Laboratory Phonology
Journal of the Association for Laboratory Phonology

Franich, K. 2022. How we speak when we speak to a beat: The influence of temporal coupling on phonetic enhancement. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 13(1), pp. 1–51. DOI: https://doi.org/10.16995/labphon.6452

Open Library of Humanities

How we speak when we speak to a beat: The influence of temporal coupling on phonetic enhancement

Kathryn Franich, Harvard University, Cambridge, MA, USA, kfranich@fas.harvard.edu

Stressed syllables in languages which have them tend to show two interesting properties: They show patterns of phonetic 'enhancement' at the articulatory and acoustic levels, and they also show coordinative properties. They typically play a key role in coordinating speech with co-speech gesture, in coordination with a musical beat, and in other sensorimotor synchronization tasks such as speech-coordinated beat tapping and metronome timing. While various phonological theories have considered stress from both of these perspectives, there is as yet no clear explanation as to how these properties relate to one another. The present work tests the hypothesis that aspects of phonetic enhancement may in fact be driven by coordination itself by observing how phonetic patterns produced by speakers of two prosodically-distinct languages—English and Medʉmba (Grassfields Bantu)—vary as a function of timing relations with an imaginary metronome beat. Results indicate that production of syllables in time (versus on the 'offbeat') with the imaginary beat led to increased duration and first formant frequency—two widely observed correlates of syllable stress—for speakers of both languages. These results support the idea that some patterns of phonetic enhancement may have their roots in coordinative practices.

1. Introduction

Word stress has been characterized in the vast majority of linguistic literature in terms of the phonetic properties it associates with, in particular various forms of acoustic or articulatory 'enhancement' found on stressed syllables as compared with unstressed syllables (Edwards & Beckman, 1988; Beckman, Edwards, & Fletcher, 1992; Cho, 2005; Fujimura, 1990; Ladefoged, 1967; Sluijter, van Heuven, & Pacilly, 1997). Stressed syllables across languages are found to be produced, for example, with longer duration, increased jaw lowering, more extreme fundamental frequency, and greater intensity (e.g., de Jong & Zawaydeh, 1999; Fry, 1955, 1958; Gordon, 2004; Kleber & Klipphahn, 2006; Hualde, Lujanbio, & Torreira, 2008; Lieberman, 1960; Lindblom, 1963; Sluijter & Van Heuven, 1996; Vogel, Athanasopoulou, & Pincus, 2016; see also Gordon & Roettger, 2017). A less-explored but equally intriguing property of stress-based languages is the fact that stressed syllables play a key role in the coordination of speech, as well as between speech and other systems. For example, stressed syllables (or a subset of them which also carry phrase-level prominence) serve as the locus of coordination in many languages with co-speech gestures of the hands and head: In several languages, including English, Brazilian Portuguese, and Catalan, the 'apex' (point of maximal excursion) of a co-speech gesture is consistently found to be temporally anchored to a syllable bearing stress or phrase-level pitch accent (Esteve-Gibert, Borràs-Comes, Asor, Swerts, & Prieto, 2017; Kendon, 1980; Loehr, 2012; Leonard & Cummins, 2011; Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008). Stressed syllables also play an important role in musical text-setting, or the mapping of speech to musical rhythms. Specifically, stressed syllables are found to map consistently to musically-strong beats in several languages (Dell & Halle, 2009; Lerdahl & Jackendoff, 1983; Morgan & Janda, 1989; Temperley & Temperley, 2012), though this mapping constraint is more stringent in some languages than others. Stressed syllables also tend to show privileged status for coordination in speech-motor tasks such as rhythmic hand-tapping to speech (Allen, 1972; Rathcke, Lin, Falk, & Dalla Bella, 2021) and for alignment with an external stimulus such as a metronome (Cummins, 1997; Cummins & Port, 1998; Tajima, 1998; Tajima & Port, 2003). Within speech itself, stressed syllables are found to constrain articulatory movements; for example, timing of the velum lowering gesture for word-internal intervocalic nasals is found to be 'attracted' to syllable nuclei in stressed syllables, as opposed to unstressed ones (Byrd, Tobin, Bresch, & Narayanan, 2009; Krakow, 1993).

Despite the parallels between enhancement[1] patterns of stress and its coordinative properties across languages, little work has attempted to understand the nature of this link. Within feature-based theories of phonology, a dominant perspective about metrical prominence and phonetic properties has revolved around the role of stress in speech perception: Metrically-or accentually-

[1] Throughout the paper, the term 'enhancement' is used to refer to the distinctive phonetic properties of stressed/metrically-prominent syllables; however, we use this term to refer both to active enhancement of a metrically-prominent syllable or to reduction of non-prominent syllables with the effect of making prominent ones more phonetically distinctive.

prominent positions are considered to be phonologically 'privileged' in the grammar, and a set of rules or constraints can be applied to enforce the production of such syllables so that they are maximally perceptually salient or distinct from surrounding non-prominent syllables, either through requiring privileged positions to contain perceptually-salient phonetic patterns, or through the avoidance of reduction/neutralization of contrasts in privileged positions (Beckman, 1997; Crosswhite, 2001; Smith, 2002, 2004). The privileged status of stressed/accented syllables is traditionally seen to derive from the status of these syllables as metrical heads (e.g. Liberman & Prince, 1977; Hayes, 1995; Beckman, 1996; Ladd, 1996), though some recent work within Autosegmental-Metrical/ToBI theory, in particular, has focused more on the role of these syllables as phonetic prominence-bearers—namely, bearers of different varieties of pitch accents which link to information-structural functions (Baumann & Röhr, 2015; Cole, Mo, & Hasegawa-Johnson, 2010; Gussenhoven, 2021). Either way, coordination of speech with other systems can be conceptualized within these theories as an *alignment* of events which share some aspect of prominence. A shared notion of prominence across systems is more straightforward in some cases than others: In the case of text-setting, elements across domains which share similar phonetic prominence profiles in terms of e.g., pitch, duration, and loudness can be clearly mapped to one another (Gussenhoven, 2021). In the case of co-speech gesture, however, the motivation for alignment is less clear, since the notion of prominence at the level non-speech gestures has not been well-defined. Definitions of perceptual prominence based on height and direction of pitch movement (e.g., Baumann & Röhr, 2015), for example, do not seem to map straightforwardly to gestures of the hands, arms, and head. Furthermore, much of this work has focused on the notion of phonetic prominence from the perspective of non-tonal languages, making it unclear how constraints on coordination might be regulated for languages which do not show the same pitch-based correlates of prominence. It has been suggested, alternatively, that kinematic similarities in speech and co-speech gesture profiles may provide a more direct link (Krivokapić, Tiede, & Tyrone, 2017; Shattuck-Hufnagel & Ren, 2018). Regardless of how we define prominence across these different domains, the link between phonetic enhancement and coordination within these theories is indirect: Prominence arises due to abstract grammatical properties, and prominences across systems or modalities (e.g., speech and music, or speech and co-speech gesture) are aligned during communication through similarly abstract rules or constraints.

Among researchers working from an articulatory perspective, investigation into the role of stress has largely focused on either its coordinative role *or* its effects on the spatial position of articulators and movement duration. In terms of intergestural coordination, stress is found to influence articulatory coordination patterns (Byrd et al., 2009), as well as the degree of variability in intergestural timing (Tilsen, 2009). Stress also has an impact on the coordination of other prosodic events such as boundary tones with vowel timing (Katsika, Krivokapić, Mooshammer, Tiede, & Goldstein, 2014). Recent work has shown that F0 peaks in pitch-accented syllables are coordinated in time with the apex of finger pointing co-speech gestures (Esteve-Gibert

& Prieto, 2013; Krivokapic, Tiede, Tyrone, & Goldenberg, 2016), as well as other types of manual gestures (Kendon, 1980; Loehr, 2012; Leonard & Cummins, 2011; Rochet-Capellan et al., 2008). In the spatial domain, several studies have shown greater articulatory displacement at metrically-strong positions (Beckman et al., 1992; Cho, 2005; Erickson & Kawahara, 2016; Keating, Lindblom, Lubker, & Kreiman, 1994; Van Summers, 1987); this effect has been found to be largest in English for low vowels, as opposed to high vowels (Harrington & Palethorpe, 1996). Maximum displacement of articulators is known to vary as a function of speech style and rate—generally, slower speech rate is linked with greater jaw displacement (Linville, 1982; Sonoda, 1987; Mefferd, 2017), possibly as a result of a general link between slowed speech rate and hyperarticulation (Lindblom, 1990). Articulatory displacement of the tongue has also been shown to scale with peak velocity of articulator movement (Kent & Moll, 1972; Kuehn & Moll, 1976; Ostry & Munhall, 1985)—reflecting the level of *gestural stiffness* in articulation—though this relationship has been shown to be individual- and speech rate-dependent (Gay & Hirose, 1973; McClean & Tasko, 2003). Other factors such as vowel tenseness are also known to play a role in stress-related articulatory enhancement/reduction effects (Mooshammer & Fuchs, 2002).

While these studies demonstrate clear effects of stress on many aspects of temporal and spatial patterning in the articulatory domain, only recently have researchers begun to try to account for these types of enhancement effects from a grammatical standpoint. The aspect of stress-related enhancement which has received the most attention from this perspective is increased syllable duration, which has most recently been treated within the framework of Articulatory Phonology through the application of various types of 'clock slowing' gestures which, based on a coupled oscillator model, serve to temporally modulate the oscillatory timing of speech gestures over which they are activated. For example, Saltzman, Nam, Krivokapić, and Goldstein (2008) model durational asymmetries between stressed and unstressed syllables in stress-timed languages using the coupled-oscillator account developed by O'Dell and Nieminen (1999), in which syllable- and foot-level oscillators can be asymmetrically coupled to one another to produce foot-internal duration reduction, with the addition of a temporal modulation gesture (the μ_T-*gesture*) which is activated during the stressed syllable only, and which leads to oscillator slowing during that portion of the stress foot (see also Byrd & Saltzman, 2003). While the approach does not directly explain all reported enhancement effects linked to stress, some other effects, such as increased jaw lowering, can be predicted to fall out from clock slowing due to the fact that more time is afforded to the jaw to reach peak displacement (Byrd & Saltzman, 2003). The general nature of the internal clock can also be used to account for results on co-speech gesture which show that manual gestures show increased duration when timed to occur with stressed syllables or prosodic phrase boundaries (Krivokapić et al., 2017; Parrell, Goldstein, Lee, & Byrd, 2014; see also Rusiewicz, Shaiman, Iverson, & Szuminsky, 2013). Within this account, durational enhancement

effects are achieved through the application of the temporal modulation gesture; without it, no durational differences between stressed and unstressed syllables are predicted to emerge. In sum, while coordination and coupling relations form a central part of the theory of Articulatory Phonology, their link to phonetic enhancement processes such as durational lengthening of stressed syllables is somewhat indirect.

1.1. Coordination patterns as a potential source of phonetic enhancement

An alternative to the view that phonetic enhancement is driven exclusively by rule or by activation of a clock-slowing gesture is that some or all aspects of phonetic enhancement may be intrinsically related to coordinative properties themselves. The idea that coordination patterns can drive changes in movement stems from observations by von Holst (1973) of oscillatory movements of fish pectoral fins, based on which he hypothesized that absolute synchrony of coordination between movements of the two fins is associated with increased movement amplitude—a condition which he referred to as *superimposition*. Schwartz, Amazeen, and Turvey (1995) tested this hypothesis among humans by examining the effects of coordination on patterns of limb movements. In the experiment, the researchers asked participants to oscillate hand-held pendulums in three different coupling modes, including an uncoupled mode with just a single pendulum being manipulated with one arm, a coupled mode with the two pendulums operating in-phase (at a 0° angle) with the two arms, and a coupled mode in which the two pendulums were operated anti-phase (at a 180° angle) with the two arms. Amplitude of pendulum swings was found to be greatest in the in-phase coupled position, and movements were also found to be less temporally-variable in that condition.

Schwartz et al. (1995) demonstrate how the conditions favoring superimposition—namely, those associated with increased movement stability—can be interpreted within a dynamical model of intersegmental coordination elaborated in Kelso (1994) and Schöner (1994). To describe coordinated movement between the two arms, for example, we can define coordination dynamics by the velocity vector field of a collective variable with relative phase $\phi = (\theta_i - \theta_j)$, in which the two θs represent the phase angles of the left and right arms. The first order differential equation that characterizes the evolution of the collective variable is:

$$\dot{\phi} = \Delta\omega - a\sin(\phi) - 2b\sin(2\phi) + \sqrt{Q}\zeta_t \tag{1}$$

where the overdot represents the derivative, or rate of change, in ϕ. For in-phase (1:1) frequency-locked behavior, a solution to equation (1) is the stable state of ϕ given the current coordination parameters (Haken, Kelso, & Bunz, 1985; Kelso, DelColle, & Schöner, 1990; Schöner, Haken,

& Kelso, 1986). The ratio of b/a in the sine functions defines the control parameter, in this case limb movement frequency, which will influence the strength of the stable states of ϕ. The term $\Delta\omega$ represents the difference between the preferred movement frequencies of the two arms—essentially, it represents 'competition' between the two arms, which was manipulated by Schwartz et al. by differing the eigenfrequencies of the two manual pendulums used in their arm-swinging experiment. Where $\Delta\omega = 0$ and $b/a > .25$, the two stable states of ϕ will be at or near $\phi = 0^{\circ}$ (in-phase coordination) and at or near $\phi = 180^{\circ}$ (antiphase coordination) (Haken et al., 1985). The stable state of $\phi = 0^{\circ}$, termed the 'global attractor,' can be shown to be overall more stable than $\phi = 180^{\circ}$; many experiments on human motor control have confirmed this (Fuchs & Kelso, 2018; Kelso, Southard, & Goodman, 1979; Kelso, 1984; Schmidt, Carello, & Turvey, 1990). Where $\Delta\omega > 0$ and b/a decreases, limbs become increasingly 'detuned,' meaning the relative phase of the limbs will shift away from the canonical stable states of $\phi = 0^{\circ}$ and $\phi = 180^{\circ}$. The term $Q\,\zeta_t$ represents a Gaussian white noise process ζ_t with a strength of $Q > 0$.

Given all of this, von Holst's hypothesis boils down to the idea that superimposition is favored where movement is most stable, namely when the two limbs are frequency-locked and coupled at $\phi = 0^{\circ}$, the global attractor. That movement amplitude should be maximized under these conditions was not initially reflected in the equation in (1); however, Kudo, Park, Kay, and Turvey (2006) demonstrate how amplitude can be incorporated as an additional variable within the model such that the degree of shift away from the stable states of ϕ (which is minimized at the global attractor) is positively related to the magnitude of $|\lambda|$, where $1/\lambda$ is the time it takes for the arms to relax to the attractor phase position following perturbation. Movement amplitude can be shown to be directly related to λ. Their analysis thus posits a direct relationship between movement stability and amplitude (see also de Poel, Roerdink, Peper, Lieke, & Beek, 2020 for a recent overview and discussion of the relationship between amplitude and stability in interlimb coordination).

Importantly, additional work has found that the effects of movement synchronization on movement stability and amplitude extend beyond coordination within the individual to coordination with an external stimulus, such as a metronome. Generally speaking, it has been found that movements display less temporal and spatial variability when they are coordinated with a metronome beat than when they are performed without coordinating to an external stimulus (Byblow, Carson, & Goodman, 1994; Carson, 1990); this phenomenon is known as *anchoring*. Moreover, stability of movement is found to be further increased where multiple points of anchoring are present. For example, when oscillating the fingers or the limbs in a back-and-forth motion with a metronome beat, stability is increased where the points of peak movement amplitude in both the forward and backward directions are timed to occur with a metronome beat, as compared with conditions where coupling with the beat only takes

place in one direction of movement (Fink, Kelso, Jirsa, & de Guzman, 2000; Jirsa, Fink, Foo, & Kelso, 2000; Kudo et al., 2006). Under these conditions of enhanced stability introduced by an external stimulus, it has also been found that movements are performed with greater amplitude, with similar associations between stability and amplitude established for bimanual finger wagging (Fink et al., 2000), forearm movement (Kudo et al., 2006; Pellecchia, Shockley, & Turvey, 2005; Peper, de Boer, de Poel, & Beek, 2008) and circle drawing (Ryu & Buchanan, 2004). In sum, the effects of superimposition on movement stability are similar regardless of whether an individual is coordinating their own movements internally or with an external stimulus.

