) showed more extreme positions in the CSG condition. While this explanation is appealing given that gesture presence is known to be associated with more effortful modes of speech (such as those involved in prosodic focus)[9,52], the observed relationship between coupling and articulatory stability in our data is somewhat less straightforward to explain. One possibility, following Tilsen[53], is that exertive force modulates articulatory task-associated neuronal ensembles, the size of which can, in turn, influence coupling strength. Under the assumption that articulatory stability results from greater coupling strength, exertion could be seen as conditioning both gesture presence and articulatory stability. Thus, gesture need not have a direct influence on articulation. While this account seems generally plausible, it remains to be seen whether gesture occurrence can be attributed to exertion-related factors, what those factors might be, and why co-speech gesture, in particular, should be implicated in more exertitive states. We also note that in-phase coupling is seen to be associated with relatively low-exertion states in Tilsen’s model, which is not compatible with our finding that in-phase coupling between CSG and oral articulatory gestures is strongest where oral gestures are most extreme (and thus, arguably, requiring the greatest exertion).

Our findings on the effects of CSG on the magnitude of speech gestures also shed light on the relationship between co-speech gestures and speech prosody. While our results demonstrate an effect of hyperarticulation in comparing stressed and unstressed syllables (§3.1), the effects of gesture on speech are distinct from those of stress: we find a uniformly lower tongue position across vowel types in the presence of a co-speech gesture, consistent only with sonority expansion effects. Meanwhile, we find no evidence to support an added hyperarticulation effect in the presence of a co-speech gesture. Thus, it does not appear that gesture presence leads to an increase in the magnitude of stress correlates in speech. Our results also provide counterevidence to the idea that speech and gesture serve as redundant and complementary cues to prosodic prominence in the way that e.g. voice onset time, vowel duration, and fundamental frequency complement one another in cuing voicing contrasts in some languages[54]; if this were the case, we would hypothesize that oral gestures should be less extreme in the context of a CSG. However, we find no evidence of a trading relation between speech and co-speech gesture in marking prominence in our data. Instead, it appears that the presence of gesture has an effect on articulation which does not implicate phonology or prosody directly [c.f. Franich[35]]. We term this effect *coupling enhancement*.

Our findings also point to new avenues for future research on the relationship between speech articulation, gesture coordination, and rhythm in language. Our findings indicated that the effects of co-speech gestures on speech dynamics were overall stronger for words with final stress than for words with initial stress. Though we cannot say definitively why this difference arose across stress conditions, we look to research on the timing of articulatory gestures at syllable boundaries for some clues. Disyllabic words with initial stress have been shown to display greater variability of timing in tongue movements for medial consonants, as medial consonantal gestures (which would normally be expected to serve as onsets to the final syllable) tend to show an attraction to the preceding stressed syllable[55,56]. Variability of this magnitude is not observed with disyllabic final stress words. We hypothesize that this variability in coordination reduces the effect of CSG on magnitude and stability for initial stress tokens in our study.

Finally, we believe our findings can help to bridge a gap between theories of speech-motor coordination and prosody. Prior work has proposed that speech and CSG are jointly controlled by the same prosodic planning mechanism[57,58]. For example, Krivokapić et al.[58] find that both speech articulatory gestures and co-speech gestures display differences in duration as a function of prosodic prominence. They propose that the same clock-slowing mechanism for inducing longer durations on speech articulations under prominence may also operate over co-speech gestures. Our own results suggest that coupling speech to gesture can have effects that are independent of prosodic effects. However, the effects we observe in our data could act as a precursor to such a prosodic control mechanism. Assuming a more direct role for co-speech gesture in influencing speech articulation, the duration-enhancing effect of gesture-speech coupling could be grammaticalized and applied predictably to different prosodic environments, similar to the grammaticalization of lexical tone based on f0 perturbations over time[59]. Indeed, the idea that gesture can entrain articulatory and acoustic cues to prosodic prominence has been postulated on an even shorter timescale, as young children have been shown to employ prosodically appropriate co-speech gestures before they learn to produce the acoustic cues to prominence that accompany such gestures in adult speech[60]. If this hypothesis proves to be correct, it will have important implications for the study of prosody and language typology. Concretely, given cross-linguistic variability in the timing between co-speech gestures and speech, we may expect to see correlations between gesture timing strategies and prosodic patterns across the world's languages. A better understanding of gesture as a potential phonetic precursor may help to shed light on patterns of sound change and the evolutionary relationship between speech and co-speech gesture.

