Research Article

Interarticulator Speech Coordination: Timing Is of the Essence

Matthew Masapollo[a] and **Susan Nittrouer**[a]

[a] Department of Speech, Language, and Hearing Sciences, University of Florida, Gainesville

ARTICLE INFO

Article History:
Received October 15, 2022
Revision received December 1, 2022
Accepted December 3, 2022

Editor-in-Chief: Cara E. Stepp
Editor: Raymond D. Kent

https://doi.org/10.1044/2022_JSLHR-22-00594

ABSTRACT

Purpose: In skilled speech production, sets of articulators, such as the jaw, tongue, and lips, work cooperatively to achieve task-specific movement goals, despite rampant contextual variation. Efforts to understand these functional units, termed *coordinative structures,* have focused on identifying the essential control parameters responsible for allowing articulators to achieve these goals, with some research focusing on temporal parameters (relative timing of movements) and other research focusing on spatiotemporal parameters (phase angle of movement onset for one articulator, relative to another). Here, both types of parameters were investigated and compared in detail.
Method: Ten talkers recorded nonsense, disyllabic /tV#Cat/ utterances using electromagnetic articulography, with alternative V (/ɑ/–/ɛ/) and C (/t/–/d/), across variation in rate (fast–slow) and stress (first syllable stressed–unstressed). Two measures were obtained: (a) the timing of tongue-tip raising onset for medial C, relative to jaw opening–closing cycles and (b) the angle of tongue-tip raising onset, relative to the jaw phase plane.
Results: Results showed that any manipulation that shortened the jaw opening-closing cycle reduced both the relative timing and phase angle of the tongue-tip movement onset, but relative timing of tongue-tip movement onset scaled more consistently with jaw opening-closing across rate and stress variation.
Conclusion: These findings suggest the existence of an intrinsic timing mechanism (or "central clock") that is the primary control parameter for coordinative structures, with online compensation then allowing these structures to achieve their goals spatially.
Supplemental Material: https://doi.org/10.23641/asha.22144259

When we listen to a talker produce speech, we perceive a succession of qualitatively and temporally discrete phonetic segments, like beads on a string. For example, when we listen to a talker utter the word "cat," we perceive a tidy consonant–vowel–consonant (CVC) sequence: /k/➔/æ/➔/t/. However, contrary to this intuitive notion that phonetic segments exist as monolithic, encapsulated units, when we examine a talker's articulatory movements, or the acoustic consequences of them, we do not observe such temporal discreteness (Fowler, 1980). Rather, different types of articulatory gestures—constricting actions of the articulators at specified degrees of closure at specific locations within the vocal tract—are executed in partially overlapping time frames, and thus there are no clear "boundaries" between one phonetic segment and another in the physical result. Throughout the years, a range of theoretical accounts have been offered as explanations for this mismatch in physical structure and perceptual phenomenon (e.g., Browman & Goldstein, 1992; Fowler, 1980; Galantucci et al., 2006; Guenther et al., 1998; Holt & Lotto, 2008; Liberman & Mattingly, 1985; Stevens, 1998). According to one theoretical framework, termed *articulatory phonology* (AP; Browman & Goldstein, 1992; Fowler, 1980; Goldstein & Fowler, 2003; Goldstein et al., 2006; Kelso et al., 1986; Saltzman & Munhall, 1989), phonetic structure is conveyed by functional groupings of independent articulators such as the jaw, lips, and tongue, whose exquisitely coordinated movements impart

Correspondence to Matthew Masapollo: mmasapollo@phhp.ufl.edu.
Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

to the resulting acoustic signal the structure the listener requires to recover discrete phonetic segments. The stable and lawful nature of these interarticulator "synergies" (Kelso, 1986) provides skilled listeners with the necessary cues to decode the signal and reconstruct the string of intended phonetic units. Thus, the core premise of the AP view is that skilled talkers execute movements of multiple articulators in asynchronous, but precisely timed, patterns to dynamically shape the vocal tract to produce constrictions during speech; sets of articulators cooperating in this manner have been referred to as *coordinative structures* (Fowler, 1980; Kelso et al., 1984, 1986; Kelso, 1986; Turvey et al., 1978).

While numerous experimental studies (e.g., Abbs & Gracco, 1984; Folkins & Abbs, 1975; Gracco, 1988; Kelso et al., 1984, 1986; Nittrouer, 1991; Nittrouer et al., 1988; Sorensen et al., 2016; Tuller et al., 1982) have bolstered the AP view by providing direct evidence that independent articulators function as unitary ensembles during the act of speaking, a comprehensive understanding of the essential control parameters that govern those patterns of interarticulator motion is still lacking. Moreover, much of the research on speech kinematics outside the AP framework has instead concentrated on quantifying the movement stability of single articulators (e.g., jaw or upper lip) separate from other vocal tract structures, utilizing the kinematic spatiotemporal index first introduced by Smith et al. (1995). Although descriptive of the movement consistency typical across repetitions of a given utterance, this index does not characterize relations among the various articulators involved in generating a phonetic string. This leaves a gap in our understanding of the critical control parameters that underlie the complex kinematics involved in speech production, as well as how those kinematics instantiate phonetic structure.

Over the years, technological advances, such as electromagnetic articulography (EMA; Perkell et al., 1992; Rebernik et al., 2021), electroglottography (Herbst, 2020), ultrasound imaging (Whalen et al., 2005), and real-time magnetic resonance imaging (Narayanan et al., 2014; Sorensen et al., 2016), have enabled speech researchers to directly track the skilled, sound-producing movements of the vocal tract, both intraoral and laryngeal articulators normally hidden from view (the tongue, velum, and glottis) and orofacial articulators directly visible on talkers' faces (the lips and jaw). Despite these advances, however, progress in characterizing spatiotemporal coordination of these various structures has remained limited due to several challenges. First, quantifying and then relating the movements of articulators that operate on very different timescales, due to different physical properties such as mass and intrinsic velocity, has made the characterization of dynamic speech difficult. Consider, for example, relating dynamics of the tongue tip (TT) and the jaw: the TT is a soft tissue that is less massive than the bone comprising the jaw and, therefore, moves with greater velocity. Second, how and when a gesture will occur heavily depends on the context and starting position of the articulators. For example, the articulation of a TT constriction in syllable-initial position (when the vocal tract begins to open) versus syllable-final position (when the vocal tract begins to close). These different contexts will result in very different types of tongue movements, with different starting positions and different distances to travel to target constriction positions (possibly also at different movement speeds). Also consider a TT constriction 7in the context of the tongue body positions for front versus back vowels (as in /di/ and /du/). The same TT constriction is produced in both cases, but the contribution of the TT and tongue body will differ, due to the differing demands of the flanking vowels on the tongue. A further difficulty is that no single instrument is capable of simultaneously measuring all the articulators involved in speech production, resulting in various metrics utilized to assess different types of articulatory movements. Explicating how sets of articulators work cooperatively to achieve task-specific goals, in spite of rampant contextual variation, is therefore a crucial endeavor for further development of the AP framework (Browman & Goldstein, 1992; Fowler, 1980; Goldstein & Fowler, 2003; Goldstein et al., 2006).

To begin to address these conceptual and methodological issues, the current research represents the first step in a systematic line of investigation designed to explore how movements of sets of articulators operate in concert to achieve a phonetic target. The focus here is on the particular case of TT and jaw coordination during the production of VCV utterances, which has previously been investigated by Nittrouer (1991) using articulatory data obtained from the x-ray microbeam (XRMB) system (Westbury, 1994). We present a refined technique for quantifying temporal and spatiotemporal coordination of the TT and jaw across a variety of speaking manipulations, using newer EMA technology (Perkell et al., 1992; Rebernik et al., 2021) with a larger sample of talkers.