1.1.1 Coordination in speech and analogues to limb movement amplitude

Speech, like limb movement, is a highly complex coordinative act typically involving controlled expulsion of air from the lungs with simultaneous laryngeal adjustments to regulate vocal fold tension, coordinated with overlapping movements of the intraoral articulators such as the jaw, lips, and tongue to create syllables. Coordination of speech with other systems is also highly ubiquitous in daily use—even when interlocutors cannot see one another, as when speaking on the phone, they still coordinate their speech with co-speech gestures (Wei, 2006). Blind speakers have also been found to show language-specific use of co-speech gesture which is similar to that of sighted speakers (Özçalışkan, Lucero, & Goldin-Meadow, 2016), suggesting that co-speech gesture is not learned through visual cues, but rather reflects a coordinated element which is acquired naturally through the act of speaking. Add to this the fact that speech is often being coordinated in other ways, such as to music or within a conversation with another speaker, and there are myriad opportunities for coordination to influence speech.

There are a number of dimensions on which we might compare changes in limb amplitude to analogous changes in speech articulation and acoustics. For example, we might expect more extreme displacement of the oral articulators during speech, which could serve to influence vowel formant frequencies: For example, lower jaw position during vowel production can lead to higher F1 values (Erickson, 2002; Harrington, Fletcher, & Beckman, 2000; Lindblom & Sundberg, 1971). As mentioned previously, increased jaw lowering may also lead to longer syllable duration. Erickson (2002) also shows that the tongue dorsum shows greater displacement in the front-back dimension during stressed versus unstressed syllables, an effect that we might expect to be enhanced based on coordination patterns. Another speech-related analog might be increased sound wave amplitude and intensity induced through greater subglottal pressure during sound production resulting from contraction of the intercostal muscles (Ladefoged & McKinney, 1963). Increased subglottal pressure could also lead to raised fundamental frequency (F0) due to increased vocal fold vibration rate.

The present work seeks to examine whether coupling of speech to a metronome may elicit some or all of the phonetic effects described above. Of course, many of these hypothetical effects are similar to the kinds of phonetic enhancement effects found to be associated with syllable stress cross-linguistically (see e.g., Gordon & Roettger, 2017). Therefore, if speech is found to change in these ways as a function of coupling, our results would provide evidence for a direct link between coordinative properties of stress and the observed acoustic/articulatory properties associated with stress.

1.1.2 Prior work on coupling and speech

A variety of studies have examined how coupling affects speech timing in terms of variability. For example, articulatory timing has been found to be less temporally variable when speech is coupled more strongly with a metronome beat (Tilsen, 2009). Metronome coupling has been found to be highly effective in inducing greater speech fluency in certain speech and language disorders which impact speech timing, including stuttering and dysarthria (Andrews et al., 2012; Mainka & Mallien, 2014). Speech spoken synchronously with another individual or group of individuals has also been found to be more temporally consistent in terms of syllable and pause durations (Cummins, 2002, 2009; Zvonik & Cummins, 2002). Aside from timing variability, however, there has been little work which explores how different acoustic or articulatory properties of speech are influenced under these types of coordination. One study by Parrell et al. (2014) found that finger taps produced concurrently with syllables were produced with greater movement amplitudes and usually longer durations when paired with stressed (as opposed to unstressed) syllables; since coupling of stressed syllables and manual gestures was found to be stronger/less variable than between unstressed syllables and gestures in the same study, this finding is consistent with the proposed link between coupling and movement amplitude. The authors account for their findings by proposing that coupling of speech and tapping combines the two tasks into a single coordinative structure, such that speech and tapping are mutually influenced by a single prosodic clock-slowing gesture which modulates duration and amplitude of movements across domains. Interestingly, though, the authors also found that smaller modulations in speech and tapping amplitude even on unstressed syllables were correlated across the two domains, suggesting that there may be a more general effect of coupling on amplitude and timing which is not the result of a prosodic gesture, per se. Since all conditions in the study involved synchronous speech and tapping, it's not clear how much coupling itself across the two modes may have affected speech or tapping dynamics. The present study aims to investigate more directly the influence of coupling on speech production in order to identify whether or not coupling itself can shape speech production.

1.2. The role of linguistic structure

An additional focus of the present study is on the degree to which linguistic structure may shape coupling effects on speech production. To that end, we investigate these effects on two languages with distinct prosodic structures: English (a dominant US variety), a stress-based language which also utilizes pitch accent to mark phrase-level prominence, and Medʉmba, a Grassfields Bantu language spoken in Cameroon, which is tonal and which does not show clear phonetic evidence for word stress in terms of the typical cues outlined in Section 1. Unlike in English, fundamental frequency and intensity do not play as large a role in prominence marking in Medʉmba; the dominant role of these cues is instead to signal contrasts in lexical and grammatical tone. While duration has been found to be an acoustic correlate of some types of phrase-level prominence in the language (Franich, 2019), these durational effects are quite small in comparison with languages that display clear evidence of both stress and phrase-level accent (e.g., as found by Prieto, Vanrell, Astruc, Payne, & Post, 2012). Medʉmba also shows a durational profile more consistent with 'syllable-timing'—where durations between successive syllables show relatively lower variability—in contrast with the 'stress-timed' variety of English examined here, in which durations between successive *stressed* syllables are more consistent than successive syllables (Abercrombie, 1967; Grabe & Low, 2002; Pike, 1945).

Despite a lack of clear stress cues in Medʉmba, the language patterns with other Grassfields Bantu languages and other Central and West African languages in exhibiting positional prominence effects, such that stem-initial syllables bear a greater number of consonantal and vocalic contrasts than do non-initial and non-stem syllables (see Hyman et al., 2019 and references therein). Franich (2021) and Franich and Lendja Ngnemzué (2021) show that the phonological patterning of stem-initial syllables in Medʉmba is consistent with their status as heads of metrical feet, and that these syllables show aspects of rhythmic behavior which are similar to English syllables bearing metrical stress, both in speech production and in some aspects of musical text-setting. Of particular interest in the present study is whether any effects of coupling on acoustic properties of speech can also be found in a language lacking typical cues to lexical stress, and whether distinctive aspects of the structures of Medʉmba and English—such as the use of lexical tone—may influence acoustic reflexes of coupling.

2. Method

2.1. Study design

The present study investigates the effects of coupling on speech using a metronome synchronization-continuation task. In the task, speakers repeat a word in time to a metronome for several beats, and then continue to repeat the word for several more repetitions once the beat has stopped, attempting to maintain the same pace and phasing, as if to a silent continuation of

the metronome. The motivation for using this type of a task, as opposed to e.g., a speech and tapping task, is that we can control for the possibility that changes in speech during coupling result from prosodic mechanisms alone (e.g., as proposed by Parrell et al., 2014). Since timing of the metronome is of course not controlled by the same system as timing of speech in our task, any effects of coupling must be explained through coordination alone. As described below, only the data from the continuation portion of the task is analyzed, allowing us to eliminate the possibility that speech changes may result simply from participants trying to speak over the metronome.

2.2. Stimuli

Stimuli consisted of 22 words for each language, containing a mixture of disyllabic and trisyllabic words. An additional 16 phrases were also included as fillers, to be analyzed for a separate experiment. English stimuli varied between the two most common stress patterns of SWW and WSW for trisyllabic words and SW and WS patterns for disyllabic words. Medʉmba stimuli varied between the three tone patterns found for trisyllabic words, HHL, LHL, and HLH, and the four tone patterns found for disyllabic words, HH, HL, LH, and LL. Words were also varied in terms of the vowels they contained in each position, though a fully balanced set of vowel qualities across conditions was not possible due to the limited inventory of polysyllabic words in Medʉmba and concerns about matching for segmental quality across stress positions in English. A full list of stimuli is given in Appendix A. Sample stimuli are provided in **Tables 1** and **2**. Note that Medʉmba has few non-compound native words longer than two syllables, so trisyllabic words are limited to English loanwords. Furthermore, due to the strong restrictions on vowel quality in non-initial non-compound native words, some prosodic words were incorporated among the Medʉmba stimuli which include a pronominal enclitic or which are likely derived from compounds. In order to explore potential differences in coupling-related speech changes on Medʉmba stem-initial syllables based on word position, words were also varied in terms of whether they were prefixed (such that stem-initial syllables occurred in non-initial position) or not (such that stem-initial syllables occurred in initial position). This manipulation was only possible for words bearing LH melodies due to limitations on possible tone patterns on prefixes and stems.

Stress Pattern	***Trisyllabic – SWW***	***Trisyllabic – WSW***	***Disyllabic – SW***	***Disyllabic – WS***
Words	bítterly cábinet	banána connéction	décade présent (noun)	decáy presént (verb)

Table 1: Sample of English Stimuli.

Tone Pattern	***Trisyllabic – HHL***	***Trisyllabic – LHL***	***Trisyllabic – HLH***	
Words	bítálì ‘bitter leaf’ máŋkə́lù ‘mango’	bànánà ‘banana’ tòmátù ‘tomato’	k͡xísəmít ‘Christmas’ tósɨ̀dé ‘Thursday’	
Tone Pattern (cont.)	***Disyllabic – HH***	***Disyllabic – HL***	***Disyllabic – LH (unprefixed)***	***Disyllabic LL***
Words (cont.)	kə́bə́ ‘cut’ jʉ́ní ‘see him’	bíbà ‘paper’ mɛ́nù ‘your child’	<u>Unprefixed Words</u> làbə́ ‘hit’ mìnú ‘cat (derogatory)’ <u>Prefixed Words</u> nə̀-bə́ ‘to be’ nə̀-nú ‘to drink’	gə̀ptə̀ ‘cut’ mɛ̀nɔ̀m ‘my person’

Table 2: Sample of Medʉmba stimuli.

2.3. Participants and procedure

Twelve native speakers of a northeastern US variety of English (7 identifying as female; mean age 35) and 12 native speakers of Medʉmba (8 identifying as female; mean age 42) were recruited for the study. Data collection with English speaking subjects took place at the University of Delaware while data collection with Medʉmba speaking subjects took place in a mobile laboratory in Bangangte, Cameroon. Medʉmba speakers were all from in and around the town of Bangangte. Participants completed a brief demographic survey which also included questions about possible hearing loss or speech disorders; no participants reported any problems with hearing or speech. On each trial, subjects were presented with a metronome beat played through an external Altec Lansing Series100 speaker. The metronome sound consisted of a synthetic drumbeat created in version 2.1.2 Audacity® (Audacity Team, 2018) recording and editing software, an open-source program for sound editing. The drumbeat had a 125 ms decay and a center frequency of 100 Hz. Width of the noise band was set to 600 Hz. Speakers wore a head-mounted Shure SM10A-CN dynamic cardioid microphone and were recorded on a Zoom H6n Pro digital audio recorder in .wav format at a sampling frequency of 48 kHz. Separate time-aligned channels were used to record the speaker’s voice and the metronome beat they were repeating to.

The beat was played a total of eight times per trial at two different speeds, with the slower speed consisting of a 1320 ms inter-stimulus interval (ISI) and the faster speed consisting of a 900 ms ISI. Participants completed four blocks of trials in total, with the first two blocks utilizing the slower metronome speed and the second two utilizing the faster speed. Participants were asked

to listen to the first four beats of the metronome and then to begin repeating with the metronome on the fifth beat, repeating the target word eight times total. This meant that participants would continue to repeat the word four times after the beat had stopped sounding; they were asked to continue repeating with the same timing to the imaginary beat as they had maintained when the beat was playing. Two phasing modes with the metronome were used in the task: For one slow and one fast block, participants were asked to align the metronome beat with the first syllable of the word (the 'onbeat' condition); for the other slow and fast blocks, participants were asked to repeat the word so that its first syllable occurred at about a third of the way through the metronome cycle, or at a 120 degree angle with the metronome beat (the 'offbeat' condition). Auditory examples were provided of each target phasing relation to familiarize the participants with the target metronome timing patterns (**Figure 1**), and participants were given several practice trials to get used to repeating in the different phasing relations at the start of each block. An experimenter was present for the duration of the experiment with each participant to guide them through the list of words. Attempts to correct participants' performance were limited: If participants failed to repeat a word on a particular trial in the appropriate phasing mode, however, they were asked to repeat the trial.

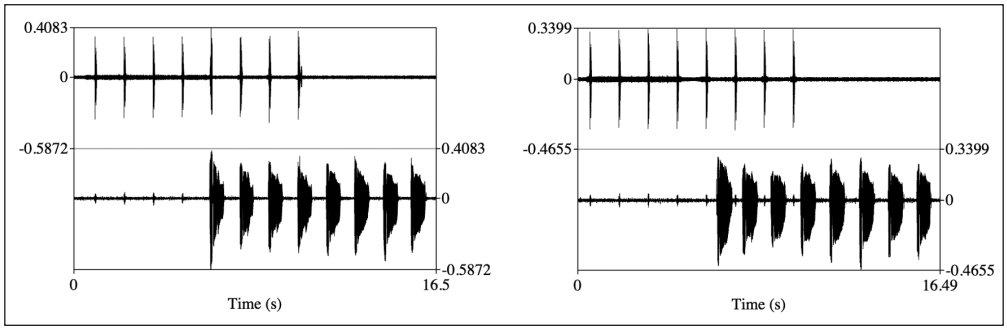

Figure 1: Examples of the word 'bitterly' uttered in the Onbeat (left) and Offbeat (right) phasing conditions with both metronome (top) and speech (bottom) channels shown.

The phasing mode of the first block was randomized by participant; they would then alternate between the two phasing modes for the slow metronome speed, continue with the mode they had last used as they began with the fast metronome speed, and then finish with the phasing mode they had begun with for the last fast metronome block.

In a follow-up session, participant were asked back to the lab to provide an additional set of repetitions of the same target words in an uncoordinated condition, without the metronome beat. In this condition, participants were asked to repeat the target words eight times at a comfortable pace. Due to COVID-19-related data collection interruptions during the English-speaking portion of the study, some subjects participated remotely in this last phase of the study, recording data on their home computers in Praat and sending it to the experimenter via email. These participants

were instructed to record using the same parameters as had been used for other participants who provided data in the lab. One participant was unavailable to provide data for the follow-up session.

2.4. Data processing

Audio recordings of participants' repetitions were segmented at the phone and word level for both languages using the FAVE-align forced aligner for English data (Rosenfelder et al., 2015) and via hand-segmentation in Praat for the Medʉmba data. Reliability between annotators for the Medʉmba data was achieved by having annotators segment a single file in which their phone boundary alignments were required to occur less than 3 ms from those in a sample file pre-annotated by a highly trained phonetician who was blind to the study goals. Annotators' alignments were also consistently checked for accuracy. Once datafiles were fully segmented, the following measures were extracted from vowels in both datasets using Praat scripts:

1. Vowel duration
2. Intensity at vowel midpoint
3. F0 at vowel midpoint
4. F1 at vowel midpoint
5. F2 at vowel midpoint

In addition, in order to evaluate alignment patterns with the metronome beat, Praat scripts were used to automatically mark metronome beats and extract their start times. Start times for silent beats in the continuation phase of the experiment were calculated by adding 1-4 ISI values to the start time of the final metronome beat. Vowels were extracted from both English and Medʉmba datasets for analysis of acoustic patterns and metronome alignment; vowels, as opposed to syllable onsets, were selected for alignment measures due to the fact that vowels approximate the location of *perceptual centers* (or 'p-centers'), the point in a syllable that speakers and listeners tend to intuit as the 'moment of occurrence' of the syllable, and the landmark which tends to align most consistently with the beat in metronome alignment studies for both English and Medʉmba (Franich, 2018b; Morton, Marcus, & Frankish, 1976; Scott, 1993). F0 was log-transformed. Intensity, log-transformed F0, and vowel formant values were z-scored by subject. Euclidean distance was also calculated based on Bark-transformed formant frequencies, and took the difference in the F1xF2 space from each vowel token to the center of the speaker's vowel space, calculated as the speaker's mean Bark-transformed F1 and F2 for all vowels (this approximated the average formant frequencies for schwa in either language). Outliers for any acoustic variable lying farther than 2.5 standard deviations from the mean were trimmed from the dataset; this resulted in a total reduction of less than 5% of the data for either language.