Methods

Participants

Prior approval from the Institutional Review Board (IRB) was obtained for this study and appropriate protocols were followed for data collection. All methods were performed in accordance with relevant guidelines and regulations. This included obtaining participants' written informed consent to participate in the study. Ten subjects (1 male, 8 female, 1 unspecified) participated in the study. Participants consisted of students and young professionals recruited in the greater Boston, Massachusetts area; ages ranged from 20 to 40. All participants reported US English as their native language and none of the participants reported any history of speech/language or vision impairment.

Materials

Stimuli consisted of nonce words of the shape CVbV, controlling for vowel quality: /i, o, a ~ ə/, initial consonant: /s, p, l/, stress (initial vs. final): /sóbo/ or /sobó/. Vowels were chosen to mimic stressed and unstressed vowels in English while sampling a range of the total vowel space. As will be detailed below, target words were produced in one of two conditions: with or without a concurrent co-speech gesture. Examples of the target words are provided in Table 2. All target words were produced in the carrier phrase *I saw the____ today*, to situate the target word in the environment V#CVCV#C, to maximize clarity in articulatory data parsing. The task was blocked by stress condition and by co-speech gesture condition, with the order of blocks randomized by participant. Half of participants produced the gesture block followed by the non-gesture block and the other half produced the non-gesture block followed by the gesture block. Each token was repeated 6 times throughout the task block and the order of tokens was randomized within the task block. A total of 216 tokens per subject were produced (9 word shapes × 2 stress conditions × 2 gesture conditions × 6 repetitions).

	/s/	**/p/**	**/l/**
/i /	/síbi, sibí/	/píbi, pibí/	/líbi, libí/
	< seeybee, seebeey >	< peeybee, peebeey >	< leeybee, leebeey >
/o /	/sóbo, sobó/	/póbo, pobó/	/lóbo, lobó/
	< sohbo, soboh >	< pohbo, poboh >	< lohbo, loboh >
/a ~ ə/	/sábə, səbá/	/pábə, pəbá/	/lábə, ləbá/
	< suhbah, sahbuh >	< puhbah, pahbuh >	< luhbah, lahbuh >

Table 2. Target word shapes included in the study, where IPA transcripts are provided on the first line and orthographic transcripts seen by participants are included on the second line.

Procedure

Stimuli were displayed using OpenSesame software[61] on a computer monitor at a comfortable distance from each participant. A member of the research team manually progressed through each experimental item to ensure that each token was produced correctly before moving on to the next item. If the participant made a speech error, the researcher prompted the participant to try again and the target utterance was immediately repeated.

Each task block contained a short set of instructions that told the participant what to expect during the task and provided an example sentence to illustrate stress production. The instructions for each task block were read aloud by a researcher. Sample sentences on instruction screens contained words that replicated the same stress construction, but with phones that were not present in the stimulus set. In order to ensure that the prosodic context was similar between gesture and non-gesture conditions, for all items, participants were given the same prompt, shown in Ex 1.

1. Imagine you're traveling in a foreign land and see a famous landmark while out exploring. You run into a friend in your travels and excitedly tell them about your experience.

Prior to the start of the gesture condition experiment block, participants were presented with a single demonstration video of a researcher naturally producing a sample sentence with a bimanual co-speech gesture produced synchronously with the target word. Participants were asked to model their gesture after the video, and to use both hands for the gesture, but were not explicitly told to copy all aspects of the model gesture. Participants were also not explicitly told when they should begin or end the gesturing within the spoken sentence. Participants were only instructed to produce a manual gesture as they read each sentence aloud.

Data collection

Acoustic data collection

The primary acoustic data used for acoustic analyses in this study were collected using a Rode NTG2 shotgun microphone. The microphone was attached to a boom arm mounted to the desk and positioned above the participant's head. This data is time-aligned with the EMA data files, where for the primary acoustic data and EMA data, each utterance was saved individually. However, the video data was not time aligned to the primary acoustic data and EMA recordings; thus, secondary acoustic data was also recorded for the video data in order to time align the two data sets, as discussed in §4.5.1. Secondary acoustic data time-aligned to the video was recorded using a Zoom Q8 Handy Video Recorder Microphone.

EMA data collection

EMA data was collected using the NDI Wave system, a point-tracking system with accuracy within approximately 0.5 mm[62] and a sampling rate of 100 Hz. NDI Wave 5 Degree of Freedom (5DoF) sensors were attached to the face and mouth to capture movement of the articulators. Reference sensors were used to track the position and movement of the head in order to correct articulatory data for head movement, as discussed in more detail in §4.4.2. Reference sensors consisted of five NDI Wave 5DoF sensors. Three 5DoF reference sensors were placed on the right mastoid (RMA), left mastoid (LMA) and on the bridge of the nose (NAS), as illustrated in Fig. 8a; these sensors remained in place for the entirety of the experiment. Two additional sensors were used to record the occlusal plane of each participant. Sensors were attached along the sagittal midline to a wax bite plate with one sensor aligned with the front incisor (OS) and the other aligned with the back molar (MS), as illustrated in Fig. 8b. Prior to the presentation of the stimuli, participants were asked to hold still and held the bite plate between their teeth for five seconds while a recording was made of their position. The bite plate was then removed and set aside for the remainder of the experiment.