Quantifying Temporal and Spatiotemporal Characteristics of Interarticulator Coordination

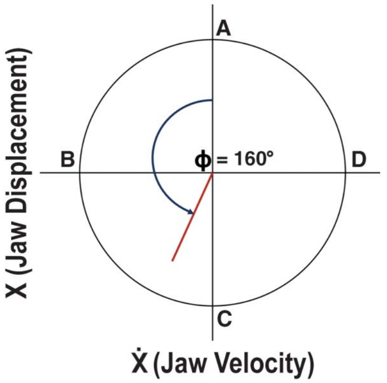
To adequately characterize interarticulator coordination, it is first necessary to determine the nature of the phonetic targets the speech production mechanism is attempting to achieve. Speech movements possess spatiotemporal characteristics: They are executed in articulatory space and evolve over time. Thus, the coordination of articulatory movements must involve an attempt by the speech motor controller to achieve a desired spatial structure ("where" to be in articulatory space) and temporal structure ("when" to be somewhere in articulatory space). The AP view posits that sets of articulators are functionally yolked into precisely timed coordination patterns to achieve phonetic targets (e.g., Goldstein et al., 2006; Kelso et al., 1986). Other theoretical approaches, such as Guenther's Directions Into Velocities of the Articulators model (Guenther, 1998, 2016), propose instead that it is the *acoustic signal* generated by constrictions of the vocal tract that is the primary target of speech production. As we are interested here in testing the conceptual merits of the AP view, we focus our theoretical discussion on AP and supporting evidence.

With regard to spatial targets, results obtained from experiments in which articulator trajectories are dynamically perturbed during ongoing speech production (e.g., Abbs & Gracco, 1984; Folkins & Abbs, 1975; Kelso et al., 1984; Tremblay et al., 2003) provide evidence that the *positions of the articulators* are the primary target (or control parameter) in speech production. To take one example, Abbs and Gracco (1984) reported upper lip compensation for a downward lower lip perturbation during the bilabial constriction for the initial consonant in /bɑ/. In another example, Kelso et al. (1984) applied downward perturbations of the jaw during constrictions for the final consonant in two utterances, /bæb/ and /bæz/. These authors reported compensatory adjustments that were specific to the utterance being produced: lip compensation was found for the utterance ending in the bilabial stop /b/, whereas TT compensation was found for the utterance ending in the alveolar consonant /z/. Such compensatory movements demonstrate that the intricate dynamics among the articulators are fluid enough to make live adjustments in concert to achieve a spatial goal.

Yet, at the same time, there is other evidence that suggests that *the time at which an articulator begins moving toward a constriction site* in the vocal tract is a critical control parameter during speech production. The evidence relevant to this claim comes from a series of articulatory investigations by Tuller and colleagues (e.g., Harris et al., 1986; Kelso et al., 1984; Tuller & Kelso, 1984; Tuller et al., 1982, 1983) that found the relative timing of independent articulators to be remarkably stable across certain scalar changes during speaking. To take one example, Tuller and Kelso (1984) examined the orchestration of jaw and upper lip movements over time during the production of /ba#Cab/ utterances, where C was voiceless /p/ or voiced /b/ or /w/ (i.e., /babab/, /bapab/, /bawab/). Jaw and upper lip movements were tracked using flesh point motion tracking signals (i.e., light-emitting diodes) across manipulations in production rate (fast vs. slow) and stress pattern (first syllable stressed vs. unstressed). Such manipulations engendered variation in jaw and lip movement amplitude, duration, and velocity. To examine the effects of these scalar changes on the coordination between these two articulators, the authors measured the time between onsets of jaw lowering for successive vowels, and the time between onsets of jaw lowering at which the upper lip began to descend for the medial bilabial constriction (affiliated with /b/, /p/, or /w/) across those manipulations. These two temporal intervals—the jaw vowel cycle (henceforth, JVC) and upper lip onset latency—were found to be highly correlated across production rate and stress pattern in all the talkers tested. More precisely, any manipulation that resulted in a shorter JVC (e.g., fast production rate or unstressed syllable) also tended to serve to reduce the latency of upper lip movement onset. Thus, although the *absolute* timing between the jaw and upper lip varied considerably across these scalar changes, their *relative* timing appeared to be lawful and systematic, independent of changes in the absolute timing of movement. That is, timing variation in the JVC was accompanied by proportional changes in the timing of the upper lip. On the basis of these findings, Tuller and Kelso (1984) hypothesized that the period between successive jaw lowering movements for vowels serves as a fundamental unit of articulatory organization with the planning and execution of consonantal gestures timed relative to such periods. One idea borne out of this account is the existence of scaled interarticulator timing patterns that serve as the basis of phonetic structure across different scales (i.e., temporal targets in speech production).

However, Kelso et al. (1986) subsequently argued that relative timing analyses provided inadequate descriptions of interarticulator relations because they failed to incorporate information about the *full spatiotemporal trajectories* of pairs of articulators: They do not include the entire trajectory of one articulator once it has initiated movement and is instead based solely on onset of movement of the other articulator. To further complicate matters, it was argued that JVCs and consonant latencies may not always be linearly related, as a matter of logic. For example, two hypothetical utterances having the same consonant latencies could nevertheless have different JVC durations. Indeed, Tuller and Kelso's (1984) findings were reanalyzed and reinterpreted by Kelso et al. (1986) following a set of detailed kinematic analyses showing constant *spatiotemporal phasing* relations between the upper lip and jaw across the same scalar changes. To quantify interarticulator phase relations, Kelso et al. (1986) represented the movement of the jaw on a phase plane (with displacement on one axis and velocity on the other), as schematized in Figure 1, so that the onset of movement of the upper lip could be given as an angle on that phase plane. In Figure 1, downward jaw movements would be displayed as downward movements of the phase path (going from A to C). The vertical crosshair indicates zero velocity, and the horizontal crosshair indicates midway between minimum and maximum jaw displacement. As the jaw moves

Figure 1. Interarticulator spatial phasing between the tongue tip and the jaw. Onset of tongue tip movement toward an intervocalic consonant represented as an angle (Φ) on the jaw position-velocity phase plane. Jaw displacement is given by the *y*-axis, and velocity of jaw movement is given by the *x*-axis. Thus, the movement of the jaw is depicted as moving from its highest initial point (A) to its lowest point (C) and back again. Maximum velocity values during jaw lowering and raising are given by (B) and (D), respectively. The red line indicates the onset of tongue tip movement within the phase plane, adapted from Kelso and Tuller (1987) and Nittrouer (1991). In this tutorial example, Φ is 160° (see main text for further explanation).



from its highest point (A) to its lowest point (C), velocity increases to a local maximum (B) then decreases to zero (C) when the jaw changes direction of movement. The onset of movement of the upper lip was then represented as an angle (Φ) on this JVC phase plane. In the schematic given in Figure 1, the onset of movement of the upper lip begins at a phase angle of 160° (in the first half cycle affiliated with jaw lowering). Using this approach, Kelso et al. (1986) reanalyzed the jaw-upper lip data discussed earlier (Tuller & Kelso, 1984) and found that the interarticulator phase angle was constant across manipulations in speech production rate and stress pattern. Kelso and colleagues interpreted this consistency in phase angle for upper lip movement onset as indicating that it is the primary factor explaining phonetic stability across variability in timing across articulatory structures. These authors further argued, as a matter of logic, that phase angle was a superior description of interarticulator relations because two hypothetical utterances could have the same JVC durations and consonant latencies and nevertheless have different phase angles, if the vowel-related jaw lowering and raising is asymmetrical (i.e., the jaw lowers faster than it raises, or vice-versa).[1] Thus, in this view (Kelso et al., 1986), descriptions of interarticulator relations based solely on time are thought to be inappropriate, because it is assumed that talkers attempt to control both how the articulators are coordinated in space (e.g., where and how far to move in space) and time (e.g., when to move and for how long) to achieve phonetic targets.