2.5 Statistical modeling

From the metronome-coordinated portion of the study, only data corresponding to the continuation portion of the task were analyzed; this was in order to avoid the possible confound of speakers trying to 'compete' with the metronome sound during the synchronization phase. Data were analyzed using a series of linear mixed effects models utilizing the *lmer* package for *R* statistical software (Bates, Mächler, Bolker, & Walker, 2015). Separate models were built for each language of interest. For both languages, dependent variables included Metronome Distance (distance, in ms, between the target vowel and the corresponding metronome beat), vowel Duration, F0, Intensity, F1 Frequency (a correlate of jaw height), F2 Frequency (a correlate of tongue backness), and Euclidean Distance. For English models, predictor variables included the factors PHASING (2 levels: Onbeat versus Offbeat), STRESS (2 levels: Stressed versus Unstressed), and WORD POSITION (3 levels: Initial versus Medial versus Final). The F1 model also included the factor VOWEL HEIGHT (3 levels: High versus Mid versus Low), and the F2 model included the factor VOWEL BACKNESS (3 levels: BACK, CENTRAL, FRONT); models also included interaction terms for all of these variables. All models except the Duration models also included VOWEL DURATION as a co-variate. Medʉmba models were identical to English models except that the factor TONE (2 levels: High versus Low) was substituted for STRESS. Finally, a subset of Medʉmba data is analyzed in Section 3.6 in which the position of metrically-strong syllables was manipulated to occur either word-initial or non-initial; dependent variables of Metronome Distance and Duration are examined as a function of this variable, PROMINENCE (2 levels: Initial and NonInitial) as well as PHASING and POSITION, and their interactions.

Since speech rate could not be controlled for in the uncoordinated speech condition, direct comparison of speech between the two metronome phasing conditions and the uncoordinated condition was only carried out for a subset of acoustic parameters (see Section 3.7). For this comparison, we incorporated an additional level to the PHASING variable, for three levels in this analysis: Onbeat, Offbeat, and NoBeat. An additional dependent measure of RELATIVE DURATION, or the ratio of the duration of stressed versus unstressed vowels in each word, was used for analyses of English.

All categorical predictors were sum-coded, while continuous predictors were mean-centered. By-subject random intercepts were included in all models, and initial models included random slopes for all predictor variables. The lme4 optimizer was set to 'bobyqa' with the maximum number of iterations set to 50,000. These maximal models were found to be singular (i.e., variances of one or more linear combinations of effects were near zero); therefore, following Barr, Levy, Scheepers, and Tily (2013), only those random slope terms whose absence eliminated singularity were removed. In most cases, this amounted to removing by-subject random slopes for STRESS/TONE and POSITION. Model *p*-values for fixed effects were derived using Satterthwaite's degrees of freedom method, implemented with the lmerTest package for R (Kuznetsova, Brockhoff, & Christensen, 2017). Bonferroni-corrected *p*-values are reported ($\alpha = 0.05$) where multiple comparisons were conducted.

2.6. Hypotheses

We predict that the more stable mode of coupling—i.e., in-phase (Onbeat) coupling—will lead to greater phonetic enhancement effects on those syllables which are synchronized with the (silent) metronome beat in the task. Given that English speakers already display considerable enhancement effects in the presence of stress, it may be the case that coupling-induced changes would be weaker overall in English than in Medʉmba, where these effects are not already present. Should phonetic enhancement occur as a result of coupling, we predict that significant interactions should be observed between the factors PHASING × POSITION for some or all of the dependent variables presented in Section 2.5. Specifically, in cases where coupling influences phonetic properties, it is predicted that word-initial syllables—those that speakers were instructed to coordinate with the beat—should show enhancement effects, but other syllables should show lesser or no effects. However, given the strong drive that English speakers often feel to align stressed syllables with a beat, it may be that interactions between PHASING × POSITION × STRESS will also emerge. Medʉmba speakers could show a similar interaction between PHASING × POSITION × PROMINENCE where word-position of metrically-strong syllables is manipulated (see Section 3.6) if these syllables also show an attraction to the metronome beat. An interesting question concerns whether Medʉmba speakers, in particular, show patterns of coupling-induced phonetic enhancement which look similar to those found for stressed syllables in other languages.

3. Results

Results across the two different metronome rates in the task followed similar patterns for all variables; we therefore present results of data collapsed across metronome speeds. Below, we begin with an overview of metronome alignment patterns exhibited across speakers of the two languages, followed by results for each of the acoustic variables of interest. We highlight results that are of particular interest and direct readers to Tables B1–B14 in Appendix B for full model results. For all graphs presented, error bars represent 95% confidence intervals.

3.1. Cross-linguistic metronome alignment patterns

Despite the similar overall trends in alignment patterns modeled for Onbeat and Offbeat conditions across the two languages, speakers of English and Medʉmba nonetheless gravitated towards quite different alignment strategies with the metronome (**Figures 2** and **3**). English stressed initial vowels in the Onbeat condition were produced very close to the metronome beat, trailing the beat slightly, by an average of about 7 ms. English unstressed syllables anticipated the beat by an average of 71 ms. Medʉmba speakers tended to anticipate the beat even more, placing initial high and low toned vowels an average of 104 and 65 ms before the beat, respectively. Note that even examining the data by initial segment type and word length, these differences in alignment across languages persisted. Conversely, initial syllables of Medʉmba speakers' Offbeat

repetitions were generally closer to the metronome beat than English speakers' by about 250 ms, suggesting that speakers of the two languages opted for quite different alignment strategies for this condition. As predicted, however, timing to the beat was less variable in the Onbeat versus the Offbeat condition for both languages as indicated by standard deviations (242 ms versus 316 ms for English; 278 ms versus 292 ms for Medʉmba), suggesting that coupling in the Onbeat condition was more stable than in the Offbeat condition regardless of alignment strategy. For both languages, an effect of PHASING was observed, with Onbeat repetitions occurring significantly earlier than Offbeat repetitions, as expected (English: $\beta = -243.15$, $t = -11.163$, $p < .001$; Medʉmba: $\beta = -138.76$, $t = -7.96$, $p < .001$).

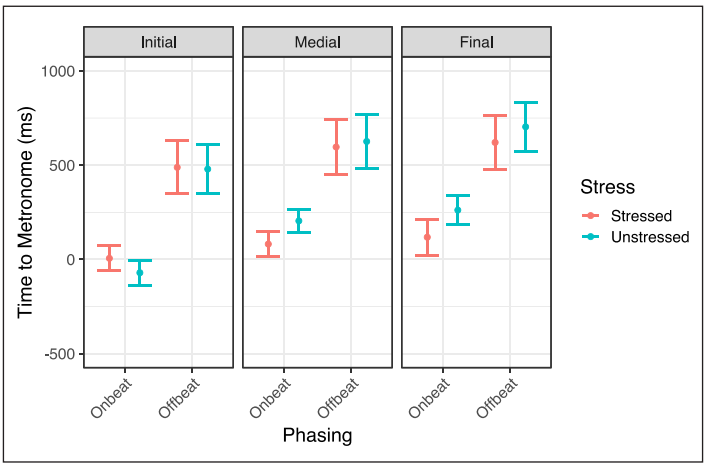

Figure 2: Metronome Distance as a function of phasing, word position, and stress; English speakers.

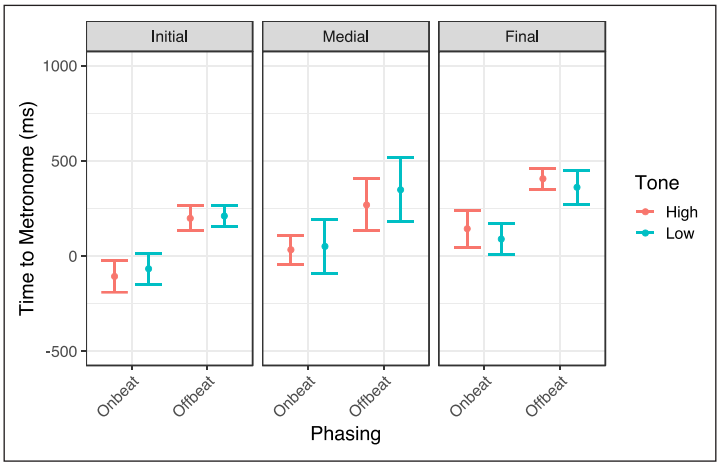

Figure 3: Metronome Distance as a function of phasing, word position, and tone; Medʉmba speakers.

Another striking aspect of English speakers' alignment patterns was the effect of STRESS on alignment: Stressed syllables—even when occurring in medial or final positions—occurred significantly closer to the metronome beat than did unstressed syllables (β = –24.86, t = –9.60, $p < .001$); this pattern is reflective of the fact that English speakers struggled to align unstressed initial syllables with the metronome beat, in some cases allowing a medial or final stressed syllable to align with the beat instead. As mentioned, high toned vowels in Medʉmba occurred slightly earlier than low toned vowels (β = –10.62, t = –2.57, $p < .05$). This difference could stem from differences in laryngeal timing for high and low tones (Erickson, 2011), or from differences in the perceptual centers of F0 patterns in Medʉmba (Franich, 2018b). A two-way interaction between PHASING × STRESS (β = –6.75, t = –2.61; $p < .01$) was also found for English speakers, reflecting the fact that timing differences between stressed and unstressed syllables were larger in the Onbeat condition than the Offbeat condition. A three-way interaction between PHASING × POSITION × STRESS among English speakers reflected the fact that, while unstressed syllables showed earlier timing than stressed syllables on the Onbeat in initial position, they showed later timing than stressed syllables in medial position (β = –16.34, t = –4.11; $p < .001$) as well as final position (β = –8.19, t = –2.24; $p < .05$). It is likely that this, too, is a reflection of English speakers' tendency to produce stressed syllables more closely to the metronome beat in the Onbeat condition, even when these syllables did not occur in word-initial position.

3.2. Duration

Overall, vowel duration was roughly similar between English speakers and Medʉmba speakers, with an average duration of 113 ms for Medʉmba speakers, and an average duration of 110 ms for English speakers. Effects of POSITION for both languages reflected that final syllables (which were both word- and phrase/utterance-final) were significantly longer than medial syllables (English: β = 22.93, t = 26.75; $p < .001$; Medʉmba: β = 12.05, t = 22.60, $p < .001$), and initial syllables (English: β = 47.15, t = 61.36, $p < .001$; Medʉmba: β = 31.20, t = 59.75, $p < .001$) (**Figures 4** and **5**). As expected, an effect of STRESS for English speakers reflected that stressed vowels were generally much longer in duration than unstressed vowels (β = 23.47, t = 42.34; $p < .001$). Medʉmba speakers showed somewhat longer duration for high tone vowels than low tone vowels (β = 10.24, t = 23.15; $p < .001$), though an interaction between TONE and POSITION indicates that this effect is greatest in final position (β = 11.39, t = 26.51; $p < .001$). This finding is intriguing given that past work has shown low and high tone syllables to have similar durations in monosyllabic words uttered in isolation in Medʉmba (Franich, 2018b; see also Franich, 2016). This pattern could stem from the fact that some of the word structures with final high tones are prosodically

complex, such that the final high tone syllable constitutes its own metrical foot (Franich, 2021).

English speakers showed no overall effect of PHASING on duration ($\beta = 1.55$, t = 1.98; $p = .07$), but did show a significant two-way interaction between PHASING × POSITION, such that initial syllables showed longer duration when occurring in the Onbeat condition than in the Offbeat condition, whereas medial vowels did not show this difference ($\beta = 1.47$, t = 2.03; $p < .05$); patterns in medial and final position did not differ from each other ($\beta = 0.39$, t = 0.46; $p = 0.64$). While the three-way interaction between PHASING × POSITION × STRESS did not reach significance, stressed syllables did show numerically larger differences between Onbeat and Offbeat conditions when in initial position compared with unstressed syllables ($\beta = 1.15$, t = 1.49; $p = .14$). In contrast with English speakers, Medumba speakers did show overall longer durations for vowels in the Onbeat condition ($\beta = 3.42$, t = 4.42; $p < .001$), but also showed an interaction between PHASING × POSITION, such that differences between the Onbeat and Offbeat condition were larger in word-initial position than either medial position ($\beta = 2.41$, t = 4.51; $p < .001$) or final position ($\beta = 2.71$, t = 5.18; $p < .001$). Thus, in neither language was it the case that phasing affected duration in all positions equally; this suggests that the observed effects of phasing on duration cannot be chalked up exclusively to differences in overall speech rates across phasing conditions.

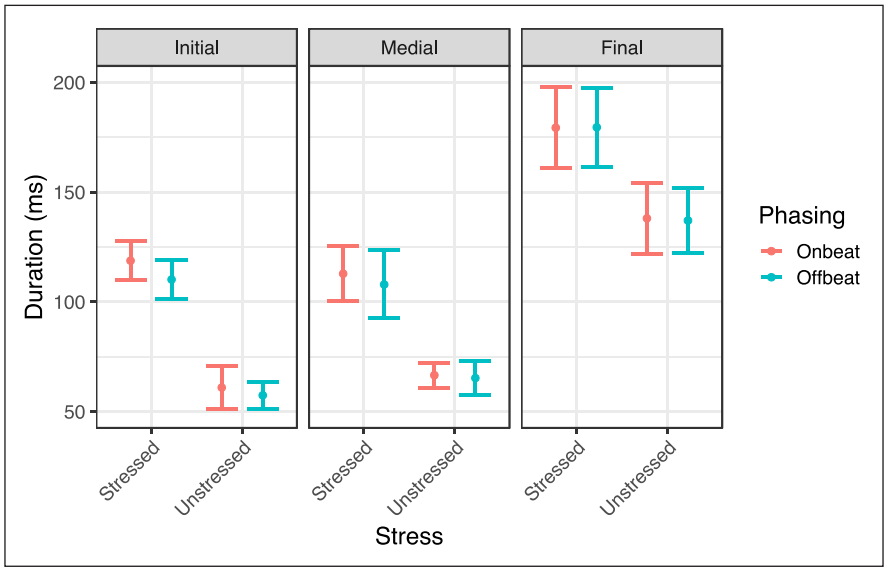

Figure 4: Vowel duration as a function of phasing, word position, and stress; English speakers.

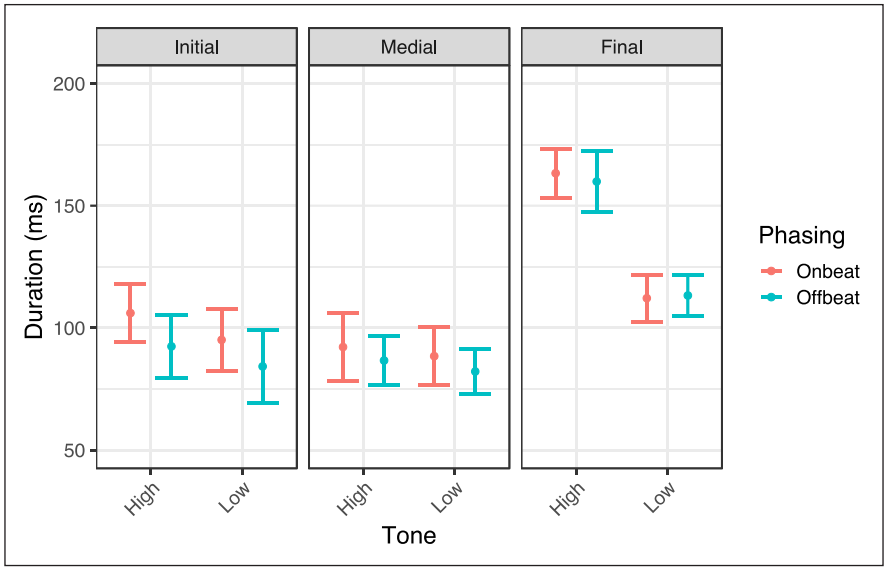

Figure 5: Vowel duration as a function of phasing, word position, and tone; Medumba speakers.