Movement of the oral articulators were tracked using six 5DoF sensors attached to the incisors, lips, and tongue, as shown in Fig. 9. Sensors were attached to the upper and lower lips near the vermilion border, lower jaw to the gums below the lower incisor (JW), and three sensors were attached along the sagittal midline of the tongue at the tongue tip (TT), blade (TB), and dorsum (TD). The front-most sensor (TT) was attached less than 1 cm from the tip of the tongue, the back-most sensor (TD) was attached as far back as was comfortable for the participant, typically 4–6 cm from the tongue tip, and the third sensor (TB) was placed midway between the TT and TD sensors.

The participant was seated in front of the experiment monitor and beside the EMA magnet. A researcher then created a mold of the participant's bite and reference sensors were attached before collecting the bite plate recording. Next, positions for the oral sensors were marked before gluing to ensure consistency in placement in the case of re-gluing and then attached accordingly. Tongue and jaw sensors were attached using glue and lip sensors were attached using medical-grade tape. Before beginning stimuli collection, the researcher engaged the participant in conversation for several minutes to help them adjust to speaking with the attached sensors.

Throughout the study, one researcher was assigned the task of controlling the experiment program while another was assigned the task of monitoring the sensor tracking to ensure that all sensors were actively tracking. The researcher monitoring the sensor tracking would interrupt the experiment to notify the other researcher of a sensor that began to behave irregularly. In such cases, sensor stability was assessed and sensors were secured as needed.

Video data collection

Video data was collected using a Zoom Q8 Handy Recorder with a 160° wide-angle lens and a 30 fps framerate. The camera was mounted directly above the participant's monitor and angled to ensure it captured the entirety of a participant's manual gestures, from resting position in the participant's lap to full extension during the manual beat gesture. Examples of still images from the video recording are provided in Fig. 10, where (a) demonstrates

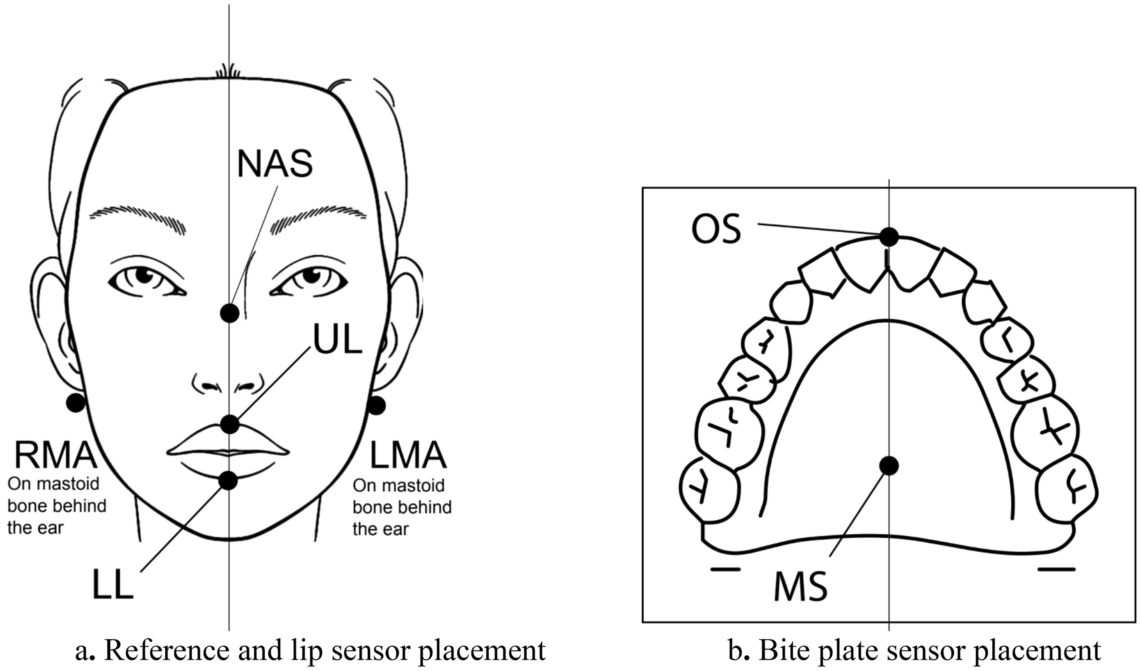

NAS
UL
RMA
On mastoid bone behind the ear
LMA
On mastoid bone behind the ear
LL
OS
MS
a. Reference and lip sensor placement
b. Bite plate sensor placement

Fig. 8. Reference and lip sensor placement.