In a subsequent experiment, Nittrouer (1991) used Kelso et al.'s (1986) phase angle approach with articulatory data obtained from the XRMB system to quantify phase relations between the jaw and TT in /tV#Cɑt/ utterances across manipulations in phonetic structure (where V was /ɛ/ or /ɑ/; C was /t/ or /d/), production rate, and stress pattern. Nittrouer (1991) reported that the onset of TT movement toward the intervocalic alveolar consonant, given as an angle on the phase plane, varied with the duration of the JVC: Any manipulation that shortened the JVC also reduced the phase angle at which the TT began movement. A second finding of the XRMB study by Nittrouer (1991) showed that the within-condition JVC duration variability did not correlate with the within-condition variability in phase angle (see Nittrouer et al., 1988, for similar findings involving coordination of the jaw and upper lip). Rather, within a given utterance, TT movement was found to initiate at the same angle within the JVC phase plane with only slight, random variability, indicating utterance-specific targets. Critically, however, Nittrouer (1991) did not quantify onset of TT movement as a temporal event, and no study since has directly compared interarticulator stability using relative timing versus phase angle measures. Clearly more research is needed to fully explicate the essential control parameters that underlie the generation of multi-articulator movement patterns.

The Current Research

As discussed above, previous efforts to understand coordinative structures for speech focused on identifying the essential control parameters responsible for allowing articulators to achieve desired goals, with some research focusing on temporal parameters (Tuller & Kelso, 1984) and other research focusing on spatiotemporal parameters (Kelso et al., 1986; Nittrouer, 1991; Nittrouer et al., 1988). The purpose of the current research was to replicate and extend the XRMB study by Nittrouer (1991) using newer motion capture technology to quantify and compare both the temporal *and* spatiotemporal relations of jaw and TT movements—across manipulations in segmental and suprasegmental structure—in detail. We tested a broader sample of talkers whose productions were assessed using state-of-the-art EMA technology (Perkell et al., 1992; Rebernik et al., 2021). This study also extended and solidified the methodology used in Nittrouer (1991) to include a refined two-step, dual coder process for reliably identifying and extracting key gestural landmarks in EMA data needed for interarticulator coordination analyses. As in Nittrouer's (1991) study, the segmental structure, rate of production, and stress pattern were

[1] Although, it is worth noting that the extent to which vowel-related jaw lowering and raising is symmetrical has not been directly examined in this type of speaking task.

manipulated to test whether the spatiotemporal relations between the jaw and TT systematically vary with changes in JVC durations. Specifically, we quantified the latencies and phase angles within the JVC at which the TT begins to raise for the medial consonant in /tV#Cɑt/ sequences. We reasoned that, if timing variation in the jaw is accompanied by proportional changes in the timing of the TT (Tuller & Kelso, 1984), then that would indicate that talkers impart phonetic structure to the acoustic signal by coordinating the *timing* between articulators. In this scenario, the TT should begin moving at the same proportion of the way through the JVC across scalar changes in JVC duration. If, however, the phase angles of TT movement onset are highly stable across variations in JVC durations (Kelso et al., 1986), then that would indicate that talkers are more precise in *spatiotemporally* coordinating the articulators. An alternative, but not mutually exclusive, possibility is that the spatial and/or temporal relations among articulators may uniquely vary for each utterance. If this is the case, then TT latencies or phase angles will be strongly related to each speaking condition but not within-condition variability (Nittrouer, 1991).

Method

All experiments complied with the principles of research involving human subjects as stipulated by the University of Florida.

Participants

Ten adult speakers (eight women, aged 19–23 years, M_{age} = 20.4 years) served as participants for this experiment. These participants were graduate or undergraduate students at the University of Florida. All were native speakers of American English, and none reported a history of a speech, language, hearing, or other neurological disorder.

Speech Stimuli

As detailed above, Nittrouer (1991) reported that the onset of TT movement began relatively sooner on the jaw phase plane for utterances with shorter JVCs. We used the same set of utterances chosen for that study in order to engender comparable variation in JVC durations. Specifically, the target utterances for this study consisted of nonsense, disyllabic /tV#Cɑt/ sequences where (a) V was short /ɛ/ or long /ɑ/; (b) C was voiceless /t/ or voiced /d/; (c) the first syllable was stressed or unstressed; and (d) production rate was fast or normal. Variation in JVC durations were also obtained by using the inherently long vowel (/ɑ/) and shorter for the inherently short vowel (/ɛ/),[2] and by using the voiced and voiceless versions of the alveolar stop in the medial consonant position (/d/ and /t/), as prior acoustic analyses reported shorter vocalic segments preceded voiceless compared to voiced stops (e.g., Klatt, 1976). Production rate and stress pattern were also manipulated to provide variability in JVC durations without affecting the segmental composition of the utterances.

To control the pre- and post-vocal-tract configurations, all stimuli were embedded in the carrier phase, “It’s a ____ again.” The target utterances were cued audiovisually by a model speaker to the subjects on a color computer monitor (24-in. curved screen) with full-screen video.[3] To create the test stimuli, we made a digital audio–video (AV) recording of a female speaker of American English producing the target utterances with each stress pattern and at each production rate. The model talker was audio- and video-recorded in a soundproof booth with bright lighting and a plain blue background. We asked the talker to produce clear and distinct speech while emphasizing either the first or second syllable. The camera was centered on the talker’s face and was framed above the top of her head to just below her larynx. The video was recorded using a digital camcorder (Sony Exmor R). The video stream was digitized at the standard frame rate (30 images per second) and the audio signal was digitized at a frequency of 44100 Hz. The sound was played through an 8-in. speaker (Yamaha HS5) mounted to the right of the monitor. Subjects were seated approximately 2 ft from the monitor.

Instrumentation, Experimental Design, and Procedure

Jaw and tongue movements were recorded using a Carstens AG501 EMA system (Carstens Medizinelektronik GmbH) in a quiet room. Speech movement tracking is performed with EMA using weak and diffuse magnetic fields, generated by a series of transmitter coils, to localize the positions of small sensor coils temporarily fixated on the surfaces of the articulators during the act of speaking (Hoole, 1996; Perkell et al., 1992; Rebernik et al., 2021). The transmitter coils, positioned above the head of the speaker at different orientations, generate and radiate out a set of magnetic fields, which induce a current in the sensor coils. The magnetic fields generated by the transmitters oscillate at fixed kHz range radio frequencies, which permits their superimposition and subsequent detection in the composite signal induced in each sensor. The intensity of the current depends on the distance and orientation of the

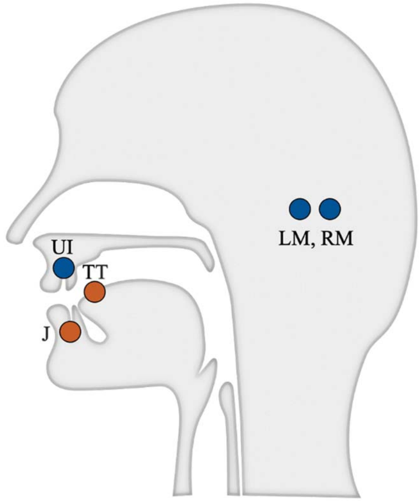
[2] /ɑ/ is also inherently lower than /ɛ/, which was expected to engender variation in extent of jaw displacement.

[3] Note that Nittrouer (1991) cued the target utterances to subjects using orthographic (i.e., text-only) stimuli.

sensor coils from the transmitter coils. This voltage-distance relation, combining the separate signals induced by each transmitter, is then used to compute the locations and orientations of the sensors and, therefore, the articulatory surfaces to which they are attached, in near real time in 3D. Proprietary Carstens software is used to convert from voltages to Cartesian coordinates.

As shown in Figure 2, articulatory sensors were placed on the TT and the jaw with static, head reference sensors placed on the gingiva above the upper incisor (UI) and behind each ear on the left mastoid (LM) and right mastoid (RM) processes. The UI-reference sensor was placed intraorally on the gingiva of the UIs using a piece of Stomahesive wafer. The LM- and RM-reference sensors were firmly placed using medical tape. Prior to the speaking task, each subject's occlusal plane was obtained by attaching three sensors to the Carstens biteplane that subjects held between their upper and lower teeth while data were recorded for these three sensors, as well as the reference sensors. After the occlusal plane was established, sensors were attached to the midsagittal surface of the jaw and TT. The jaw-movement sensor was placed intraorally on the gingiva of the lower incisors using a piece of Stomahesive wafer. The TT-movement sensor was placed 1 cm behind the anatomical TT using a nontoxic dental glue (EPIGLU, Meyer-Haake). Participants were engaged in spontaneous conversation for approximately 10 min with an experimenter prior to the start of the test session to allow some time to habituate to talking with the sensors (see Dromey et al., 2018; Weismer & Bunton, 1999).