3.3. Fundamental frequency

Turning now to fundamental frequency, Medumba speakers showed a somewhat higher average F0 than English speakers, with mean F0 at 154 Hz for English speakers and 181 Hz for Medumba speakers (mean for high tones = 198 Hz; mean for low tones = 165 Hz). As expected, high tone vowels had much higher F0 than low tone vowels in Medumba (β = 0.54, t = 43.27; $p < .001$) (**Figure 7**). No significant difference in F0 was found between English stressed and unstressed syllables, and in fact the pattern trended toward lower F0 for stressed syllables than unstressed ones (β = –0.03, t = –1.94; p = .05) (**Figure 6**). Several English speakers appear to have assigned low pitch accents to the prominent syllables in the task, though speakers varied in this respect.

Metronome phasing had distinct effects on F0 across the two languages. While English speakers overall showed slightly higher F0 in the Onbeat condition than the Offbeat condition (β = 0.05, t = 4.06; $p < .001$), Medumba speakers showed the opposite effect, with lower F0 exhibited in the Onbeat condition than in the Offbeat condition (β = –0.05, t = –2.63; $p < .05$). For English speakers, an interaction between PHASING × STRESS reflected that the difference in F0 between conditions was greater for unstressed vowels than for stressed vowels (β = .05, t = 3.68; $p < .001$). For Medumba, a three-way interaction between PHASING × STRESS × TONE reflected the fact that phasing differences were especially large for high tone vowels occurring in final position (β = .03, t = 2.19; $p < .05$). The fact that F0 was more dramatically influenced by phasing in the later portion of words for Medumba, and that positional effects were not found at all for English, suggests that the effects of phasing on F0 are not necessarily about coupling of the

vowel to the beat per se, but rather representative or more general differences in performance across the two task conditions (see Section 4 for further discussion of this pattern).

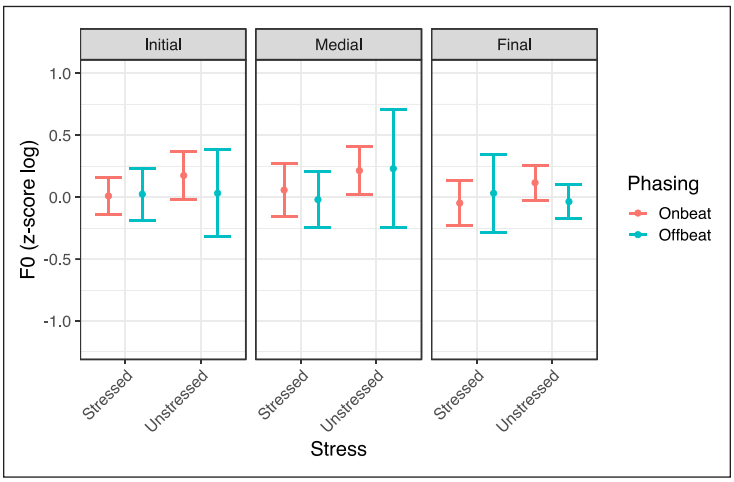

Figure 6: Vowel F0 as a function of phasing, word position, and stress; English speakers.

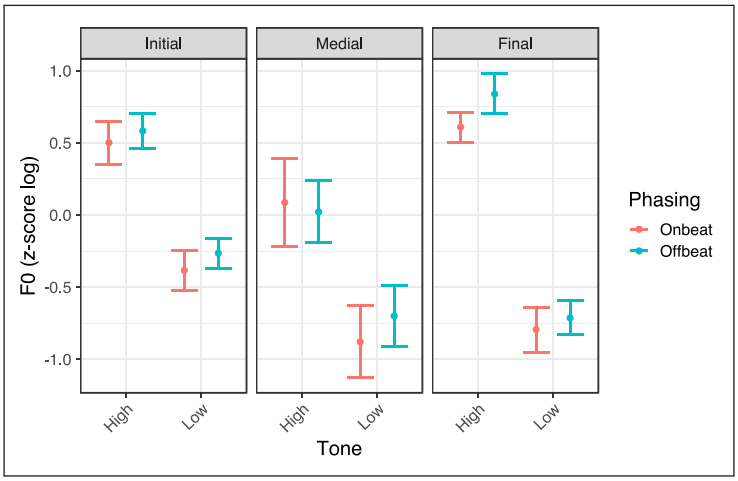

Figure 7: Vowel F0 as a function of phasing, word position, and tone; Medumba speakers.

3.4. Vowel formants

3.4.1. F1 and F2

Mean F1 values were similar across the two groups, though the mean was slightly higher for English speakers at 595.28 Hz versus for Medumba speakers at 525.38 Hz. Note that none of the English words examined had low vowels in final position. In both languages, an effect of POSITION was found: In both cases, F1 was found to be greater in initial position than medial position (English: $\beta = .06$, t $= 10.34$, $p < .001$; Medumba: $\beta = 0.10$, t $= 14.90$, $p < .001$)

(**Figures 8** and **9**), and higher in final position than in medial position (English: $\beta = 0.03$, t = 5.19, $p < .001$; Medʉmba: $\beta = –0.08$, t = –16.10, $p < .001$). Interactions between POSITION and VOWEL HEIGHT for both languages indicated that this pattern was not uniform across vowel heights, however: In English, low vowels in medial position displayed lower F1 than in initial position, while high vowels in medial position displayed higher F1 than in initial position ($\beta = .09$, t = 10.24 $p < .001$); comparison between mid and high vowels revealed the opposite trend ($\beta = –0.04$, t = –7.99 $p < .001$). In Medʉmba, the difference between medial and final vowels was larger for high vowels than for low vowels ($\beta = 0.02$, t = 3.26, $p < .001$) and smaller for high vowels than for mid vowels ($\beta = –0.001$, t = –2.68, p < .01).

No main effect of PHASING was found for F1 in either language (English: $\beta = –0.01$, t = –1.06, $p = .31$; Medʉmba: $\beta = –.006$, t = –1.68, $p = .09$). For English, a two-way interaction was found between PHASING × POSITION reflecting the fact that vowels produced in the Onbeat condition generally had slightly higher F1 in initial position, but not for medial position ($\beta = .01$, t = 2.21, $p < .05$); patterns between medial and final positions did not differ ($\beta = .00$, t = 0.07, $p = .94$). For both languages, a three-way interaction was found between PHASING × POSITION × VOWEL HEIGHT (English: $\beta = 0.02$, t = 2.08, $p < .05$; Medʉmba: $\beta = .01$, t = 2.18, $p < .05$). In both cases, low vowels in initial position exhibited higher F1 in the Onbeat condition, whereas this difference was absent or even reversed in medial position; for Medʉmba, no differences in patterning were observed between medial and final positions for low vowels by phasing condition ($\beta = –0.01$, t = –1.69, $p = .09$).

Mean F2 values were similar across the two groups, though the mean was slightly higher for English speakers at 1793.69 Hz versus for Medʉmba speakers at 1727.31. English initial syllables had higher F2 than medial syllables ($\beta = 0.05$, t = 17.28, $p < .001$) and lower F2 than final syllables ($\beta = –0.06$, t = –17.03, $p < .001$). Medʉmba medial syllables had higher F2 than both initial and final syllables (βs > 0.10; *t*s >4.15; *p*s < .001). We note, however, that back vowels were sparse in both initial and medial position for English, and back and front vowels were lacking in medial position for Medʉmba (**Figures 10** and **11**).

An effect of PHASING was found for English ($\beta = 0.02$, t = 4.39, $p <.001$), with higher F2 in the Onbeat condition, though a significant two-way interaction between PHASING × POSITION indicated that the effect was stronger in medial position than in initial position ($\beta = 0.01$, t = 2.88, $p < .01$). However, we note again that data were sparse in these positions for F2. The effect was reversed in final position ($\beta = –0.01$, t = –2.23, $p < .05$). A two-way interaction between PHASING and VOWEL BACKNESS indicated that F2 raising in the Onbeat condition was more pronounced for front vowels than for central vowels ($\beta = 0.02$, t = 4.10, $p < .001$). No effect of PHASING was found for Medʉmba ($\beta = 0.00$, t = 0.03, $p = .98$), nor was there a significant PHASING × POSITION interaction ($\beta = 0.00$, t = 0.28, $p = 0.78$). A significant three-way interaction between PHASING, POSITION, and VOWEL BACKNESS was not found for either language.

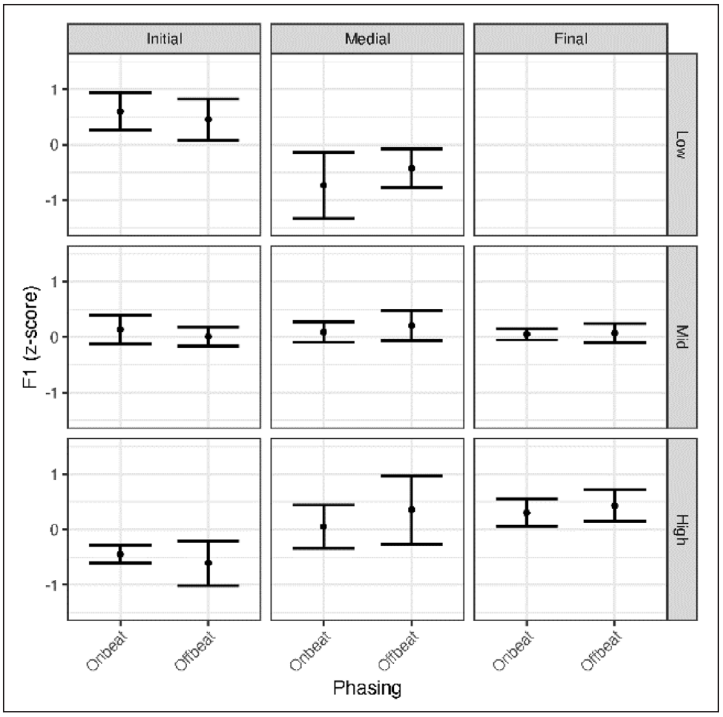

Figure 8: Vowel F1 as a function of phasing, word position, and stress; English speakers.

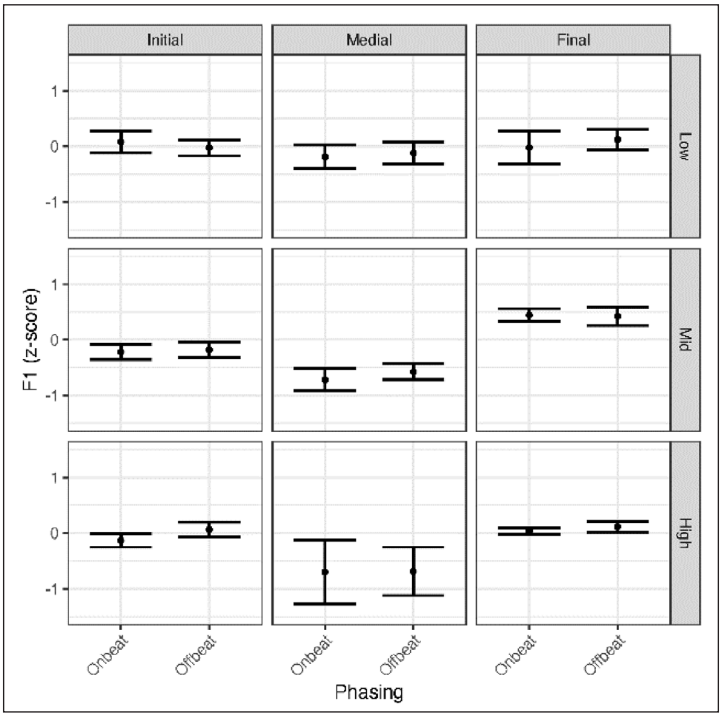

Figure 9: Vowel F1 as a function of phasing, word position, and tone; Medumba speakers.

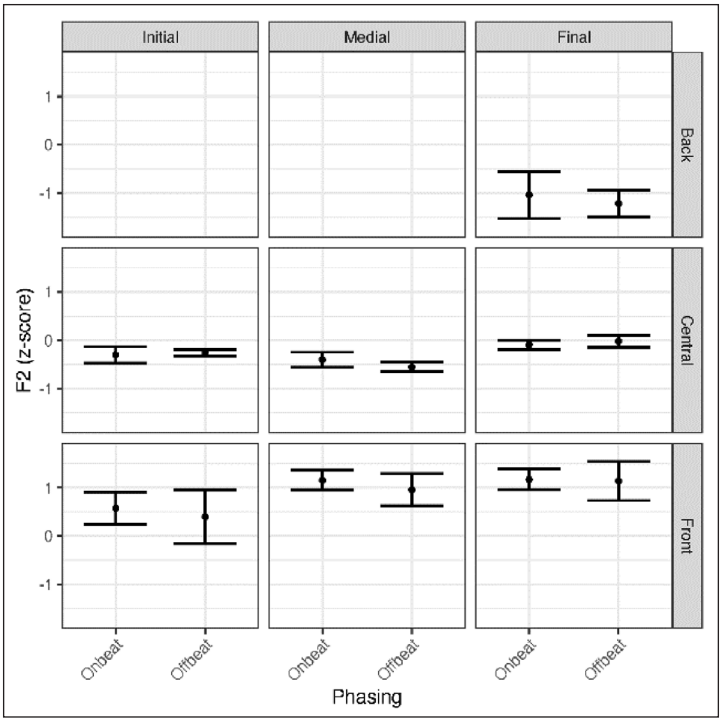

Figure 10: Vowel F2 as a function of phasing, word position, and stress; English speakers.

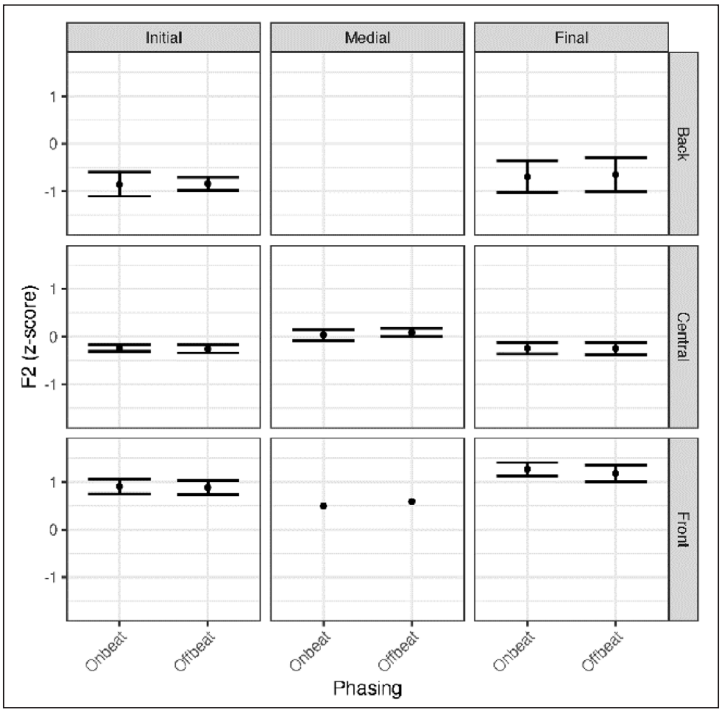

Figure 11: Vowel F2 as a function of phasing, word position, and tone; Medumba speakers.

3.4.2. Euclidean Distance

We now examine how speakers' overall vowel space may have been affected by phasing by looking at Euclidean Distance, a measure of how far tokens of each vowel occurred from the center of the speaker's vowel space. As expected, vowel duration had a significant effect on vowel space for both languages, with more expanded vowel space in the presence of longer vowels (English: $\beta = 0.13$, t = 9.22, $p < .001$; Medumba: $\beta = 0.07$, t = 5.37, $p < .001$). Neither group showed an overall effect of phasing on vowel space (English: $\beta = .004$, t = 0.05, p = .96; Medumba: $\beta = –.003$, t = –0.216, $p = .83$) (**Figures 12** and **13**). Surprisingly, English stressed vowels did not show significantly more expanded vowel space overall than unstressed vowels ($\beta = –0.02$, t = –1.77, $p = .08$). However, a two-way interaction between POSITION × STRESS reflected that stressed vowels in medial position did show greater Euclidean Distance than unstressed vowels ($\beta = –.07$, t = –4.41, $p < .001$). We note that this is the position within the word (and the phrase, given these words were uttered in isolation) which is least susceptible to edge-related lengthening effects (Fougeron & Keating, 1997; Byrd & Saltzman, 2003). One possibility is that stress-related differences in vowel space were neutralized in initial and final positions, where lengthening would apply, hence the lack of an overall effect. Three-way interactions between PHASING × POSITION × STRESS reflected that unstressed initial syllables showed increased vowel space in the Onbeat condition versus the Offbeat condition compared with medial syllables ($\beta = 0.05$, t = 2.92, $p < .01$). The expanded vowel space on unstressed syllables was greater in final position Onbeat vowels than initial position Onbeat vowels ($\beta = 0.06$, t = 4.16, $p < .001$). In Medumba, word-initial position was found to have more expanded vowel space than word-medial position ($\beta = 0.11$, t = 7.07, $p < .001$), and final vowels were found to have more expanded vowel space than initial vowels ($\beta = 0.23$, t = 12.35, $p < .001$). Low tone vowels were found to have overall smaller vowel space than high tone vowels ($\beta = –0.04$, t = –2.80, $p < .01$), although an interaction between POSITION × TONE indicated that this pattern was reversed in word-medial position ($\beta = –0.08$, t = –3.29, $p < .01$). No significant interactions were found involving PHASING for Medumba.