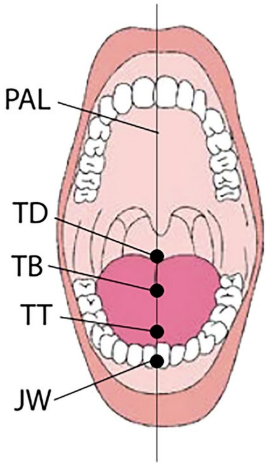


Fig. 9. Tongue, lip, and jaw sensor placement.

a no-gesture condition where the hands were kept at rest in the lap and (b) demonstrates a gesture condition at the point of maximum extension during the beat gesture production. Each task block comprised its own video recording.

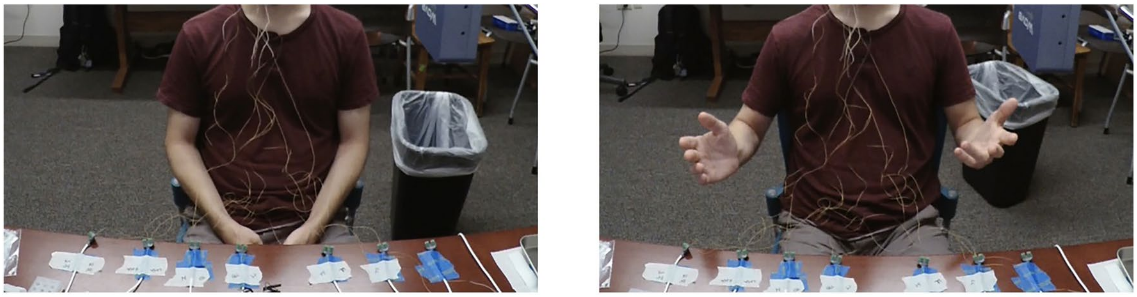
Informed consent was obtained from all subjects and/or their legal guardian(s) for publication of identifying information/images in an online open-access publication; however, out of consideration for the participant, we cropped the face to protect their privacy.

Data processing

Acoustic data processing

Primary acoustic data was force-aligned using the Montreal Forced Aligner (MFA)[63]. Pitch maxima were extracted from each vowel in target words using a script for Praat[64].

Primary acoustic data, which was time-aligned to the EMA recordings, and secondary acoustic data, which was time-aligned to the video data, was merged using the Python tool *audalign*, which identifies similarities across audio files in order to allow for time alignment of signals[65]. The alignment between primary and secondary acoustic data was reviewed by the researchers to identify and correct any misalignments between the two data



a. No gesture condition

b. Gesture condition

Fig. 10. Example stills of video camera positions demonstrating target production during the gesture and no gesture conditions.

	b	p	l	s	a ~ ə	i	o
Primary	LL/UL	LL/UL	TT	TT	JW	JW	JW
Secondary	–		TB	TB	TB	TB	TB
Tertiary						TT	LL/UL

Table 3. Articulators used in determining target achievement for a given segment. The secondary tier of the table was only used if no target could be identified from the primary level, and the tertiary tier was only used if no target could be identified from either the primary or the secondary level.

sets. Corrections were made by identifying a unique acoustic landmark within the utterance that could be used to synchronize the two recordings.

EMA data processing

The EMA data was rotated along the mid-sagittal plane and head movements were corrected for in Python[66] using the reference and bite plate sensors[67], such that the origin of the spatial coordinates corresponds to the front teeth. All articulatory trajectories were smoothed using Garcia's robust smoothing algorithm[68]. All EMA data was visually inspected by a research assistant to ensure that no recordings containing major tracking errors or detached sensors were included in the data analysis. Major errors in the recording were identified by ensuring that all movements were consistent with possible movement of the articulators. Recordings that included any major errors in tacking were removed from the dataset prior to analysis.

Following a similar procedure to that implemented in[69], relevant gestural targets, including the point of minimum and maximum displacement in the horizontal and vertical plane, as well as local x and y speed minima, were automatically identified in Python using a window determined by the acoustic signal. The articulator used to determine target achievement differs depending on the segment quality and the kinematic profile of the utterance. Target identification proceeded hierarchically based on the articulatory parameters of the segment, where if a target could not be detected on the basis of the first articulator, a target was assessed using the second and finally third articulators, as detailed in Table 3. Typically, the secondary and tertiary levels were only needed in the case of unstressed vowels, which are not the focus of this study.