Figure 2. Midsagittal schematic view of the vocal tract with the locations of the electromagnetic articulography sensors used in the current experiment. Orange dots mark dynamic, articulatory movement sensors; blue dots mark static, head reference sensors. LM = left mastoid; RM = right mastoid; UI = upper incisor; J = jaw; TT = tongue tip.



Participants sat in front of a computer monitor while positioned under the transducer coils of the EMA system. Each trial consisted of the following sequence of events. First, the video model of a given utterance was presented on screen. Subjects only saw and heard each video model once. Then, after the offset of the video, participants repeated the target utterance. The utterances were randomized across trials. Subjects were instructed to repeat each utterance as seen and heard in the video model at the appropriate (self-determined) production rate (normal or fast), while making sure to produce the correct vowel and medial consonant and emphasis the appropriate syllable. Before beginning the experiment, participants were exposed to multiple practice trials with experimenter feedback to confirm that they understood the instructions and were able to perform the task.

Subjects were recorded while producing 15 repetitions of each utterance, resulting in 240 tokens collected per subject: 2 Vs × 2 Cs × 2 rates × 2 stress patterns × 15 repetitions. The trials were blocked by speaking rate: normal first and then fast. V, C, and stress pattern were varied randomly within blocks. PsychoPy software (Version 3.0; Peirce, 2007) was used to sequence the experiment and display the AV speech stimuli. Simultaneous audio (sampled at 48 kHz) was recorded using a shotgun microphone (t.bone EM9600) and the EMA sensor signals (sampled at 250 Hz) were recorded using Carstens' CS5RECORDER and CS5VIEW programs.

Data Processing and Analysis

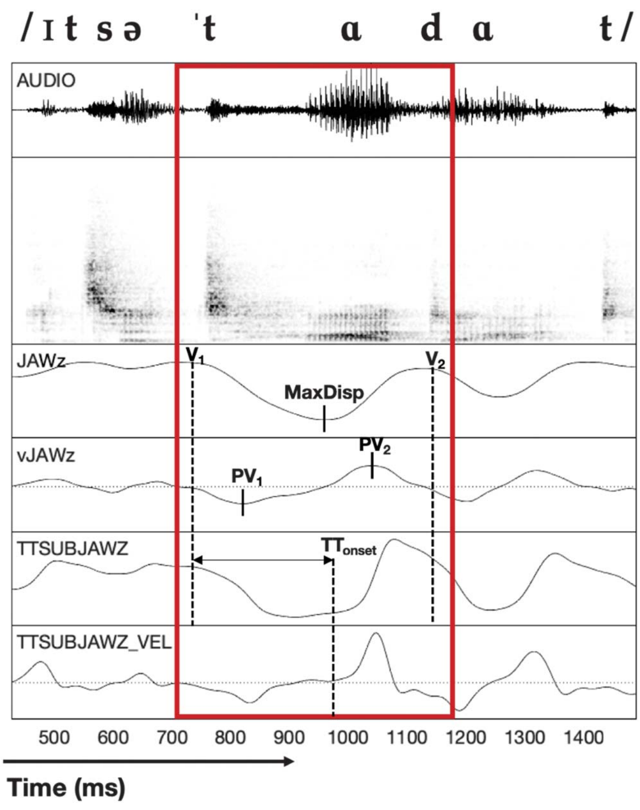
The raw acoustic (.wav) and kinematic (.pos) data were processed, visualized, and analyzed using the MATLAB-based *Mview* algorithms (Tiede, 2005, 2010). The raw kinematic data first underwent a series of standardized preprocessing steps to rotate and translate each position signal to a consistent maxillary frame of reference (based on the location of UI-reference sensor), and to correct for head motion artifacts (on the basis of the three static, head reference sensors). The acoustic and kinematic signals were then synchronized and visualized together in the same analysis space.

Each token was auditorily evaluated for accuracy by a phonetically trained coder. Tokens were discarded if stress was inappropriately placed, the incorrect vowel was clearly substituted in the first syllable, or incorrect consonant voicing was clearly used. Tokens were not discarded if the production rate was uncertain, a schwa was used in the first syllable when it was unstressed, or the medial consonant was flapped, because those realizations of articulation represent how speakers actually talk. To remove them would be removing natural variability. Instead, the goal was to remove only tokens in which the target was produced incorrectly.

For the retained tokens, phonetically trained coders manually inspected the processed articulatory displacement and velocity time series (in the vertical dimension relative to UI) obtained for the jaw- and TT-sensor locations. Figure 3 shows the acoustic signal and displacement-time and velocity-time functions for the jaw- and TT-sensors for one token of /tɑ’dɑt/, produced with stress on the first syllable. Since the tongue rides on top of the jaw, jaw displacement was subtracted from TT displacement to isolate lingual movements that occurred independently of jaw-raising/lowering movements. As illustrated in Figure 3, these time series were used to identify the following six temporal landmarks in each token:

1. JVC onset for first vowel in /tVCat/.
2. Peak velocity during jaw lowering.
3. Maximum displacement of the jaw.
4. Peak velocity during jaw raising.
5. Onset of TT raising.
6. JVC onset for second vowel in /tVCat/.

Figure 3. Observed trajectories for the vertical position of four articulator measures presented in tandem with the resulting acoustic speech signal (waveform and spectrogram) as displayed in the Mview graphical user interface. The four measures are (from top to bottom): Jaw displacement, jaw velocity, tongue tip (TT) displacement, and TT velocity. The utterance is /tɑ’dɑt/. The temporal kinematic landmarks identified by coders are overlaid: V_1 = first JVC onset; V_2 = second JVC onset; MaxDisp = maximum jaw displacement; PV_1 = peak velocity during jaw lowering; PV_2 = peak velocity during jaw raising; TT_{onset} = onset of tongue tip raising within JVC.



These landmarks were labeled for all tokens in *Mview* using a two-step, dual coder process. For each token, two coders first identified the location of the first JVC onset. The standard procedure for determining JVC onset in prior articulatory investigations using optotrack or XRMB data (Nittrouer, 1991; Nittrouer et al., 1988; Tuller & Kelso, 1984) was to identify the zero crossing(s) associated with lowering (or raising) movements in the jaw-sensor velocity record. While these procedures worked in most cases, there were some instances where patterns of articulation made it unreasonable to use the zero crossing criterion. Sometimes the talker would appear to gradually initiate jaw lowering, before achieving a more robust pattern of lowering. When this occurred for stressed first syllables, the onset of jaw lowering was defined as the point at which the velocity of the jaw-sensor lowering passed 1 cm/s (this was the same criterion used by Nittrouer, 1991). For unstressed syllables, however, that criterion was often inappropriate because talkers might not lower their jaws very far or very rapidly for that unstressed vowel. In these cases, the onset of jaw lowering was defined as the point at which the jaw-sensor displacement reached a value of 0.5 mm below maximum height (i.e., height at closure prior to jaw lowering). Finally, for unstressed syllables, some talkers would not raise the jaw after lowering for the first vowel. Instead, they might lower the jaw minimally for that unstressed vowel, and then proceed to lower the jaw further for the second vowel, allowing the TT to achieve contact with the alveolar ridge through its own raising. In these cases, the onset of jaw lowering for the second vowel was defined as the point at which the velocity of the jaw-sensor lowering passed 1 cm/s. Using these procedures, two coders performed this initial step of labelling the first JVC onset independently, and then a custom MATLAB script compared the values and flagged any that differed by more than 5 ms.