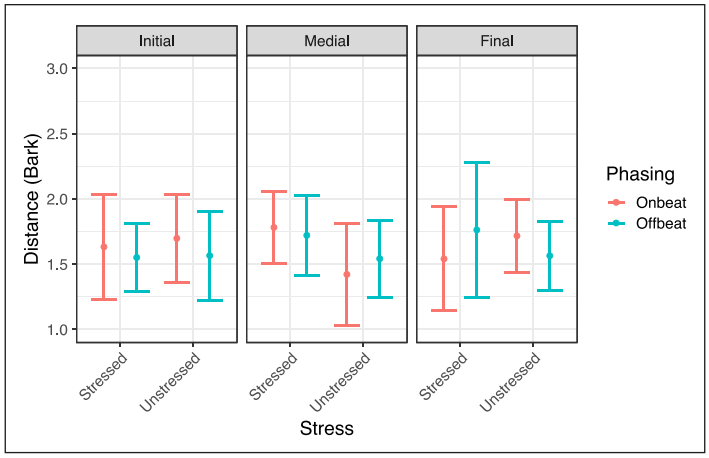

Figure 12: Euclidean distance as a function of phasing, word position, and stress; English speakers.

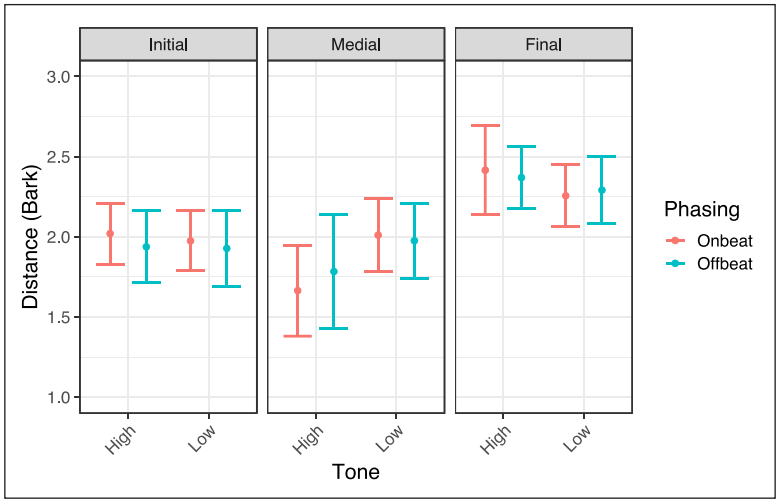

Figure 13: Euclidean distance as a function of phasing, word position, and tone; Medumba speakers.

3.5. Intensity

English and Medumba speakers showed similar overall patterns of intensity, both demonstrating a significant effect of POSITION, such that initial vowels had greater intensity than final vowels (English: $\beta = 1.92$, t = 25.23, $p < .001$; Medumba; $\beta = 1.82$, t = 23.12, $p < .001$); Medumba speakers also showed greater intensity in initial vowels compared to medial vowels, but this effect did not reach significance for English speakers (English: $\beta = 0.12$, t = 1.65, $p = .09$; Medumba *:* $\beta = 0.07$, t = 6.74, $p < .001$) (**Figures 14** and **15**). English speakers also showed greater intensity on stressed versus unstressed vowels ($\beta = 1.20$, t = 23.23, $p < .001$); Medumba speakers meanwhile showed greater intensity on high versus low tone vowels ($\beta = 1.35$, t = 22.86, p < .001). In Medumba, an effect of PHASING was found, with intensity found to be lesser in Onbeat condition than in the Offbeat condition ($\beta = –0.22$, t = –3.87, $p < .001$); English trended in the same direction, but the effect of phasing was not significant ($\beta = –0.30$, t = –1.48, $p = .17$). In both languages, an interaction between PHASING × POSITION was found, such that differences between Onbeat and Offbeat conditions were larger in final position than in initial position (English: $\beta = 1.70$, t = 2.61, $p < .01$; Medumba: $\beta = .21$, t = 3.01, $p < .01$). For English, a three-way interaction between PHASING × POSITION × STRESS indicated that effects of PHASING were more pronounced for stressed vowels in initial position than in medial position ($\beta = 0.12$, t = 2.47, $p < .05$).

3.6 Medumba stem-initial syllables by position

Finally, we turn our attention to a subset of the Medumba speakers' data in which we investigate how patterns of positional prominence interact with metronome coupling. Recall from Section 1 that stem-initial syllables in Medumba show evidence of greater rhythmic prominence than

non-stem syllables or stem-final syllables. We therefore sought to investigate a) whether there was any evidence that prominent syllables showed greater attraction to the metronome beat than prefix syllables or stem-final syllables; and b) whether syllable duration varied as a function of prominence and metronome phasing. Since this distinction is only represented in a small number of disyllabic words in our dataset containing a LH tone melody, we excluded TONE as a factor in the analysis and also did not analyze differences in fundamental frequency, vowel formants, or intensity.

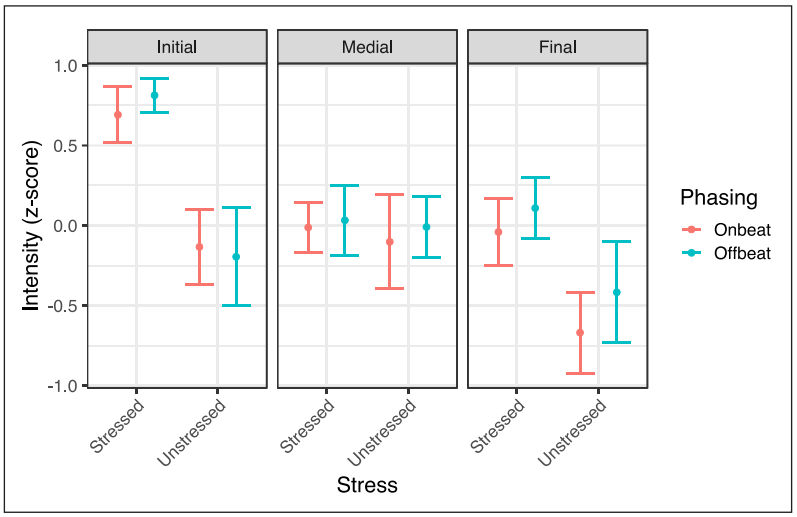

Figure 14: Vowel intensity as a function of phasing, word position, and stress; English speakers.

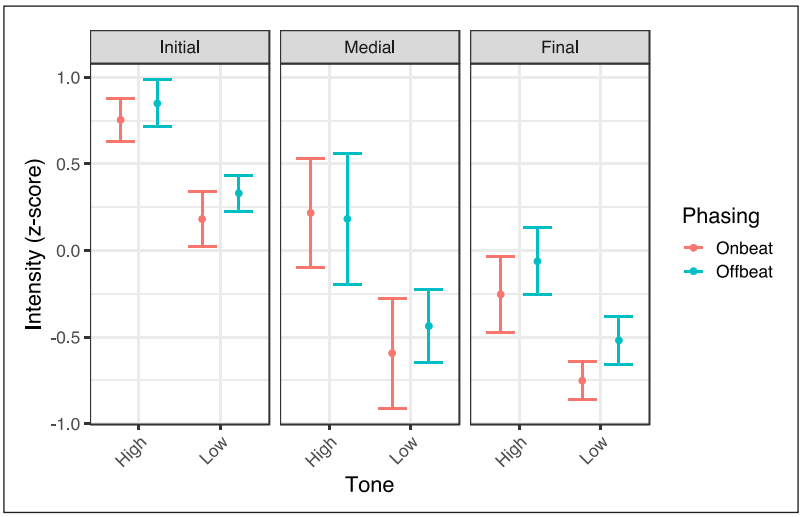

Figure 15: Vowel intensity as a function of phasing, word position, and tone; Medumba speakers.

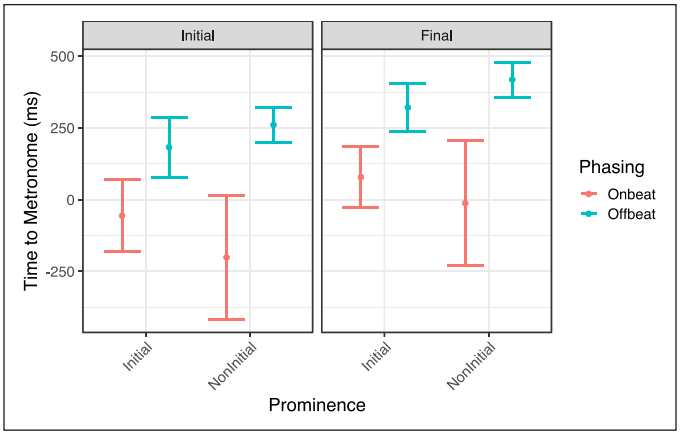

Figure 16: Metronome difference as a function of phasing, word position (Initial versus Final), and prominence (Initial versus NonInitial); Medumba speakers.

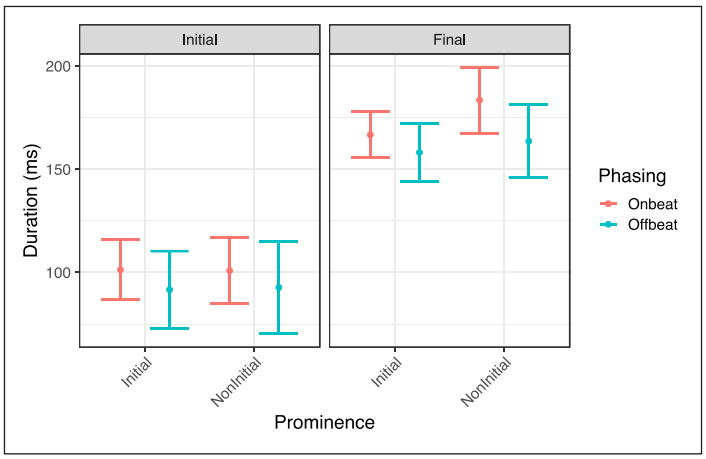

Figure 17: Vowel duration as a function of phasing, word position (Initial versus Final), and prominence (Initial versus NonInitial); Medumba speakers.

However, we can see interesting differences in the timing of vowels across conditions with respect to the metronome. Examining first differences in the timing of vowels cross conditions with respect to the metronome, as expected, vowels in the Onbeat condition were overall earlier than those in the Offbeat condition (β = –177.70, t = –5.51; p < .001), and vowels in initial position of the word were repeated earlier and closer in time with the metronome than those in final position (β = –77.81, t = –6.59; p < .001) (**Figure 16**). While we found no overall effect of PROMINENCE on metronome timing (β = 15.85, t = 1.29; p = .20), a significant two-way interaction between PHASING × PROMINENCE indicated that differences between Initial and NonInitial prominence conditions were reversed between the Onbeat condition than in the Offbeat condition (β = 56.40, t = 4.58; p < .001). **Figure 14** shows that while timing

of syllables in the NonInitial prominence condition was somewhat earlier than in the Initial condition when produced in the Onbeat phasing condition, words with noninitial prominence were uttered later with respect to the beat in the NonInitial condition in the Offbeat phasing condition. Interestingly, final (prominent) vowels in the NonInitial condition occurred right around the metronome beat on the Onbeat condition, whereas the initial (nonprominent prefix) syllable of that word occurred quite a bit earlier (~200 ms) than the beat; this is consistent with the idea that participants were timing their utterances earlier in order to ensure closer alignment of the prominent syllable with the metronome beat in the Onbeat condition.

Turning to results for duration, we find, similar to the larger Medumba dataset, effects of PHASING and POSITION on duration, such that vowels produced in the Onbeat condition were generally longer than those in the Offbeat condition ($\beta = 5.42$, t = 3.53; $p < .01$) and vowels produced in final position of a word were generally longer than those produced in initial position of a word ($\beta = 35.72$, t = 37.30; $p < .001$) (**Figure 17**). An effect of PROMINENCE indicated that vowels in words with non-initial prominence were longer than those with initial prominence, but an interaction between POSITION × PROMINENCE indicates the effect was greater in final position than initial position ($\beta = 2.60$, t = 2.72; $p < .01$); in other words, prominent syllables were longer than non-prominent syllables in word-final position, but not in word-initial position, possibly due to the overriding effects of initial vowel strengthening/lengthening (Fougeron, 2001; Turk & Shattuck-Hufnagel, 2000). Though no significant three-way interaction between PHASING × POSITION × PROMINENCE was found ($\beta = 1.28$, t = 1.34; $p = .18$), the numerically highest mean duration values were found for word-final prominent syllables in the Onbeat condition, consistent with the idea that these syllables may have been slightly lengthened due to their greater attraction to the metronome beat. We note that the difference in duration between final syllables in the Offbeat condition between words with Initial and NonInitial prominence was quite small—only 6 ms, on average—just slightly higher than the just noticeable difference for vowel duration observed in various languages (Nooteboom & Doodeman, 1980).

3.7 Comparisons between metronome-coordinated and uncoordinated speech

The primary goal of this study is to understand how coordination relations involving different levels of stability—i.e., in-phase (onbeat) and out-of-phase (offbeat) relations—contribute differently to phonetic enhancement effects. Thus far, we have seen evidence that in-phase coordination is linked with greater enhancement in the domains of vowel duration and F1 raising than out-of-phase coordination. What we have not yet explored is how these various forms of metronome-coordinated speech may differ from more naturalistic speech which is not coordinated to an external timekeeper. It could be, for example, that speech that is coordinated in any mode will show differences from uncoordinated speech. It could also be that enhancement effects observed in 'onbeat' speech are in fact similar to those found in naturalistic speech, and

that 'offbeat' speech rather leads to phonetic *reduction* of sorts. Examining speech spoken without the metronome will help to tease apart these possibilities. To that end, we provide additional analysis of patterns vowel duration and F1 frequency across the two metronome-coordinated Onbeat and Offbeat conditions, as well as a third 'NoBeat' condition in which speakers spoke without the timekeeper.

Average duration for speakers of both languages was longer in the NoBeat condition than in either the Onbeat or Offbeat condition (Medumba: 120 ms versus 116 ms and 110 ms, respectively; English: 130 ms versus 111 ms and 108 ms, respectively). This difference between NoBeat and Onbeat conditions was significant in both languages (English: $\beta = 15.20$, t = 21.91; $p < .001$; Medumba: $\beta = 6.21$, t = 11.63; $p < .001$). Differences in duration between initial and medial syllables in Medumba in the NoBeat condition were similar to those in the Offbeat condition, around 6 ms. For English, a two-way interaction between PHASING and POSITION revealed that differences between the NoBeat and Onbeat condition were smaller in initial position than medial position ($\beta = –6.86$, t = –6.51; $p < .001$), and final position ($\beta = –3.77$, t = –4.25; $p < .001$). As shown in **Figure 18**, a three-way interaction between PHASING, POSITION, and STRESS revealed that the difference between NoBeat and Onbeat conditions was particularly small for stressed syllables in initial position ($\beta = 6.52$, t = 6.19; $p < .001$). In Medumba, there was a significant two-way interaction between PHASING and POSITION, reflecting the fact that differences in duration between the Onbeat and NoBeat conditions were larger in initial position than in final position ($\beta = 3.57$, t = 6.07; $p < .001$); differences were also numerically higher between these two conditions in initial position compared with medial position, though the effect did not reach significance ($\beta = 1.53$, t = 1.83; $p = .07$). A three-way interaction was found between PHASING, POSITION, and TONE, indicating duration was highest in the Onbeat condition in initial position for high tones ($\beta = 3.98$, t = 4.75; $p < .001$) (**Figure 19**).