Video data processing

Co-speech gestures were coded by a team of researchers trained in gesture coding using ELAN[70] following the MIT Gesture Studies Coding Manual[71], which outlines several phases of the gesture including preparations, strokes, holds, and recoveries based on[7].

The apex of the gesture was automatically extracted based on manual annotations of gesture strokes using MultiPose[72], which uses MediaPipe[73] to track pixel movement in the video recording to extract a set of xy coordinates for a given articulator. In this study, we identified the right wrist as the most stable articulator for identifying the apex of the co-speech gesture (though there was one left-handed participant in the sample, he produced all gestures with both hands). The MultiPose workflow can identify several key kinematic landmarks for a given articulator. In this study, we defined the CSG apex as the xy speed minimum, which closely corresponds to the point of maximum extension.

Statistical analysis

The data was analyzed in R[74] using a combination of Pearson Correlation Testing, Linear Mixed Effects Models (lmer)[75], ANOVA model comparisons[76], and Generalized Additive Mixed Models (GAMM)[77].

Linear mixed effects models and ANOVA model comparisons

Linear mixed effects (lme) models along with a series of ANOVA model comparisons were used to evaluate differences in displacement of oral articulators at specified timepoints and duration of vowels in target words. We used an lme model to model the interaction of stress and vowel quality on TB vertical and horizontal displacement at the point of achievement during production of stressed and unstressed vowels (§2.1). We likewise used lme models and ANOVA model comparisons to predict the interactional effect of stress and gesture on a number of variables including stressed vowel duration (§2.2.2) and synchronization between TB and JW within the stressed vowel (§2.3).

Subject was included as a random intercept for all lme models. Following[78], models were initially fit with maximal random effects structures, and random slope parameters were only reduced from the model if they eliminated singularity[78]. The resulting models are illustrated in (2), where X is determined by the analysis as outlined above.

A nested series of ANOVA model comparisons was conducted in a stepwise fashion to assess the significance of the predictors in the model. ANOVA comparisons were conducted in a pairwise fashion beginning with the baseline model (3.a/b.1) and increasing in complexity incrementally. The reported results included two commonly used metrics for ANOVA model comparison (1) the ΔAIC value, calculated as the AIC value of a given model minus the AIC value of the best fit model within the set, and (2) and Likelihood Ratio Tests with associated *p*-values[76].

2. Linear mixed effects model structures:

    a. lmer(X ~ vowel_quality*stress + (1 + stress|Subject))
    b. lmer(X ~ stress*gesture + (1 + stress|Subject))

3. ANOVA model comparison structures:

    a. lmer(X ~ vowel_quality*stress + (1 + stress|Subject)

        1. lmer(X ~ 1 + (1 +|Subject))
        2. lmer(X ~ vowel_quality + (1|Subject))
        3. lmer(X ~ vowel_quality*stress + (1|Subject)
        4. lmer(X ~ vowel_quality*stress + (1 + stress|Subject)

    b. lmer(X ~ stress*gesture + (1 + stress|Subject))

        1. lmer(X ~ 1 + (1|Subject))
        2. lmer(X ~ stress + (1|Subject))
        3. lmer(X ~ stress*gesture + (1|Subject))
        4. lmer(X ~ stress*gesture + (1 + stress|Subject))

Generalized additive mixed models

Generalized Additive Mixed Model (GAMM) analyses were implemented using the *bam* package to assess differences in displacement and velocity of the oral articulators over time during target word production between the gesture and the no gesture conditions. All GAMM analyses used the basic model formula provided in (3), where X is either the vertical displacement, horizontal displacement, or vertical velocity of the relevant sensor, where all variables were z-scored by subject. Time is normalized such that the onset of the target word corresponds with zero and the offset of the target word corresponds with 1. This method of time normalization provided the best alignment between target achievement of the stressed vowel targets without making assumptions about the nature of the kinematic trace. These models predict a given articulatory trajectory with gesture presence, smoothed time, and time smoothed by gesture as fixed effects, the random intercepts of subject and gesture, and the random smooths of subject and time. This model gives both the nonlinear and the constant difference between the gesture tasks.

4. Generalized additive mixed model structure:

    a. bam(X ~ gesture + s(time) + s(time, by = gesture, bs = "tp", k = 10) + s(subject, gesture) + s(time, subject)

Across all data in the GAMM analyses, time was normalized to the production of the target word, and articulatory trajectories and velocities were compared in the presence of a CSG (gesture condition) and the absence of a CSG (no-gesture condition).

Defined measures of analysis

In this section, we define each of the measures used in the analysis of this study.