Those flagged were reviewed by both coders so that they arrived at a consistent and reliable JVC onset. Following the JVC onset quality check, the same two coders then independently identified and labeled the five subsequent landmarks. The JVC onset for the second syllable was identified using the same procedures to identify the JVC onset (described above). The onset of TT raising was identified as the point on the TT-sensor displacement record within the JVC at which an upward movement began. This point was generally determined by the zero crossing in the TT velocity record, but if velocity remained close to zero for a period of time, then it was marked at the point at which the velocity reached a criterion value of 1 cm/s. The maximum jaw displacement was identified as the jaw-sensor's maximum vertical displacement (below the UI) within the JVC. In some cases, the jaw would reach a maximum degree of vertical displacement and then plateau. In these cases, we placed the maximum jaw displacement landmark at the temporal midpoint of the plateau. The peak velocities during jaw lowering and raising were identified as maximum values (cm/s) in the jaw velocity record associated with the lowering or raising movements within the JVC.

All labels were laid down in *Mview* for all tokens (i.e., repetitions) of a given utterance before moving onto another utterance. A custom MATLAB script then compared the labels for the two coders for all tokens of a single talker and flagged any cases where the labels were more than 10% different. Those discrepant labels were reviewed by both coders to arrive at a consistent and reliable set of labels for the token.

For each token, these six articulatory landmarks were then used to produce measures of JVC durations, maximum jaw displacements, TT latencies, and TT phase angle (as defined in Kelso et al., 1986, pp. 44–45, footnote 5). Following Nittrouer (1991), jaw movement for the JVC was represented on a phase plane, with both displacement and velocity normalized to the interval of −1 to +1 (as schematized in Figure 1). Normalization of displacement was computed over the entire cycle, whereas normalization of velocity was computed only for the half cycle in which TT raising began. The onset of TT raising was then given as an angle (Φ) on this phase plane.

Results

Correlations Between JVC Durations and TT Latencies and Phase Angles

Our first set of analyses focused on examining correlations between JVC durations and TT movement onset latencies, and between JVC durations and TT phase angles.[4] Figures 4 and 5 show scatterplots (with calculated linear regression lines) relating TT latencies and JVC durations, and TT phase angles and JVC durations, respectively, for each talker and each utterance. Each point in both figures represents one token of an utterance type. A Pearson correlation coefficient was calculated for each talker's distribution (pooled across all utterances). Each correlation is based on articulatory measurements extracted from approximately 240 utterances. The computed correlations, given in the first column of Tables 1 and 2, respectively, were all highly significant ($p < .001$), but were consistently higher for TT latencies than TT phase angles for all 10 talkers, indicating that the *relative timing* of TT movement onset was more systematic across scalar changes than its *spatiotemporal position with the JVC*. A Fisher's r-to-z transformation was performed to confirm that the difference in size of the correlations between TT latencies and JVC durations and the correlations between TT phase angles and JVC durations (pooled across all 10 talkers) were significant ($z = 2.302$, $p = .011$).

Main Effect of Utterance Type on TT Latencies and Phase Angles

Our second set of analyses focused on examining the effects of utterance on TT latencies and phase angles. We performed separate analyses of variance (ANOVAs) with utterance type (Utterance 1–16) as a within-subjects factor on TT latencies and TT phase angles.[5] The results of these analyses are given in Table 3. The eta-squared (η_p^2) effect size values index how strongly interarticulator coordination between the jaw and TT is associated with the manipulations in segmental (vowel, medial consonant) and supra-segmental structure (production rate, stress pattern). These values were consistently higher for TT latencies than TT phase angles for all 10 talkers, further indicating that the *temporal* characteristics of jaw and TT movements were more strongly affiliated with utterance identity than were the *spatiotemporal phasing* characteristics.

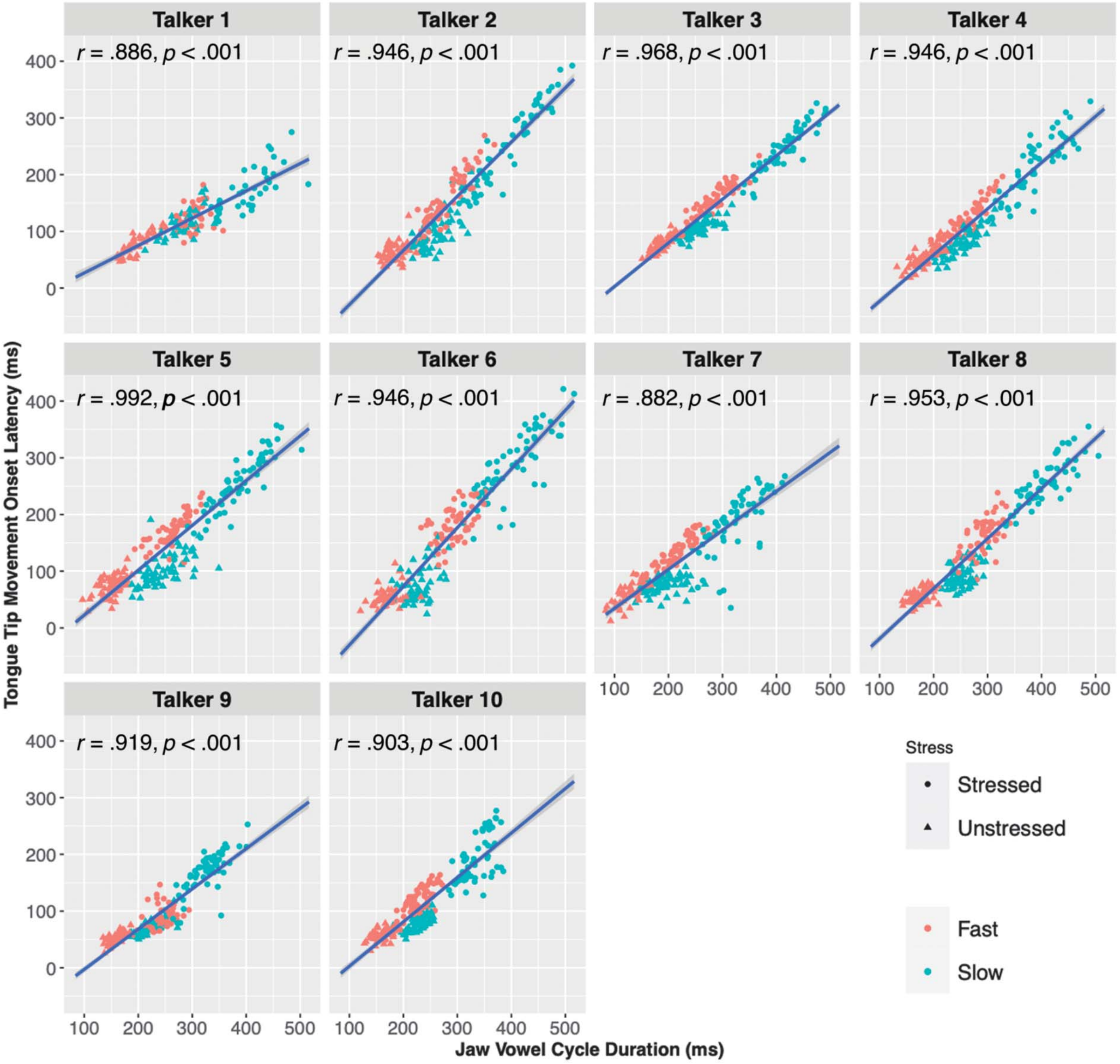
Main Effect of Utterance or Continuous Scaling

Our third set of analyses addressed whether the main effect of utterance type demonstrated in the aforementioned ANOVA analyses (given in Table 3) represented *discrete* changes in both JVC durations and TT latencies/phase angles as a function of the segmental and suprasegmental manipulations (Nittrouer, 1991), or whether they more closely represented a *continuous scaling* across these measures (Tuller & Kelso, 1984). The first column in Tables 1 and 2 displays Pearson correlation coefficients relating TT latencies and JVC durations, and TT phase angles and JVC durations, respectively, for each talker. Recall that the results of these correlation analyses indicated that both TT latencies and phase angles generally increased with JVC durations (although TT latencies were found to scale more strongly across changes in JVC durations), but it is not clear from these analyses alone if this across-token

[4] See the supplemental materials for a breakdown of the token means of JVC durations, TT movement onset latencies, and TT phase angles for each talker and each utterance and rate/stress condition (SS = slow/stressed; SU = slow/unstressed; FS = fast/stressed; FU = fast/unstressed).