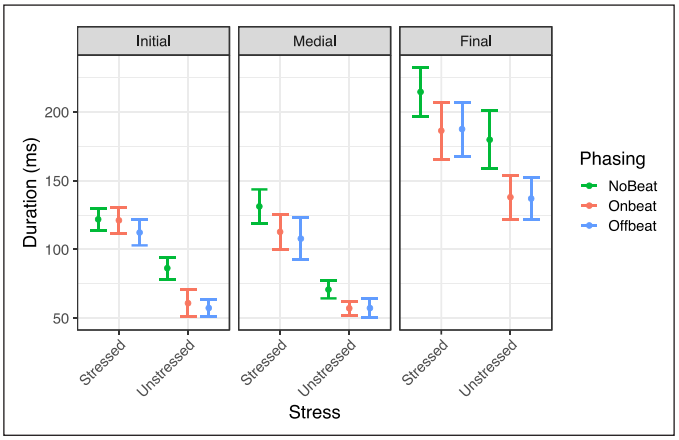

Figure 18: Metronome difference as a function of word phasing (3-way), word position, and stress, English speakers.

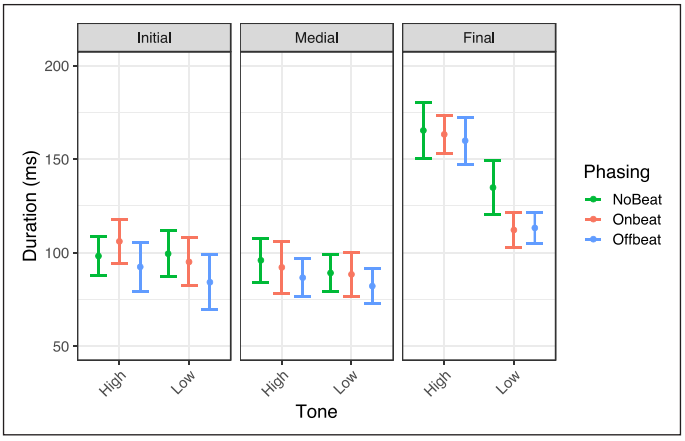

Figure 19: Vowel duration as a function of word phasing (3-way), word position, and tone, Medumba speakers.

Given the exceptionally large difference between durations in the NoBeat condition compared to the two metronome coordinated conditions in the English data (an indicator of overall slower speech rate in this condition for English speakers), an additional analysis was conducted on *relative duration*, treated as the ratio of stressed vowel duration to unstressed vowel duration within each word. An effect of PHASING reflected the fact that durational differences between stressed and unstressed syllables were higher in the Onbeat versus the Offbeat condition ($\beta = 0.11$, t $= 6.14$, $p < .001$) and lower in the NoBeat condition than in the Offbeat condition ($\beta = -0.20$, t $= -9.55$, $p < .001$) (**Figure 20**). However, a significant two-way interaction between PHASING and POSITION reflected that the difference in relative timing between the Onbeat versus Offbeat condition was larger in initial position than medial position ($\beta = 0.05$, t $= 2.52$, $p < .05$), and that the effect was reversed in final position ($\beta = -0.08$, t $= -3.03$, $p < .01$).

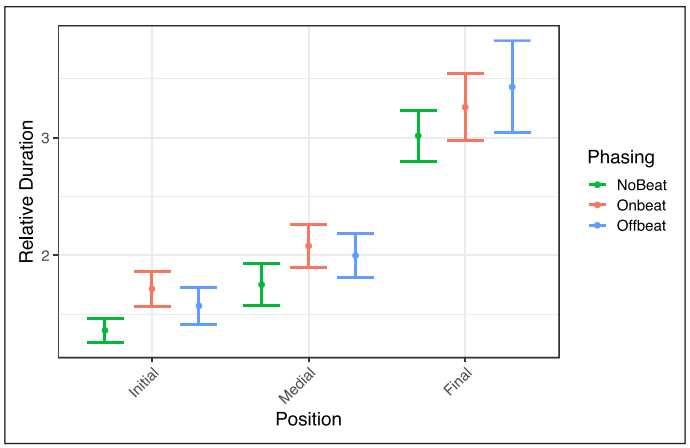

Figure 20: Relative duration (ratio of stressed to unstressed vowel duration) across positions and phasing conditions, English speakers.

Turning to first formant frequencies, results indicated that F1 was generally lower in the NoBeat condition for Medʉmba compared with both the Onbeat (β = –0.18, ts = –14.32, p < .001) and Offbeat (β = –0.22, t = –17.33, p < .001) conditions, possibly a reflection of more effortful speech production in the latter two conditions (Liénard & Di Benedetto, 1999). For English, the NoBeat condition did not differ significantly in F1 from the Onbeat condition (β = –0.03, t = –1.69, p = .09), or the Offbeat condition (β = 0.006, t = 0.31, p = .76). In English, a two-way interaction between PHASING and POSITION reflected that F1 was higher in the Offbeat condition than the NoBeat condition in medial position, but not initial position (β = 0.09, t = 3.62, p < .001); no significant difference was found between initial and final position between these two phasing conditions (β = –0.03, t = –1.83, p = .07). A three-way interaction between PHASING, POSITION, and VOWEL HEIGHT indicated that F1 was higher in the Onbeat condition than the NoBeat condition for low vowels in initial position, but not medial position (β = 0.09, t = 2.20, p < .05) (**Figure 21**). For Medʉmba speakers, a significant two-way interaction between PHASING and POSITION indicated that differences between the NoBeat condition and the OnBeat condition were larger in initial position than in medial position (β = –0.07, t = –3.30, p < .001); the same difference was found between the NoBeat and Offbeat conditions across word positions (β = –0.05, t = –2.72, p < .01). Differences between the NoBeat and Onbeat/Offbeat conditions were more pronounced in final position than initial position (βs > 0.06, ts > 4.43, ps < .001). Three-way interactions between PHASING, POSITION, and VOWEL HEIGHT reflected that positional differences in F1 between phasing conditions were more pronounced within mid vowels as compared with high vowels (β = 0.09, t = 4.68, p < .001) (**Figure 22**).

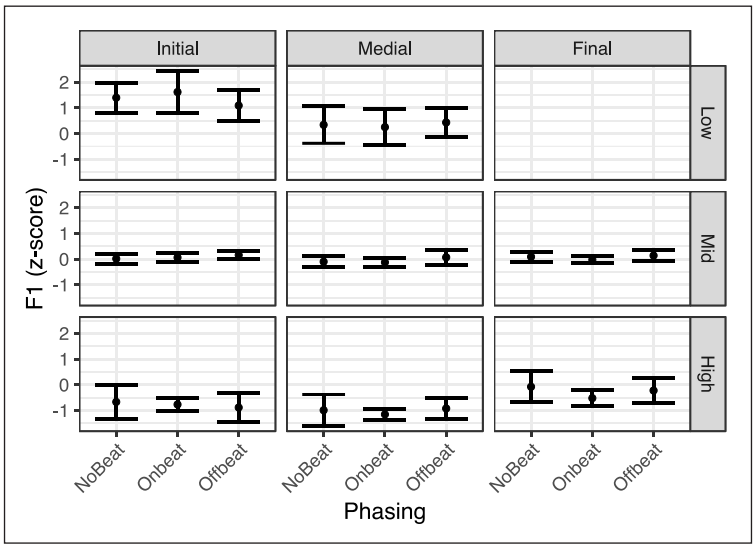

Figure 21: F1 as a function of phasing (3-way), position, and vowel height, English speakers.

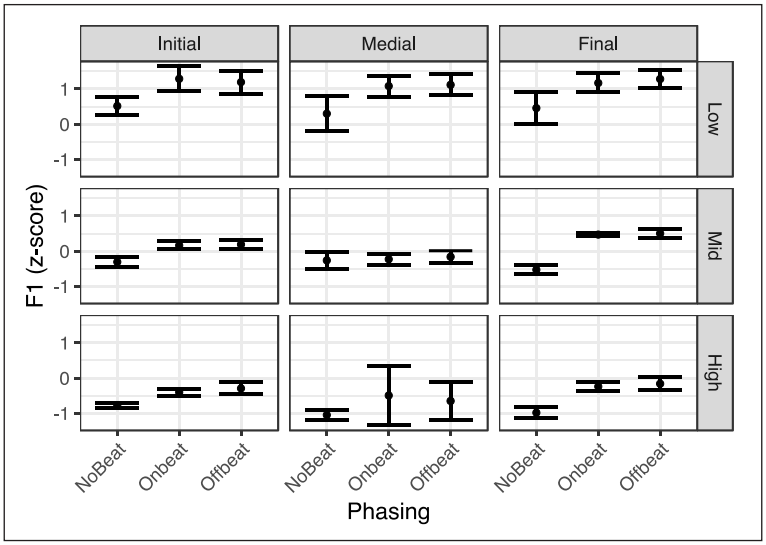

Figure 22: F1 as a function of phasing (3-way), position, and vowel height, Medʉmba speakers.

To summarize, patterns of vowel duration and F1 height differed in key ways in the NoBeat condition compared with the two metronome-timed conditions. Longer duration in the NoBeat condition reflects an overall slower speech rate in that condition as compared with the metronome-coordinated conditions. Decreased F1 was also found in this condition for Medʉmba speakers, which is consistent with greater articulatory effort in the two metronome-coordinated conditions as compared with the uncoordinated condition (Huber, Stathopoulos, Curione, Ash, & Johnson, 1999; Huber & Chandrasekaran, 2006; Traunmüller & Eriksson, 2000). Despite differences in how participants performed the task under coordinated and uncoordinated conditions, evidence suggests that performance in the Onbeat condition still differed in key ways as compared with both the Offbeat and NoBeat conditions. In particular, in spite of the overall longer duration found for vowels in the NoBeat condition, Medʉmba speakers produced initial vowels in the Onbeat condition with greater duration than those in the NoBeat condition. For English speakers, where relative duration was concerned, the ratio of stressed to unstressed vowel duration was higher in initial position in the Onbeat condition as compared with both the NoBeat and Offbeat conditions. It thus appears that syllables produced in the Onbeat condition show genuine patterns of phonetic enhancement.

4. Discussion

4.1 Phasing effects on acoustic patterning

While most variables in the study showed sensitivity to the phasing manipulation, some of these effects—such as for intensity and F0—were observed for syllables across word positions, indicating a more general effect of metronome phasing that did not relate to coupling, per se. We

begin by discussing these effects, followed by a discussion of the coupling-specific effects found for vowel duration and first formant frequency.

4.1.1 Task and language effects on intensity and F0

Despite the fact that neither intensity nor F0 showed coupling-specific effects, differences found for these variables by phasing condition and language were nonetheless interesting and worthy of comment. First off, in both languages, intensity was found to increase in the Offbeat condition as compared with the Onbeat condition. This could be a reflection of the greater effort involved in coordinating with an (imaginary) offbeat versus onbeat in the task, as increased vocal intensity has been noted as a key correlate of more effortful speech across various languages (Liénard & di Benedetto, 1999; Titze & Sundberg, 1992). In these studies, fundamental frequency has also been found to be increased under increased effort, which is consistent with what was found for English speakers in the present study, but not for Medʉmba speakers, who actually showed *reduced* F0 in the Offbeat condition. While it is the case that patterns of F0 and intensity often covary due to the potential for the vocal folds to vibrate more strongly with greater subglottal pressure (Hirano, Ohala, & Vennard, 1969; Titze, 2000), our findings are consistent with results from Tilsen (2016) who, based on inter-speaker variability patterns in covariation in F0 and intensity in English, argued that the two parameters are not related solely through physiological mechanisms. Work from Zhang (2016) has also shown that F0 manipulations can be controlled through vocal fold stiffness independently of intensity change. While our results are novel in that they suggest articulation of increased vocal effort is language-specific, it is not yet clear why F0 should be raised under these conditions in English but lowered in Medʉmba.

4.1.2 Coordination, coupling, and enhancement: Duration and F1 frequency

Results of the experiments also showed that in-phase (onbeat) coupling with the metronome beat yielded modest increases in vowel duration and first formant frequency (a correlate of jaw/tongue lowering) for both English and Medʉmba speakers. That these effects can be linked to metronome coupling specifically, as opposed to other mechanisms (such as overall variations in speech rate across phasing conditions), is confirmed by the fact that effects were localized to those syllables which were targeted for alignment with the metronome in the task. Effects of coupling on duration were more pronounced for Medʉmba speakers than for English speakers; this was also predicted given that English speakers already show a large amount of phonetic enhancement for stressed syllables, which were preferentially aligned in the task with the metronome beat. Three-way comparisons between the two metronome-coordinated conditions and an uncoordinated speech condition revealed that enhancement effects of Onbeat syllables were evident even in comparison to more naturalistic speech, despite the overall slower speech rate found in the NoBeat condition. Patterns of lowered F1 in the NoBeat condition as compared

with the Onbeat and Offbeat conditions suggest that speech in the uncoordinated condition was less effortful overall, so apparent enhancement effects in the Onbeat condition cannot be attributed to articulatory reduction in the Offbeat condition.

It is clear that metronome coupling did *not* result in English-like stress behavior among Medʉmba speakers: For example, increases in vowel duration between Offbeat and Onbeat conditions for Medʉmba speakers were 10–13% on average (around 12 ms), whereas English typically shows around a 40% increase in duration between unstressed and stressed syllables. English is of course a fairly extreme example of a stress-based language exploiting duration, however, given that it is a 'stress-timed' language with large amounts of vowel reduction (Dauer, 1983). We might expect, then, that Medʉmba, which patterns more like a typical 'syllable-timed' language (Franich, 2018a), would show enhancement effects more similar to stressed syllables in other syllable-timed languages. A difference of 10–13% between stressed and unstressed syllables is in fact on the order of what has been found for unaccented syllables in languages like Spanish (Ortega-Llebaria & Prieto, 2007, 2011). Thus, in terms of duration, the behavior elicited by metronome coupling among Medʉmba speakers is similar to a stress effect in a syllable-timed language.

These results could have important implications for our understanding of the relationship between coordination and phonetic enhancement. To begin with, it is notable that the variable that showed the strongest and most consistent effect of phasing in our data was duration, given the fact that this cue is one of the most reliable acoustic cues to stress across languages. For example, a recent cross-linguistic survey by Gordon and Roettger (2017) shows that, in a large sample of genetically-diverse languages, 85% of languages for which duration had been examined as a possible stress cue showed duration to be a key acoustic correlate of stress, compared with 70% showing use of intensity and 69% showing use of F0. Variations in formant frequency were shown to be exploited in 83% of languages in which this correlate was examined. Though the authors point out that duration is also one of the most common correlates of stress which is studied in the first place, the cross-linguistic robustness of this cue for stress is nonetheless striking.[2] Duration is also exploited as a stress cue even in many languages with contrastive vowel length, in direct opposition to Berinstein's (1979) Functional Load Hypothesis (Lunden, Campbell, Hutchens, & Kalivoda, 2017). In contrast, F0 as a stress cue is consistently unattested or only marginally present in languages where F0 is used for other means, such as tone marking (Caballero & Carroll, 2015; Chávez-Peón, 2008; Michael, 2011; Remijsen, 2002; Remijsen & van Heuven, 2005; Tallman & Elías-Ulloa, 2020).

[2] The authors also note that a common limitation of many studies (including the present study) is that stress is not sufficiently disentangled from phrase-level pitch-accent, meaning that F0 and intensity cues, in particular, may in fact be overstated in the sample.