5. **Apex (AX)** is defined as the point of minimum speed of the right wrist during the execution of a gesture.
6. **Time of max f0** is defined as the time of the pitch peak for the phone coinciding with the gesture apex.
7. **TB vertical displacement** is defined as the vertical (y) position of the TB sensor during the target achievement of the vowel.
8. **TB horizontal displacement** is defined as the horizontal (x) position of the TB sensor during the target achievement of the vowel

9. **Stressed vowel duration**: Acoustic duration of the stressed vowel as defined by the start point and end point of the parsed segment in the forced alignment process.
10. **JW to TB lag**: Lag between maximum extension of the TB and JW during stressed vowel production.
11. **JW to TB lag std**: Standard deviation of the TB and JW lag grouped by token and subject (6 repetitions/token/subject).

Data availability

Data files and scripts for statistical analysis can be found at https://osf.io/tuc86/.

Received: 11 September 2024; Accepted: 19 December 2024
Published online: 02 January 2025

References

1. Sueyoshi, A. & Hardison, D. M. The role of gestures and facial cues in second language listening comprehension. *Lang. Learn.* **55**, 661–699 (2005).
2. Hostetter, A. B. When do gestures communicate? A meta-analysis. *Psychol. Bull.* **137**, 297–315 (2011).
3. Goldin-Meadow, S. & Alibali, M. W. Gesture's role in speaking, learning, and creating language. *Annu. Rev. Psychol.* **64**, 257–283 (2013).
4. Bavelas, J., Gerwing, J., Sutton, Ch. & Prevost, D. Gesturing on the telephone: Independent effects of dialogue and visibility. *J. Mem. Lang.* **58**, 495–520 (2008).
5. Özçalışkan, Ş, Adamson, L. B., Dimitrova, N. & Baumann, S. Early gesture provides a helping hand to spoken vocabulary development for children with autism, down syndrome, and typical development. *J. Cognit. Dev.* **18**, 325–337 (2017).
6. Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M. & Prieto, P. The timing of head movements: The role of prosodic heads and edges. *J. Acoust. Soc. Am.* **141**, 4727–4739 (2017).
7. Kendon, A. Gesticulation and speech: Two aspects of the process of utterance. In *The Relationship of Verbal and Nonverbal Communication* (ed. Key, M. R.) 207–228 (De Gruyter Mouton, 1980).
8. Leonard, T. & Cummins, F. The temporal relation between beat gestures and speech. *Lang. Cogn. Processes* **26**, 1457–1471 (2011).
9. Loehr, D. P. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab. Phonol.* **3**, 71–89 (2012).
10. Rochet-Capellan, A., Laboissière, R., Galván, A. & Schwartz, J.-L. The speech focus position effect on jaw–finger coordination in a pointing task. *J. Speech Lang. Hear. Res.* **51**, 1507–1521 (2008).
11. Pouw, W., Harrison, S. J., Esteve-Gibert, N. & Dixon, J.A. 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, 1231-1247 (2020).
12. Pouw, W. & Fuchs, S. Origins of vocal-entangled gesture. Neuroscience and Biobehavioral Reviews. 141, 104836; https://doi.org/10.1016/j.neubiorev.2022.104836 (2022).
13. Mayberry, R. I. & Jaques, J. Gesture production during stuttered speech: Insights into the nature of gesture–speech integration. In *Language and Gesture* (ed. McNeill, D.) 199–214 (Cambridge University Press, 2000).
14. Devanga, S. R. & Mathew, M. Exploring the use of co-speech hand gestures as treatment outcome measures for aphasia. *Aphasiology* 1–25 (2024).