[5] See the supplemental materials for details on main effects of production rate, stress pattern, vowel, and medial consonant on JVC durations, maximum jaw displacements, TT latencies, and TT phase angles for each talker.

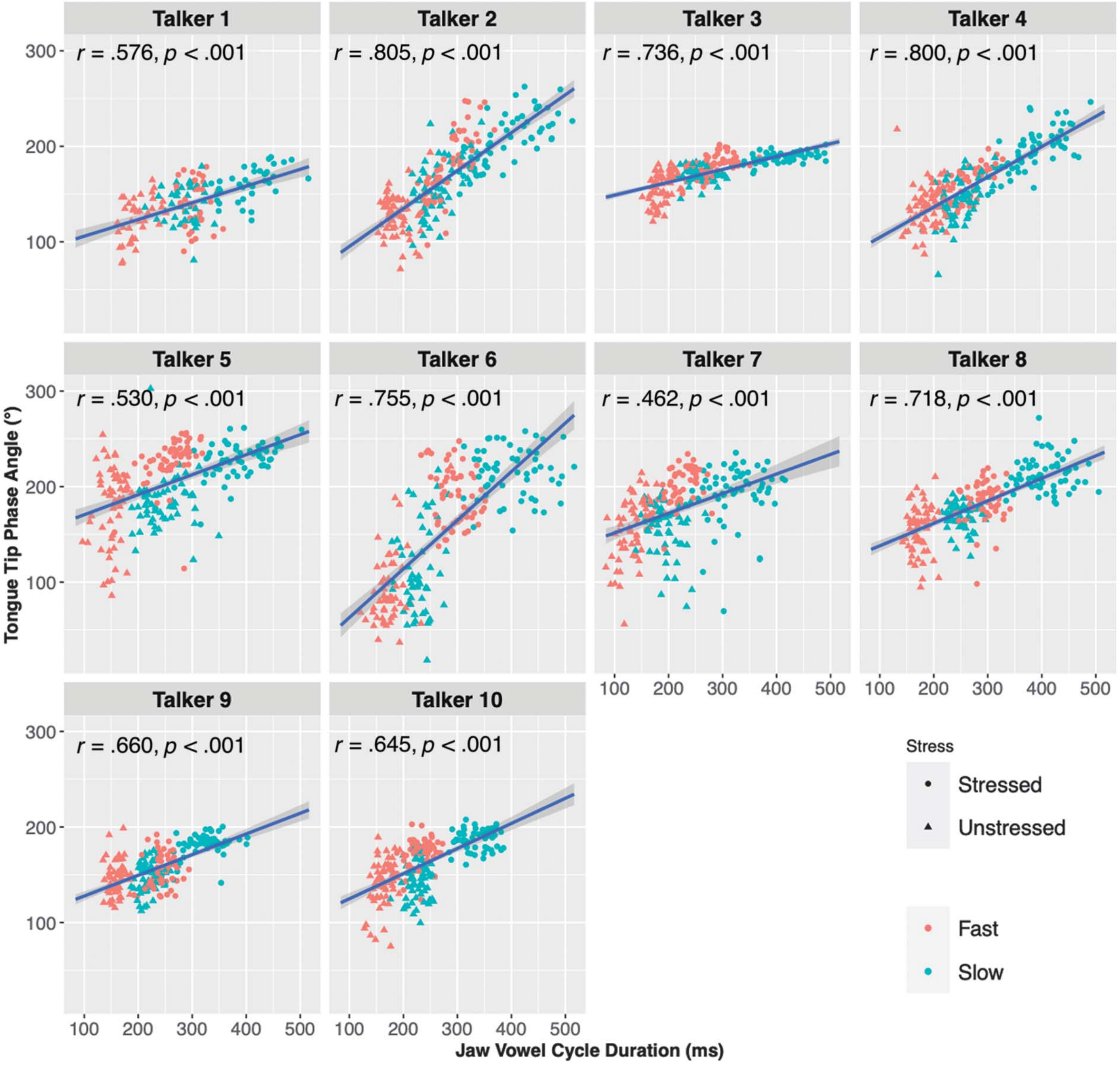
Figure 4. Scatterplots (with calculated regression lines) relating tongue tip movement onset latencies (ms) and jaw vowel cycle durations (ms) for each talker and each rate/stress condition (cyan-colored points = slow production rate, coral-colored points = fast production rate; circular points = stressed, triangular points = unstressed). Each point represents one token of an utterance type.



relation is due to a continuous scaling or to the main effect of utterance type.

A distinction between these two possibilities can be made only by comparing correlation coefficients computed for tokens within each condition (i.e., the same utterance and rate/stress pattern) and those computed for the condition means. If the relation is one of continuous scaling, then the within-condition as well as condition-mean correlation coefficients should be roughly as high as those computed across all individual tokens (even though the within-condition correlation coefficients are based on a narrower range of values than the condition-mean correlation coefficients). That is, both within-condition and across-condition relations between JVC durations and TT latencies/phase angles would have contributed equally to the coefficients obtained for all individual tokens together. If, however, the relation between JVC durations and TT latencies/phase angles is one of a main effect of utterance type, then the within-condition coefficients should be lower and the condition-mean coefficients should be higher than those obtained for individual tokens. Such a finding would indicate that the between-condition relation between JVC durations and TT latencies/phase angles was completely responsible for the significant overall coefficients, whereas

Figure 5. Scatterplots (with calculated regression lines) relating tongue tip phase angles (°) and jaw vowel cycle durations (ms) for each talker and each rate/stress condition (cyan-colored points = slow production rate, coral-colored points = fast production rate; circular points = stressed, triangular points = unstressed).



the within-conditions correlations (or lack of) were actually attenuating these values.

The second column of Tables 1 and 2 displays the mean of the 16 within-condition correlation coefficients for each talker. These values were all lower than the coefficients for individual tokens for both TT latencies and phase angles, and not all of them were statistically significant. The third column of Tables 1 and 2 displays the correlation coefficients computed between condition means for each talker. These values were all higher than the coefficients for individual tokens for both TT latencies and phase angles and were statistically significant. Taken together, these findings demonstrate that the relation between JVC durations and TT latencies/phase angles is more accurately accounted for by the main effect of utterance type, rather than continuous scaling.

Discussion

The present results replicate and extend Nittrouer's (1991) observations on the temporal coordination between the jaw and TT during speech production, using the XRMB system. Specifically, we investigated the extent to

Table 1. Pearson correlation coefficients (r) between jaw vowel cycle (JVC) durations (ms) and TT movement onset latencies (ms) for each talker.

r computed over:	All individual tokens	Tokens within a condition (*M* for 16 conditions)	JVC and TT latency condition means
Talker 1	.89	.56	.97
Talker 2	.95	.64	.97
Talker 3	.97	.83	.98
Talker 4	.95	.85	.96
Talker 5	.92	.70	.95
Talker 6	.95	.47	.97
Talker 7	.88	.66	.93
Talker 8	.95	.59	.97
Talker 9	.92	.68	.95
Talker 10	.90	.65	.92
M	.93	.66	.96
SD	.03	.11	.02
SE	.01	.03	.00
COV	.03	.17	.02

Note. *M* = mean; *SD* = standard deviation; *SE* = standard error; COV = coefficient of variation.

which temporal parameters (relative timing of articulatory movements) versus spatiotemporal parameters (phase angle of movement onset for one articulator, relative to another) govern the coordination of the articulators during the production of VCV utterances. Ten talkers recorded nonsense, disyllabic /tV#Cat/ sequences using EMA, with alternative V (/ɑ/−/ɛ) and C (/t/−/d/), across variation in rate (fast–slow) and stress pattern (first syllable stressed–unstressed). Two dependent measures were obtained: (a) timing of TT raising onset for the medial C, relative to jaw opening-closing; and (b) angle of tongue-tip raising onset, relative to the jaw phase plane. To summarize, the kinematic results showed that any manipulation that shortened the jaw opening–closing cycle reduced both the relative timing and phase angle of the tongue-tip movement onset, but relative timing of tongue-tip movement onset scaled more consistently with J opening–closing across variation in production rate and stress pattern than the phase angle. In addition, these relations between the jaw and TT were not found to be continuous in nature, but rather a main effect of utterance type (see also Nittrouer, 1991; Nittrouer et al., 1988). That is to say, the relative timing and phase angle of movement onset for the TT, relative to the jaw, was precisely organized and coordinated for each unique utterance produced. These kinematic results have important theoretical implications, which center on the nature of the control parameters that govern the creation of constrictions by various vocal tract structures across segmental and suprasegmental manipulations during speech production.