That F1 differences were also found in our data is consistent with past work showing that increased duration may arise as a result of increased jaw lowering (Flege, 1988). Of course, the extent to which jaw lowering can definitively be implicated in our results would need to be directly verified through a study using articulatory methods such as electromagnetic articulography. The idea that jaw lowering amplitude should show similar patterns to limb movement amplitude as a function of metronome phasing make sense intuitively, though, given that both of these types of movement involve similar physical systems, including hinge-style synovial joints which allow for basic muscle-controlled flexion and extension or depression and elevation. By contrast, laryngeal adjustments for the production of F0 variation result from the complex rotation and rocking movements of the arytenoid cartilages controlled by the intrinsic laryngeal muscles to narrow and lengthen or shorten the vocal folds, and the raising and lowering of the thyroid cartilage via the thyroarytenoid muscle. Subglottal pressure fluctuations resulting from changes in contraction patterns of the intercostal muscles (for the production of intensity and F0 variation) also represent a considerably different kind of physiological process than limb or jaw movement. And while it is possible that F0 and intensity would show similar patterns of phonetic enhancement were the design of the present study to have controlled more tightly for differences in articulatory effort across metronome phasing conditions, independent work examining the relationship between manual gesture coordination and enhancement patterns in Medumba supports the idea that the relationship between coordination and enhancement is language-specific for some variables, including F0 (Franich & Keupdjio, 2022).

4.2 Coordinative roots of stress?

How, then, might our observed effects of coupling relate to stress more broadly? One possibility is that the coupling-related differences such as the ones observed in the present study are representative of a broader, possibly universal biomechanically-motivated pattern of enhancement which would be expected to emerge whenever speech is coordinated in-phase with another element, whether it be internal or external to the body. From this perspective, coordination at the linguistic level can be seen to be driven at its core not by patterns of perceptual prominence, but rather by more abstract rhythmic properties of language such as foot structure. While it is certainly not the case that coordination 'causes' stress in a broad sense, from a diachronic perspective, subtle patterns of durational variability resulting from coordination might have served as a phonetic 'precursor' to phonologization (Hyman, 1976), whereby the pattern could have been enhanced in those languages in which it was grammaticalized as stress, through the application of something like a clock-slowing μ_{Γ}-gesture. Further changes could have then taken place, such as the incorporation of other acoustic cues such as amplitude and fundamental frequency, in order to enhance perceptual correlates of the existing phonological stress contrast in language-specific ways (Hall, 2011).

Another aspect of coordination that is interesting to consider from this perspective is that of articulatory coordination at different positions with a syllable or word. Research has shown that segments across different languages show patterns of strengthening in word-initial position (Byrd, 2000; Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 1999; Keating, Cho, Fougeron, & Hsu, 2003). Syllable onsets across a number of languages have also been shown to display qualitatively different timing patterns than syllable codas, with the former displaying greater temporal overlap between consonantal and vocalic articulatory gestures (Browman & Goldstein, 1988; Byrd, 1995; Goldstein, Saltzman, Chitoran, & Nam, 2009). Word-internally within polysyllabic words, patterns of articulatory timing are more variable, with consonant sequences sometimes syllabifying as onsets, and other times as coda-onset sequences (Byrd et al., 2009; Garvin, 2021). Therefore, syllable onsets in word-initial position represent units of articulatory timing which involve the greatest amount of synchronous gestural coordination and the highest level of stability, two patterns which characterize the kind of superimposition which we have found lead to enhancement effects in the present work. It is therefore interesting to consider whether word-initial strengthening effects may stem from a similar phenomenon as stress-related enhancement. Of further interest is the fact that stress is a strong predictor of syllabification in word-medial contexts, with sequences of consonantal gestures more likely to syllabify as onsets if preceding a stressed vowel (Byrd et al., 2009; Garvin, 2021); future work will be beneficial in shedding further light on this relationship. Notably, however, domain edge effects and word stress effects have been shown to display both quantitatively and qualitatively distinctive patterning, suggesting there are key differences in the mechanisms that drive them (Cho & Keating, 2009). For example, the level of contact during consonantal gestures is found to be influenced by proximity to a prosodic boundary, but not by stress.

As has been discussed throughout the paper, a special property of stressed syllables is that, in addition to displaying characteristic coordination patterns at the level of the speech articulators, they show distinctive coordination behavior with other parts of the body and with body-external stimuli, as well. Two key examples are that stressed syllables show preferential timing with co-speech gesture and with rhythmic stimuli like music. Of course, speakers of non-stress languages like Medʉmba also coordinate their speech to gesture and to music; in Medʉmba, for example, foot-initial syllables have been found to be targeted for gesture alignment and to play an important role in musical text-setting (Franich & Keupdjio, 2022; Franich & Lendja, 2021). While there is some evidence for vowel and consonant distributional asymmetries related to foot structure in Medʉmba (Franich, 2021), typical stress cues such as increased duration are not found. Thus, like many other non-stress languages, Medʉmba has clearly not phonologized patterns of phonetic enhancement in the same way as speakers of languages like English (or, indeed, a syllable-timed language like Spanish). This could pertain to the status of Medʉmba as a lexical tone language: Since pitch is known to interact with duration in ways that can

distort duration perception (Yu, 2010), tone languages generally may not be good candidates for duration-based stress development. However, many tone languages do, in fact, show presence of duration-cued stress in addition to tone (Caballero & Carroll, 2015; Chávez-Peón, 2008; Remijsen, 2002; Remijsen & van Heuven, 2005), suggesting that there is nothing that inherently precludes the two from coexisting in a given language.

Another possibility is that preferred coupling patterns themselves differ cross-linguistically in a way which could bias certain languages away from developing durational cues to stress. Ethnomusicologists have noted that the approach to rhythm in many genres of music found in West and Central Africa involves metrical subdivision patterns which can rely on complex integer or non-integer ratios (Kubik, 2010; Polak, 2010; Polak & London, 2014) as opposed to the more isochronous timing found in musics of other parts of the world, most notably within Western European traditions. The complexity of metrical subdivision has been found to directly impact coupling strength in ensemble music playing, with more complex metrical organization associated with weaker coupling (Doffman, 2013). Given the fact that rhythmic preference, like language, is developed early in life and based on the surrounding cultural context (Soley & Hannon, 2010; Morrison, Demorest, & Stambaugh, 2008), one could imagine that a preference for certain metrical patterns over others could have far-reaching implications for how individuals of different cultures interact rhythmically with their environment. Indeed, our results on metronome coordination patterns presented in Section 3.1 are a direct reflection of preferred rhythmic patterns: Even when given a simple metronome beat to coordinate to with similar instructions, English and Medʉmba speakers showed very different alignment patterns, particularly where it came to the 'offbeat' condition. Such preferences could furthermore be imagined to impact coupling strength at an intrapersonal level if alignment of, for example, speech and gesture was regulated by timing relations other than perfect synchrony. From a musical perspective, studies of body movement during music and dance suggest that variability also exists in the way that individuals coordinate movements of different parts of their own bodies while dancing or moving to music from different cultures (Haugen & Godøy, 2014; Kilchenmann & Senn, 2015).

4.3 'Rhythm' and coordination in speech and music

The picture sketched in Section 4.2 is one in which the presence of 'rhythm' is not dependent on a particular phonetic cue or quality. Rather, rhythm is viewed here as a more fundamental aspect of linguistic structure and timing—more in the sense of Liberman (1975) and Liberman and Prince (1977)–which can then be enhanced in speech production by way of coordinative patterns and phonetic enhancement. This view is in line with work within music theory which posits that beat 'prominence'—associated with the beats around which coordinated movement takes place—does not depend directly on properties of a musical stimulus, but rather on metrical expectations which shape the selective enhancement of some beats over others in perception (Nozaradan,

Peretz, & Mouraux, 2012; Tal et al., 2017). This helps to explain how individuals can hear illusory metrical accents even when none are present in a signal. Nonetheless, physical qualities of the stimulus can also serve to enhance entrainment to a beat, suggesting that expectations and signal can mutually influence one another (Lenc, Keller, Varlet, & Nozaradan, 2018). Within this literature, it has also been shown that performing body movements in time with a perceived beat enhances perception of, and neural response to, a musical beat, suggesting that the motor system plays an important role in rhythm perception more generally (Nozaradan, Zerouali, Peretz, & Mouraux, 2015; Nozaradan, Peretz, & Keller, 2016).

Language is, of course, generally less 'rhythmic' in the periodic sense than music in terms of relative timing between 'beats' (metrical prominences), so a direct comparison across the two domains does not seem plausible at first glance. However, the communicative function of language enables listeners to make predictions based on a number of other properties besides strict isochronous timing of syllables or stresses; these properties are also demonstrated to play a role in promoting entrainment to the speech signal (Riecke et al., 2018). Furthermore, beat induction has been shown to be a robust phenomenon in music even with very complex rhythms (Fiveash et al., 2020; Stupacher, Wood, & Witte, 2017), suggesting that further research into the construction of temporal expectations in language based on metrical and other properties will be fruitful.

5 Conclusion

This study has shown that coupling speech to a metronome serves to enhance certain phonetic cues across languages, including vowel duration and first formant frequency, both known correlates of increased jaw opening. Our results thus suggest that phonetic enhancement, rather than being driven purely by perceptual factors, has roots also in the speech-motor system. Likewise, results suggest that the notion of prosodic 'prominence' should be considered to involve aspects of language use which go beyond the speech system, and which may concern aspects of body movement and interaction with other systems in the environment. Understanding speech and phonological structure within this broader context may aid in our understanding of language typology more generally.

Additional files

The additional files for this article can be found as follows:

- **Appendix A:** List of stimuli. DOI: https://doi.org/10.16995/labphon.6452.s1
- **Appendix B:** Mixed effects model results tables. DOI: https://doi.org/10.16995/labphon.6452.s2

Acknowledgements

Many thanks to the study participants and to Dr. Ange Bergson Lendja for assistance in coordinating the Medumba portion of the study. I would also like to thank His Majesty Mbiandou Yogang Bernard, Chief of the Lafeng District, and His Majesty Yonkeu Kuika Jean, Paramount Chief of the Bangoulap community, for their support. Thanks to participants at the 2021 *Rhythm Production and Perception Workshop* for valuable feedback on this work. All errors are my own.

Funding information

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.

Competing interests

The author has no competing interests to declare.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Aldine and Atherton.

Allen, G. D. (1972). The Location of Rhythmic Stress Beats in English: An Experimental Study I. *Language and Speech, 15*(1), 72–100. DOI: https://doi.org/10.1177/002383097201500110

Andrews, C., O'Brian, S., Harrison, E., Onslow, M., Packman, A., et al. (2012). Syllable-timed speech treatment for school-age children who stutter: A phase I trial. *Language, Speech, and Hearing Services in Schools, 43*, 359–369. DOI: https://doi.org/10.1044/0161-1461(2012/11-0038)

Audacity Team. (2018). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 17th 2021 from https://audacityteam.org/.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Baumann, S., & Röhr, C. T. (2015). The perceptual prominence of pitch accent types in German. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland.

Beckman, J. N. (1997). Positional faithfulness, positional neutralisation and Shona vowel harmony. *Phonology, 14*(1), 1–46. DOI: https://doi.org/10.1017/S0952675797003308

Beckman, M. E. (1996). The Parsing of Prosody. *Language and Cognitive Processes*, *11*(1–2), 17–68. DOI: https://doi.org/10.1080/016909696387213

Beckman, M., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Eds.), *Gesture, Segment, Prosody* (1st ed., pp. 68–89). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511519918.004

Berinstein, Ava E. (1979). A cross-linguistic study on the perception and production of stress. *UCLA Working Papers in Phonetics 47*, 1–59

Browman, C., & Goldstein, L. (1988). Some notes on syllable structure in Articulatory Phonology. *Phonetica*, *45*(140–155). DOI: https://doi.org/10.1159/000261823

Byblow, W. D., Carson, R. G., & Goodman, D. (1994). Expressions of asymmetries and anchoring in bimanual coordination. *Human Movement Science*, *13*(1), 3–28. DOI: https://doi.org/10.1016/0167-9457(94)90027-2

Byrd, D. (1995). C-Centers Revisited. *Phonetica*, *52*(4), 285–306. DOI: https://doi.org/10.1159/000262183

Byrd, D. (2000). Articulatory Vowel Lengthening and Coordination at Phrasal Junctures. Phonetica, *57*(1), 3–16. DOI: https://doi.org/10.1159/000028456

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180. DOI: https://doi.org/10.1016/S0095-4470(02)00085-2

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics*, *37*(1), 97–110. DOI: https://doi.org/10.1016/j.wocn.2008.10.002

Caballero, G., & Carroll, L. (2015). Tone and Stress in Choguita Rarámuri (Tarahumara) Word Prosody. *International Journal of American Linguistics*, *81*(4), 457–493. DOI: https://doi.org/10.1086/683157

Carson, R. G. (1990). The dynamics of isometric bimanual coordination. *Experimental Brain Research, 105*(3). DOI: https://doi.org/10.1007/BF00233046

Chávez-Peón, M. (2008). Phonetic cues to stress in a tonal language: Prosodic prominence in San Lucas Quiaviní Zapotec. *Proceedings of the 2008 Annual Conference of the Canadian Linguistic Association*, Vancouver.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878. DOI: https://doi.org/10.1121/1.1861893

Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485. DOI: https://doi.org/10.1016/j.wocn.2009.08.001

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*, 425–452. DOI: https://doi.org/10.1515/labphon.2010.022

Crosswhite, K. 2001. *Vowel reduction in Optimality Theory*. New York: Routledge.

Cummins, F. (1997). *Rhythmic Coordination in English Speech: An Experimental Study* (Doctoral dissertation). Indiana University.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, *3*(1), 7–11. DOI: https://doi.org/10.1121/1.1416672

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16–28. DOI: https://doi.org/10.1016/j.wocn.2008.08.003

Cummins, F., & Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145–171. DOI: https://doi.org/10.1006/jpho.1998.0070

Dauer, R. M. (1983). Stress-timing and syllable-timing re-analysed. *Journal of Phonetics*, *11*, 51–62. DOI: https://doi.org/10.1016/S0095-4470(19)30776-4

de Jong, K., & Zawaydeh, B. A. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*, *27*, 3–22. DOI: https://doi.org/10.1006/jpho.1998.0088

de Poel, H. J., Roerdink, M., Peper, C. (Lieke) E., & Beek, P. J. (2020). A Re-Appraisal of the Effect of Amplitude on the Stability of Interlimb Coordination Based on Tightened Normalization Procedures. *Brain Sciences*, *10*(10), 724. DOI: https://doi.org/10.3390/brainsci10100724

Dell & Halle (2009). Comparing musical textsetting in French and in English songs. In J.-L. Aroui & A. Arleo (Eds.), *Towards a Typology of Poetic Forms* (pp. 63–78). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lfab.2.03del

Doffman, M. (2013). Groove: Temporality, awareness and the feeling of entrainment in jazz performance. In M. Clayton, B. Dueck & L. Leante (Eds.), Experience and meaning in music performance (pp. 62–85). Oxford, UK: Oxford University Press.

Edwards, J., & Beckman, Mary. E. (1988). Articulatory Timing and the Prosodic Interpretation of Syllable Duration. *Phonetica*, *45*(2–4), 156–174. DOI: https://doi.org/10.1159/000261824

Erickson, D. (2002). Articulation of Extreme Formant Patterns for Emphasized Vowels. *Phonetica*, *59*(2–3), 134–149. DOI: https://doi.org/10.1159/000066067

Erickson, D. (2011). Thai tones revisited. *Journal of the Phonetic Society of Japan*, *15*(2), 1–9.