15. Jenkins, T. & Pouw, W. Gesture–Speech Coupling in Persons With Aphasia: A Kinematic-Acoustic Analysis. Journal of Experimental Psychology: General. 152, 1469–1483 (2023).
16. Brady, J. P. Studies on the metronome effect on stuttering. *Behav. Res. Ther.* **7**, 197–204 (1969).
17. Toyomura, A., Fujii, T. & Kuriki, S. Effect of external auditory pacing on the neural activity of stuttering speakers. *NeuroImage* **57**, 1507–1516 (2011).
18. von Holst, E. The behavioural physiology of animals and man in *The collected papers of Eric von Holst*. (University of Miami Press, 1973)
19. Zhang, M., Kelso, J. A. S. & Tognoli, E. Critical diversity: Divided or united states of social coordination. *PLoS ONE* **13**, e0193843. https://doi.org/10.1371/journal.pone.0193843 (2018).
20. Haken, H., Kelso, J. A. S. & Bunz, H. A theoretical model of phase transitions in human hand movements. *Biol. Cybern.* **51**, 347–356 (1985).
21. Beek, P. J., Peper, C. E. & Stegeman, D. F. Dynamical models of movement coordination. *Hum. Mov. Sci.* **14**, 573–608 (1995).
22. Kelso, J. A. S. *Dynamic Patterns: The Self-Organization of Brain and Behavior* (MIT Press, 1995).
23. De Poel, H. J., Roerdink, M., Peper, C. E. & Beek, P. J. A re-appraisal of the effect of amplitude on the stability of interlimb coordination based on tightened normalization procedures. *Brain Sci.* **10**, 10100724. https://doi.org/10.3390/brainsci10100724 (2020).
24. Schwartz, M., Amazeen, E. L. & Turvey, M. T. Superimposition in interlimb coordination. *Hum. Mov. Sci.* **14**, 681–694 (1995).
25. Kudo, K., Park, H., Kay, B. A. & Turvey, M. T. Environmental coupling modulates the attractors of rhythmic coordination. *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 599–609 (2006).
26. Fitts, P. M. The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psychol.* **47**, 381–391 (1954).
27. Messier, J. & Kalaska, J. F. Differential effect of task conditions on errors of direction and extent of reaching movements. *Exp. Brain Res.* **115**, 469–478 (1997).
28. Jacobson, R. *Child Language, Aphasia and Phonological Universals* (De Gruyter Mouton, 1968).
29. Browman, C. P. & Goldstein, L. M. Some notes on syllable structure in articulatory phonology. *Phonetica* **45**, 140–155 (1988).
30. Löfqvist, A. & Gracco, V. L. Interarticulator programming in VCV sequences: Lip and tongue movements. *J. Acoust. Soc. Am.* **105**, 1864–1876 (1999).
31. Goldstein, L., Pouplier, M., Chen, L., Saltzman, E. & Byrd, D. Dynamic action units slip in speech production errors. *Cognition* **103**, 386–412 (2006).
32. Lee, D. N. General Tau Theory: Evolution to date. *Perception* **38**, 837–850 (2009).
33. Kramer, B., Stern, M., Wang, Y., Liu, Y. & Shaw, J. Synchrony and stability of articulatory landmarks in English and Mandarin CV sequences. *Proc. ICPhS*. 1022–1026 (2023).
34. Gafos, A. A grammar of gestural coordination. *Nat. Lang. Linguist. Theory* **20**, 269–337 (2002).
35. Franich, K. How we speak when we speak to a beat: The influence of temporal coupling on phonetic enhancement. *Lab. Phonol.* https://doi.org/10.16995/labphon.6452 (2022).
36. Cummins, F. & Roy, D. Using synchronous speech to minimize variability. *Acoustic Proceedings of the Institute of Acoustics*, 201–206 (2001).
37. Swerts, M. G. J. & Krahmer, E. J. Facial expressions and prosodic prominence: Effects of modality and facial area. *J. Phonet.* **36**, 219–238 (2008).
38. de Jong, K. J., Beckman, M. E. & Edwards, J. The interplay between prosodic structure and coarticulation. *Lang. Speech* **36**, 197–212 (1993).