Table 2. Pearson correlation coefficients (*r*) between jaw vowel cycles (JVC) durations (ms) and TT phase angles (PhA; °) for each talker.

r computed over:	All individual tokens	Tokens within a condition (*M* for 16 conditions)	JVC and TT PhA condition means
Talker 1	.58	.14	.89
Talker 2	.81	.28	.93
Talker 3	.74	.22	.90
Talker 4	.80	.43	.95
Talker 5	.53	.26	.71
Talker 6	.76	.02	.91
Talker 7	.46	.14	.70
Talker 8	.72	.05	.96
Talker 9	.66	.21	.87
Talker 10	.65	.24	.81
M	.67	.20	.86
SD	.12	.12	.09
SE	.03	.03	.02
COV	.17	.60	.11

Note. *M* = mean; *SD* = standard deviation; *SE* = standard error; COV = coefficient of variation.

Table 3. Analysis of variance results for tongue tip (TT) movement onset latencies and TT phase angles.

Talker	Effect	*F*	*df*	*p*	η_p^2
Dependent measure: Tongue tip movement onset latency					
1	Utterance	28.804	15	< .001	0.757
2	Utterance	113.458	15	< .001	0.887
3	Utterance	189.502	15	< .001	0.928
4	Utterance	76.778	15	< .001	0.840
5	Utterance	89.371	15	< .001	0.860
6	Utterance	156.290	15	< .001	0.918
7	Utterance	84.084	15	< .001	0.854
8	Utterance	203.274	15	< .001	0.932
9	Utterance	105.119	15	< .001	0.876
10	Utterance	186.385	15	< .001	0.926
Dependent measure: Tongue tip phase angle					
1	Utterance	7.245	15	< .001	0.439
2	Utterance	35.778	15	< .001	0.712
3	Utterance	27.066	15	< .001	0.649
4	Utterance	23.951	15	< .001	0.620
5	Utterance	14.890	15	< .001	0.506
6	Utterance	36.150	15	< .001	0.722
7	Utterance	8.284	15	< .001	0.365
8	Utterance	20.632	15	< .001	0.580
9	Utterance	17.300	15	< .001	0.537
10	Utterance	22.543	15	< .001	0.604

Note. Shown are the *F* value, the degrees of freedom (*df*), the *p* value, and the η_p^2 value for each effect and for each talker.

The observation that changes in TT latencies were more strongly associated with changes in JVC durations than were TT phase angles suggests that the relative timing at which articulators move is the critical control parameter that governs interarticulator movement patterns associated with phonetic structure. As such, the temporal aspects of coordination between the jaw and TT were more consistently related, and adjustments in the timing of each individual articulator were organized to occur in a relationally consistent manner. We hypothesize that the phase angles are likely not the true source of the relation between the jaw and TT, but rather a consequence of the systematic inter-articulator timing observed. That is to say, it is the precise timing among the articulators that serves to ensure the spatial goal of attaining a vocal-tract constriction of a specified degree of closure at a specific location. Evidence supporting this account was provided by the findings from dynamic articulator perturbation experiments reviewed earlier (e.g., Abbs & Gracco, 1984; Folkins & Abbs, 1975; Gracco, 1988; Kelso et al., 1984). Recall that the compensatory trade-offs documented in those experiments reveal that coordinative structures synergistically adjust the actions of all vocal-tract structures involved following the perturbation to attain a spatial goal while maintaining movement timing relations (e.g., speeding up a labial closure gesture when the jaw is perturbed downward). The present experiment provides further evidence that the articulators operate flexibly in real time to achieve spatial targets "on time." Analyses of the relative timing of the jaw and tongue (given in Figure 4) clearly demonstrated that these vocal tract structures are interdependently modulated such that timing variation in one structure is accompanied by proportional changes in the timing of the other active structure. Such flexibility and coordination is needed to deal with kinematic changes that occur as talkers are subjected to artificial perturbations in the laboratory (e.g., Abbs & Gracco, 1984; Folkins & Abbs, 1975; Kelso et al., 1984) or "natural" perturbations that occur during ongoing speech because of contextual variations (Nittrouer, 1991; Nittrouer et al., 1988; Tuller & Kelso, 1984).

In summary, the results from the current experiment provide evidence that *interarticulator timing is of the essence* to revealing phonetic structure, which is in keeping with the AP framework (Browman & Goldstein, 1992; Fowler, 1980). Although robust timing relations have been observed in other speech motor actions (Gracco, 1988; Gracco & Lofqvist, 1994; Tuller & Kelso, 1984) and limb movements (Kelso, 1986), further replication with other types of utterances involving other sets of articulators is still needed. It is unclear how general such coordinative interactions are among different vocal tract structures. Ongoing articulographic analyses are testing whether jaw movements are coupled in their timing to those of the lips and larynx, as well as the tongue, during the production of speech. If a general motor planning operation exists, then there should be high intraclass correlations between patterns of motor coordination across different pairs of articulators (e.g., tongue–jaw, lip–jaw, and larynx–jaw synergies). Such findings would suggest the existence of an intrinsic timing mechanism (or "central clock") that controls the timing of movements between articulators, and that other parameters control spatial details of the actions. Future research studies on speech production and perception designed to develop further the AP framework should also assess, for individual talkers, the strength of the relationship between JVC durations and TT latencies and sensitivity to phonological structure. We hypothesize (Masapollo & Nittrouer, 2021) that the degree to which individual talkers differentiate their speech motor patterns across across segmental and suprasegmental conditions will predict how keen their phonological representations are, because it is the precise interarticulator timing that appears to be imparting phonetic structure to the acoustic signal of speech.

In future studies, improved kinematic analysis procedures that are currently being developed will also improve our ability to measure and explore other metrics that may be critical to more fully explicating the dynamical control and stability of interarticular coordination. Examining other kinematic landmarks—such as the point at which an articulator achieves closure, the total distance travelled by that articulator on route to that closure, the speed of closure, and the duration of closure, all of which occur within the period of the JVC—will facilitate improved understanding of the nature of coordinative structures.

The current findings also offer new opportunities for more in-depth research into the breakdown of coordinative structures for speech. Deficits in articulatory timing have been reported across a wide range of clinical populations, including in prelingually and postlingually deaf talkers (Lane & Perkell, 2005), stuttering talkers (e.g., Alm, 2004; Masapollo et al., 2021; Max et al., 2003), talkers with apraxia of speech (e.g., Ziegler & von Cramon, 1986), and talkers with reading disorders, such as dyslexia (Lalain et al., 2003). Improved characterization of the ability to precisely time and coordinate multi-articulator movements could lead to improved diagnostic tools and mechanistically driven intervention techniques focused on *motor timing*. For example, degrees of interarticulator coordination deficits may differentiate between mild and severe forms of stuttering. In addition, successful management of intricate coordination of the timing between articulatory gestures could increase the ability to acquire and produce phonological contrasts, leading, therefore, to improved intelligibility among deaf talkers.

Although much remains to be learned about how multiple interleaving articulatory trajectories structure the acoustic speech signal, the existing data strongly suggest that a dynamic control regime governs speech movement patterns, and that phonetic structure emerges in large measure from a coordinative strategy aimed at controlling the relative timing of movements. The existence of temporal targets in speech production speaks to the strong dynamical nature of the speech production mechanism.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, Matthew Masapollo, upon reasonable request.