39. de Jong, K. J. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* **97**, 491–504 (1995).
40. Erickson, D. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* **59**, 134–149 (2002).
41. Cho, T. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ, i/ in English. *J. Acoust. Soc. Am.* **117**, 3867–3878 (2005).
42. Steffman, J. Contextual prominence in vowel perception: Testing listener sensitivity to sonority expansion and hyperarticulation. *JASA Express Lett.* **1**, 045203. https://doi.org/10.1121/10.0003984 (2021).
43. Esteve-Gibert, N. & Prieto, P. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *J. Speech Lang. Hear. Res.* **56**, 850–864 (2013).
44. Krivokapic, J., Tiede, M. K., Tyrone, M. E. & Goldenberg, D. Speech and manual gesture coordination in a pointing task. In: *Proc. Speech Prosody*. 1240–1244 (2016).
45. Munhall, K. G., Ostry, D. J. & Parush, A. Characteristics of velocity profiles of speech movements. *J. Exp. Psychol. Hum. Percept. Perform.* **11**, 457–474 (1985).
46. Johnson, K. Speech production patterns in producing linguistic contrasts are partly determined by individual differences in anatomy. *UC Berkeley Phonet. Phonol. Lab Annu. Rep.* https://doi.org/10.5070/P7141042483 (2018).
47. Helmuth, L. L. & Ivry, R. B. When two hands are better than one: Reduced timing variability during bimanual movements. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 278–293 (1996).
48. Saltzman, E. L. & Munhall, K. G. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1**, 333–382 (1989).
49. Drewing, K., Hennings, M. & Aschersleben, G. The contribution of tactile reafference to temporal regularity during bimanual finger tapping. *Psychol. Res.* **66**, 60–70 (2002).
50. Studenka, B. E., Eliasz, K. L., Shore, D. I. & Balasubramaniam, R. Crossing the arms confuses the clocks: Sensory feedback and the bimanual advantage. *Psychon. Bull. Rev.* **21**, 390–397 (2014).
51. Lindblom, B. Economy of speech gestures. In *The Production of Speech* (ed. MacNeilage, P. F.) 217–245 (Springer, 1983).
52. de Jong, K. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *J. Phonet.* **32**, 493–516 (2004).
53. Tilsen, S. Exertive modulation of speech and articulatory phasing. *J. Phonet.* **64**, 34–50 (2017).
54. Lisker, L. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech* **29**, 3–11 (1986).
55. Byrd, D., Tobin, S., Bresch, E. & Narayanan, S. Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *J. Phonet.* **37**, 97–110 (2009).
56. Garvin, K. *Word-Medial Syllabification and Gestural Coordination* (University of California, 2021).
57. Parrell, B., Goldstein, L., Lee, S. & Byrd, D. Spatiotemporal coupling between speech and manual motor actions. *J. Phonet.* **42**, 1–11 (2014).
58. Krivokapić, J., Tiede, M. K. & Tyrone, M. E. A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Lab. Phonol.* https://doi.org/10.5334/labphon.75 (2017).
59. Matisoff, J. A. Tibeto-Burman tonology in an areal context. In *Proceedings of the symposium: Cross-linguistic studies of tonal phenomena: Tonogenesis, typology and related topics* (ed. Kaji, S.) 3–32 (ILCAA, 1999).
60. Esteve-Gibert, N., Lœvenbruck, H., Dohen, M. & D'Imperio, M. Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Dev. Sci.* **25**, e13154. https://doi.org/10.1111/desc.13154 (2022).
61. Mathôt, S., Schreij, D. & Theeuwes, J. OpenSesame: An open-source, graphical experiment builder for the social sciences (2012).
62. Berry, J. J. Accuracy of the NDI Wave speech research system. *J. Speech Lang. Hear. Res.* **54**, 1295–1301 (2011).
63. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In: *Proc. 18th Conference of the International Speech Communication Association*. 498–502 (2017).
64. Boersma, P., and Weenink, D. Praat: Doing phonetics by computer. 6.2.23 http://www.praat.org/ (2022).
65. Miller, B. Audalign 1.2.4. https://pypi.org/project/audalign/ (2024).
66. Van Rossum, G., & Drake Jr, F. L. Python reference manual. 3.10.12 Centrum voor wiskunde en informatica Amsterdam. (1995).
67. Johnson, K. & Sprouse, R. L. Head correction of point tracking data. *UC Berkeley PhonLab Annu. Rep.* https://doi.org/10.5070/P7151050341 (2019).
68. Garcia, D. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computat. Stat. Data Anal.* **54**, 1167–1178 (2010).
69. Tiede, M. MVIEW: Multi-channel visualization application for displaying dynamic sensor movements. (2010).
70. ELAN 6.4 https://archive.mpi.nl/tla/elan (2022).
71. MIT speech communication group gesture coding manual. http://scg.mit.edu/gesture/coding-manual.html
72. Dych, W., Garvin, K., & Franich, K. Creating multimodal corpora for co-speech gesture research. *CorpusPhon.* abstr. (2024).
73. Lugaresi et al. MediaPipe: A Framework for Building Perception Pipelines. (2019).
74. R Core Team. R: A language and environment for statistical computing. (2013).
75. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
76. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
77. Wood, S. *Generalized Additive Models: An Introduction with R* (CRC Press, 2006).
78. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).

Acknowledgements

We'd like to thank all those who participated in or helped to pilot the study. We'd also like to thank the research assistants who helped with data collection and processing, including Clarissa Briasco-Stewart, Jian Cui, Walter Dych, Vitor Lacerda Siqueira, Xiaotian (Alice) Wang, and Yuxuan (Melody) Wang. Thanks, also, to audiences at LabPhon 19 and Northwestern University for for helpful feedback.

Author contributions

KG and KF conceived of the study and wrote the main manuscript text; all authors contributed to the data collection, processing, and methods section. KG prepared all statistical analyses and figures. All authors reviewed the manuscript.

Funding

This work was supported by National Science Foundation Linguistics Program Grant No. BCS- 2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this research.

Declarations

Competing interests
The authors declare no competing interests.

Ethics declarations
This research was granted ethics approval by the Institutional Review Board of Harvard University (Protocol IRB 22-1097).

Additional information
Correspondence and requests for materials should be addressed to K.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.