Acknowledgments

The research reported in this article was supported by a Research Opportunity Seed Fund award from the University of Florida (PI: Masapollo). Thanks to Jessica Goel, Kara Payne, Kiana Ferdowsiepour, Julia Ginter, and Kayleigh Burge, who assisted with EMA data collection and analysis. Mark Tiede (Haskins Laboratories), Donal Sinex, Jessica Goel, and Nicholas Salazar were instrumental in writing scripts for data processing, visualization, and analysis. Finally, this work benefited from helpful discussions with, or comments from Mark Tiede and audience members of the 181st Meeting of the Acoustical Society of America (held in Seattle, WA).

References

Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology, 51*(4), 705–723. https://doi.org/10.1152/jn.1984.51.4.705

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: A critical review of possible relations. *Journal of Communication Disorders, 37*(4), 325–369. https://doi.org/10.1016/j.jcomdis.2004.03.001

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*(3–4), 155–180. https://doi.org/10.1159/000261913

Dromey, C., Hunter, E., & Nissen, S. L. (2018). Speech adaptation to kinematic recording sensors: Perceptual and acoustic findings. *Journal of Speech, Language, and Hearing Research, 61*(3), 593–603. https://doi.org/10.1044/2017_JSLHR-S-17-0169

Folkins, J. W., & Abbs, J. H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 18*(1), 207–220. https://doi.org/10.1044/jshr.1801.207

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8*(1), 113–133. https://doi.org/10.1016/S0095-4470(19)31446-9

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review, 13*(3), 361–377. https://doi.org/10.3758/BF03193857

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. A. Arbib (Ed.), *Action to language via the Mirror neuron system* (pp. 215–249). Cambridge University Press. https://doi.org/10.1017/CBO9780511541599.008

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. S.

Meyer (Eds.), *Phonetics and phonology in language comprehension and production* (pp. 159–207). de Gruyter. https://doi.org/10.1515/9783110895094.159

Gracco, V. L. (1988). Timing factors in the coordination of speech movements. *Journal of Neuroscience, 8*(12), 4628–4639. https://doi.org/10.1523/JNEUROSCI.08-12-04628.1988

Gracco, V. L., & Lofqvist, A. (1994). Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. *Journal of Neuroscience, 14*(11), 6585–6597. https://doi.org/10.1523/JNEUROSCI.14-11-06585.1994

Guenther, F. H. (2016). *Neural control of speech*. MIT Press. https://doi.org/10.7551/mitpress/10471.001.0001

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105*(4), 611–633. https://doi.org/10.1037/0033-295X.105.4.611-633

Harris, K. S., Tuller, B., & Kelso, J. A. S. (1986). Temporal invariance in the production of speech. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 243–267). Erlbaum.

Herbst, C. T. (2020). Electroglottography - An update. *Journal of Voice, 34*(4), 503–526. https://doi.org/10.1016/j.jvoice.2018.12.014

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science, 17*(1), 42–46. https://doi.org/10.1111/j.1467-8721.2008.00545.x

Hoole, P. (1996). Issues in the acquisition, processing, reduction, and parameterization of articulographic data. *Instituts für Phonetik und Sprachliche Kommunikation, München (FIPKM), 34,* 158–173.

Kelso, J. A. S. (1986). Pattern formation in speech and limb movements involving many degrees of freedom. *Experimental Brain Research,* 105–128.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14*(1), 29–59. https://doi.org/10.1016/S0095-4470(19)30608-4

Kelso, J. A. S., & Tuller, B. (1987). Intrinsic timing in speech production: Theory, methodology, and preliminary observations. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 203–222). Earlbaum.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*(6), 812–832. https://doi.org/10.1037/0096-1523.10.6.812

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America, 59*(5), 1208–1221. https://doi.org/10.1121/1.380986

Lalain, M., Joly-Pottuz, N., Nguyen, N., & Habib, M. (2003). Dyslexia: The articulatory hypothesis revisited. *Brain and Cognition, 53*(2), 253–256. https://doi.org/10.1016/S0278-2626(03)00121-0

Lane, H., & Perkell, J. S. (2005). Control of voice-onset time in the absence of hearing: A review. *Journal of Speech, Language, and Hearing Research, 48*(6), 1334–1343. https://doi.org/10.1044/1092-4388(2005/093)

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Masapollo, M., & Nittrouer, S. (2021). MIPA: A theory of phonological acquisition and speech motor control. *The Journal of the Acoustical Society of America, 150*(4), A111. https://doi.org/10.1121/10.0007799

Masapollo, M., Segawa, J. A., Beal, D., Tourville, J., Nieto-Castañón, A., Heyne, M., Frankford, S., & Guenther, F. H. (2021). Behavioral and neural correlates of speech motor sequence learning in stuttering and neurotypical speakers: An fMRI investigation. *Neurobiology of Language, 2*(1), 106–137. https://doi.org/10.1162/nol_a_00027

Max, L., Caruso, A. J., & Gracco, V. L. (2003). Kinematic analyses of speech, orofacial nonspeech, and finger movements in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research, 46*(1), 215–232. https://doi.org/10.1044/1092-4388(2003/017)

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y. C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., & Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America, 136*(3), 1307–1311. https://doi.org/10.1121/1.4890284

Nittrouer, S. (1991). Phase relations of jaw and tongue tip movements in the production of VCV utterances. *The Journal of the Acoustical Society of America, 90*(4), 1806–1815. https://doi.org/10.1121/1.401661

Nittrouer, S., Munhall, K., Kelso, A. S., Tuller, B., & Harris, K. S. (1988). Patterns of inter-articulator phasing and their relation to linguistic structure. *The Journal of the Acoustical Society of America, 84*(5), 1653–1661. https://doi.org/10.1121/1.397180

Peirce, J. W. (2007). PsychoPy - psychophysics software in python. *Journal of Neuroscience Methods, 162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America, 92*(6), 3078–3096. https://doi.org/10.1121/1.404204

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology, 12*(1), 6. https://doi.org/10.5334/labphon.237

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Smith, A., Goffman, L., Zelaznik, H., Ying, S., & McGillem, C. (1995). Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research, 104*(3), 493–501. https://doi.org/10.1007/BF00231983

Sorensen, T., Toutios, A., Goldstein, L., & Narayanan, S. S. (2016). Characterizing vocal tract dynamics across speakers using real-time MRI. *Interspeech,* 465–469. https://doi.org/10.21437/Interspeech.2016-583

Stevens, K. N. (1998). *Acoustic phonetics*. MIT Press.

Tiede, M. (2005). *MVIEW: Software for visualization and analysis of concurrently recorded movement data*. Haskins Laboratory.

Tiede, M. (2010). *MVIEW: Multi-channel visualization application for displaying dynamic sensor movement*. Haskins Laboratory.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*(6942), 866–869. https://doi.org/10.1038/nature01710

Tuller, B., & Kelso, J. A. (1984). The timing of articulatory gestures: Evidence for relational invariants. *The Journal of the Acoustical Society America, 76*(4), 1030–1036. https://doi.org/10.1121/1.391421

Tuller, B., Kelso, J. A., & Harris, K. S. (1982). Inter-articulator phasing as an index of temporal regularity in speech. *Journal* *of Experimental Psychology: Human Perception and Performance, 8,* 460–472. https://doi.org/10.1037/0096-1523.8.3.460

Tuller, B., Kelso, J. A., & Harris, K. S. (1983). Converging evidence for the role of relative timing in speech. *Journal of Experimental Psychology: Human Perception and Performance, 9*(5), 829–833. https://doi.org/10.1037/0096-1523.9.5.829

Turvey, M. T., Shaw, R., & Mace, W. (1978). Issues in the theory of action: Degrees of freedom, coordinative structures and coalitions. In J. Requin (Ed.), *Attention and performance* (Vol. 7). Erlbaum.

Weismer, G., & Bunton, K. (1999). Influences of pellet markers on speech production behavior: Acoustical and perceptual measures. *The Journal of the Acoustical Society of America, 105*(5), 2882–2894. https://doi.org/10.1121/1.426902

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. University of Wisconsin Press.

Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research, 48*(3), 543–553. https://doi.org/10.1044/1092-4388(2005/037)

Ziegler, W., & von Cramon, D. (1986). Timing deficits in apraxia of speech. *European Archives of Psychiatry and Neurological Sciences, 236*(1), 44–49. https://doi.org/10.1007/BF00641058