Erickson, D., & Kawahara, S. (2016). Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistics Vanguard*, *2*(1). DOI: https://doi.org/10.1515/lingvan-2015-0025

Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M., & Prieto, P. (2017). The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America*, *141*(6), 4727–4739. DOI: https://doi.org/10.1121/1.4986649

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements. *Journal of Speech, Language, and Hearing Research, 56*(3), 850–864. DOI: https://doi.org/10.1044/1092-4388(2012/12-0049)

Fink, P. W., Kelso, J. A. S., Jirsa, V. K., & de Guzman, G. (2000). Recruitment of degrees of freedom stabilizes coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 671–692. https://doi-org.udel.idm.oclc.org/10.1037/0096-1523.26.2.671. DOI: https://doi.org/10.1037/0096-1523.26.2.671

Fiveash, A., Schön, D., Canette, L.-H., Morillon, B., Bedoin, N., & Tillmann, B. (2020). A stimulus-brain coupling analysis of regular and irregular rhythms in adults with dyslexia and controls. *Brain and Cognition, 140*, 105531. DOI: https://doi.org/10.1016/j.bandc.2020.105531

Flege, J. E. (1988). Effects of speaking rate on tongue position and velocity of movement in vowel production. *The Journal of the Acoustical Society of America*, *84*(3), 901–916. DOI: https://doi.org/10.1121/1.396659

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135. DOI: https://doi.org/10.1006/jpho.2000.0114

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3739. DOI: https://doi.org/10.1121/1.418332

Franich, K. (2016). Internal and contextual cues to tone perception in Med\textbaru mba. *Journal of the Acoustical Society of America Express Letters, 140*(1), 107–112. DOI: https://doi.org/10.1121/1.4955003

Franich, K. H. (2018a). *The Interaction of Prominence, Rhythm, and Tone in Medumba* (Doctoral dissertation). Univeristy of Chicago. DOI: https://doi.org/10.6082/M14X55ZW

Franich, K. (2018b). Tonal and morphophonological effects on the location of perceptual centers (p-centers): Evidence from a Bantu language. *Journal of Phonetics*, *67*, 21–33. DOI: https://doi.org/10.1016/j.wocn.2017.11.001

Franich, K. (2019). Uncovering Tonal and Temporal Correlates of Phrasal Prominence in Medumba. *Language and Speech*, 23830919887994. DOI: https://doi.org/10.1177/0023830919887994

Franich, K. (2021). Metrical prominence asymmetries in Medumba, a Grassfields Bantu language: Supplementary material. *Language*, *97*(2). DOI: https://doi.org/10.1353/lan.2021.0033

Franich, K. & Keupdjio, H. (2022). The influence of tone on the alignment of speech and co-speech gesture. *Proceedings of Speech Prosody*, Lisbon, Portugal. DOI: https://doi.org/10.21437/SpeechProsody.2022-63

Franich, K. H., & Lendja Ngnemzué, A. B. (2021). Feeling the Beat in an African Tone Language: Rhythmic Mapping Between Language and Music. *Frontiers in Communication, 6*, 653747. DOI: https://doi.org/10.3389/fcomm.2021.653747

Fry, D. B. (1955). Duration and Intensity as Physical Correlates of Linguistic Stress. *The Journal of the Acoustical Society of America*, *27*(4), 765–768. DOI: https://doi.org/10.1121/1.1908022

Fry, D. B. (1958). Experiments in the Perception of Stress. *Language and Speech*, *1*(2), 126–152. DOI: https://doi.org/10.1177/002383095800100207

Fuchs, A., & Scott Kelso, J. A. (2018). Coordination Dynamics and Synergetics: From Finger Movements to Brain Patterns and Ballet Dancing. In S. C. Müller, P. J. Plath, G. Radons & A. Fuchs (Eds.), *Complexity and Synergetics* (pp. 301–316). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-64334-2_23

Fujimura, O. (1990). Methods and Goals of Speech Production Research. *Language and Speech*, *33*(3), 195–258. DOI: https://doi.org/10.1177/002383099003300301

Garvin, K. (2021). *Word-medial syllabification and gestural coordination* (Doctoral dissertation). University of California at Berkeley.

Gay, T., & Hirose, H. (1973). Effect of Speaking Rate on Labial Consonant Production. *Phonetica*, *27*(1), 44–56. DOI: https://doi.org/10.1159/000259425

Goldstein, L. M., Saltzman, E., Chitoran, I. & Nam, H. (2009). Coupled oscillator model of speech timing and syllable structure. In G. Fant, H. Fujisaki, & J. Shen: F*rontiers in Phonetics and Speech Science*. Beijing: The Commercial Press, 239–250.

Gordon, M. (2004). A Phonological and Phonetic Study of Word-Level Stress in Chickasaw. *International Journal of American Linguistics*, *70*(1), 1–32. DOI: https://doi.org/10.1086/422264

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0007

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7*. De Gruyter. DOI: https://doi.org/10.1515/9783110197105.515

Gussenhoven, C. (2021). Just how metrical is the Autosegmental- Metrical model? Evidence from sentential pitch accents in Nubi, Persian, and English. In H. Kubozono, J. Ito & A. Mester (Eds.), *Prosody and Prosodic Interfaces*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198869740.003.0006

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*, 347–356. DOI: https://doi.org/10.1007/BF00336922

Hall, D. C. (2011). Phonological contrast and its phonetic enhancement: Dispersedness without dispersion. *Phonology*, *28*(1), 1–54. DOI: https://doi.org/10.1017/S0952675711000029

Harrington, J., Fletcher, J., & Beckman, M. (2000). Manner and place conflicts in the articulation of accent in Australian English. In M. Broe & J. Pierrehumber (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 40–51). Cambridge: Cambridge University Press.

Harrington, J., Palethorpe, S., Fletcher, J., & Beckman, M. E. (1996). Competing hypotheses concerning the articulation of stress in English. *The Journal of the Acoustical Society of America, 99*(4), 2494–2500. DOI: https://doi.org/10.1121/1.415636

Haugen, M., & Godøy, R. I. (2014). Rhythmical Structures in Music and Body Movement in Samba Performance. *International Conference on Music Perception and Cognition.*

Hayes, B. (1995). Metrical Stress Theory: Principles and Case Studies. University of Chicago Press.

Hirano, M., Ohala, J., & Vennard, W. (1969). The Function of Laryngeal Muscles in Regulating Fundamental Frequency and Intensity of Phonation. *Journal of Speech and Hearing Research*, *12*(3), 616–628. DOI: https://doi.org/10.1044/jshr.1203.616

Hualde, J. I., Lujanbio, O., & Torreira, F. (2008). Lexical tone and stress in Goizueta Basque. *Journal of the International Phonetic Association*, *38*(01). DOI: https://doi.org/10.1017/S0025100308003241

Huber, J. E., & Chandrasekaran, B. (2006). Effects of Increasing Sound Pressure Level on Lip and Jaw Movement Parameters and Consistency in Young Adults. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1368–1379. DOI: https://doi.org/10.1044/1092-4388(2006/098)

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, *106*(3), 1532–1542. DOI: https://doi.org/10.1121/1.427150

Hyman, L. M. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic Studies offered to Joseph Greenberg* (vol. 2, pp. 407–418). Anma Libri: Saratoga, CA.

Hyman, L. M., Rolle, N., Sande, H., Chen, E., Jenks, P., Lionnet, F., Merrill, J., & Baier, N. (2019). Niger-Congo linguistic features and typology. In E. Wolff (Ed.), *The Cambridge Handbook of African Linguistics and A History of African Linguistics*. Cambridge University Press. DOI: https://doi.org/10.1017/9781108283991.009

Jirsa, V. K., Fink, P., Foo, P., & Kelso, J. A. S. (2000). Parametric stabilization of biological coordination: a theoretical model. *Journal of Biological Physics*, *26*(2), 85–112. DOI: https://doi.org/10.1023/A:1005208122449

Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics*, *44*, 62–82. DOI: https://doi.org/10.1016/j.wocn.2014.03.003

Keating, P., Cho, T., Fougeron, C. & Hsu, C. (1999) Domain-initial articulatory strengthening in four languages. *UCLA Working Papers in Linguistics, 97*, 139–156.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2003). Domain-initial strengthening in four languages. In R. Local, J. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology* (pp. 143–161). Cambridge University Press.

Keating, P. A., Lindblom, B., Lubker, J., & Kreiman, J. (1994). Variability in jaw height for segments in English and Swedish VCVs. *Journal of Phonetics*, *22*(4), 407–422. DOI: https://doi.org/10.1016/S0095-4470(19)30293-1

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology*, *15*, R1000–R1004. DOI: https://doi.org/10.1152/ajpregu.1984.246.6.R1000

Kelso, J. A. S. (1994). Elementary Coordination Dynamics. In *Interlimb Coordination* (pp. 301–318). Elsevier. DOI: https://doi.org/10.1016/B978-0-12-679270-6.50020-6

Kelso, J. S., Southard, D. L., & Goodman, D. (1979). On the coordination of two-handed movements. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(2), 229–238. DOI: https://doi.org/10.1037/0096-1523.5.2.229

Kelso, Delcolle, J., & Schoner, G. (1990). Action-perception as a pattern-formation process. *Attention and Performance, 13*, 139–169. DOI: https://doi.org/10.4324/9780203772010-5

Kendon, A. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. In M. R. Key, *The Relationship of Verbal and Nonverbal Communication* (pp. 207–228). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110813098.207

Kent, R. D., & Moll, K. L. (1972). Cinefluorographic Analyses of Selected Lingual Consonants. *Journal of Speech and Hearing Research*, *15*(3), 453–473. DOI: https://doi.org/10.1044/jshr.1503.453

Kilchenmann, L., & Senn, O. (2015). Microtiming in Swing and Funk affects the body movement behavior of music expert listeners. *Frontiers in Psychology, 6,* Article 1232. DOI: https://doi.org/10.3389/fpsyg.2015.01232

Kleber, F. & Klipphahn, N. 2006. An acoustic investigation of secondary stress in German. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, *37*, 1–18.

Krakow, R. A. (1993). Nonsegmental influences on velum movement patterns: Syllables, sentences, stress, and speaking rate. In M. A. Huffman & R. A. Krakow (Eds.), *Nasals, Nasalization and the Velum (Phonetics and Phonology V)* (pp. 87–116). New York: Academic Press. DOI: https://doi.org/10.1016/B978-0-12-360380-7.50008-9

Krivokapić, J., Tiede, M., Tyrone, M., & Goldenberg, D. (2016). Speech and manual gesture coordination in a pointing task. *Proceedings of Speech Prosody*, 1240–1244. DOI: https://doi.org/10.21437/SpeechProsody.2016-255

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

Kubik, G. (2010). *Theory of African music (Vol. 2). Chicago studies in ethnomusicology*. Chicago, IL: The University of Chicago Press.

Kudo, K., Park, H., Kay, B. A., & Turvey, M. T. (2006). Environmental coupling modulates the attractors of rhythmic coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 599–609. DOI: https://doi.org/10.1037/0096-1523.32.3.599

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, *4*(4), 303–320. DOI: https://doi.org/10.1016/S0095-4470(19)31257-4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). DOI: https://doi.org/10.18637/jss.v082.i13

Ladd, D. R. (1996). *Intonational phonology*. Cambridge Studies in Linguistics 79. Cambridge: Cambridge University Press. Pp. xv + 334.

Ladefoged, P. 1967. Three areas of phonetics. Part 2: The nature of vowel quality. Oxford: University Press, 50–142.

Ladefoged, P., & McKinney, N. P. (1963). Loudness, sound pressure, and subglottal pressure in speech. *Journal of the Acoustical Society of America, 35*(4), 454–460. DOI: https://doi.org/10.1121/1.1918503

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Lenc, T., Keller, P. E., Varlet, M., & Nozaradan, S. (2018). Neural tracking of the musical beat is enhanced by low-frequency sounds. *Proceedings of the National Academy of Sciences*, *115*(32), 8221–8226. DOI: https://doi.org/10.1073/pnas.1801421115

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

Liberman, M. (1975). *The intonational system of English* (Doctoral dissertation). MIT.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*, 249–336.

Lieberman, P. (1960). Some Acoustic Correlates of Word Stress in American English. *The Journal of the Acoustical Society of America*, *32*(4), 451–454. DOI: https://doi.org/10.1121/1.1908095

Liénard, J.-S., & Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, *106*(1), 411–422. DOI: https://doi.org/10.1121/1.428140

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. DOI: https://doi.org/10.1121/1.1918816

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H & H theory, In W. J. Hardcastle & A. Marchal: *Speech Production and Speech Modeling*, 403-439, Kluwer Academic Publishers, Dordrecht. DOI: https://doi.org/10.1007/978-94-009-2037-8_16

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *The Journal of the Acoustical Society of America*, *50*(4B), 1166–1179. DOI: https://doi.org/10.1121/1.1912750

Linville, R. N. (1982). *Temporal aspects of articulation: Some implications for speech motor control of stereotyped productions* (Doctoral dissertation). University of Iowa, Department of Speech Pathology and Audiology.

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, *3*(1). DOI: https://doi.org/10.1515/lp-2012-0006

Lunden, A., Campbell, J., Hutchens, M., & Kalivoda, N. (2017). Vowel-length contrasts and phonetic cues to stress: An investigation of their relation. *Phonology*, *34*(3), 565–580. DOI: https://doi.org/10.1017/S0952675717000288

Mainka, S., & Mallien, G. (2014). Rhythmic speech cueing (RSC). In M. H. Thaut & V. Hoemberg (Eds.), Handbook of neurologic music therapy (pp. 150–160). Oxford University Press.

McClean, M. D., & Tasko, S. M. (2003). Association of Orofacial Muscle Activity and Movement During Changes in Speech Rate and Intensity. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1387–1400. DOI: https://doi.org/10.1044/1092-4388(2003/108)

Mefferd A. S. (2017). Tongue- and jaw-specific contributions to acoustic vowel contrast changes in the diphthong /ai/ in response to slow, loud, and clear speech. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3144–3158. DOI: https://doi.org/10.1044/2017_JSLHR-S-17-0114

Michael, L. (2011). The interaction of tone and stress in the prosodic system of Iquito (Zaparoan, Peru). *Amerindia*, *36*, 53–74. DOI: https://doi.org/10.5070/P70PW752Z8

Mooshammer, C., & Fuchs, S. (2002). Stress distinction in German: Simulating kinematic parameters of tongue-tip gestures. *Journal of Phonetics*, *30*(3), 337–355. DOI: https://doi.org/10.1006/jpho.2001.0159

Morgan, T. A., Janda, R. D. (1989). Musically-conditioned stress shift in Spanish revisited: Empirical verification and nonlinear analysis. In C. Kirschner & J. A. DeCesaris (Eds.), *Studies in Romance Linguistics, Selected Proceedings from the XVII Linguistic Symposium on Romance Languages*. New Jersey: Rutgers University. DOI: https://doi.org/10.1075/cilt.60.18mor

Morrison, S. J., Demorest, S. M., & Stambaugh, L. A. (2008). Enculturation Effects in Music Cognition: The Role of Age and Music Complexity. *Journal of Research in Music Education*, *56*(2), 118–129. DOI: https://doi.org/10.1177/0022429408322854

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-Centers). *Psychological Review*, *83*, 405–408. DOI: https://doi.org/10.1037/0033-295X.83.5.405

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287. DOI: https://doi.org/10.1121/1.383737

Nozaradan, S., Peretz, I., & Keller, P. E. (2016). Individual Differences in Rhythmic Cortical Entrainment Correlate with Predictive Behavior in Sensorimotor Synchronization. *Scientific Reports*, *6*(1), 20612. DOI: https://doi.org/10.1038/srep20612

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective Neuronal Entrainment to the Beat and Meter Embedded in a Musical Rhythm. *Journal of Neuroscience*, *32*(49), 17572–17581. DOI: https://doi.org/10.1523/JNEUROSCI.3203-12.2012

Nozaradan, S., Zerouali, Y., Peretz, I., & Mouraux, A. (2015). Capturing with EEG the Neural Entrainment and Coupling Underlying Sensorimotor Synchronization to the Beat. *Cerebral Cortex*, *25*(3), 736–747. DOI: https://doi.org/10.1093/cercor/bht261

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.). *Proceedings of the XIVth International Congress of Phonetic Sciences* (Vol. 2, pp. 1075–1078). New York: American Institute of Physics.

Ortega–Llebaría, M., & Prieto, P. 2007. Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. In P. Prieto, J. Mascaró & M. J. Solé (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 155–175). Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/cilt.282.11ort

Ortega-Llebaria, M., & Prieto, P. (2011). Acoustic Correlates of Stress in Central Catalan and Castilian Spanish. *Language and Speech*, *54*(1), 73–97. DOI: https://doi.org/10.1177/0023830910388014

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America*, *77*(2), 640–648. DOI: https://doi.org/10.1121/1.391882

Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2016). Is seeing gesture necessary to gesture like a native speaker? *Psychological Science*, *27*(5), 737–747. DOI: https://doi.org/10.1177/0956797616629931

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

Pellecchia, G. L., Shockley, K., & Turvey, M. T. (2005). Concurrent cognitive task modulates coordination dynamics. *Cognitive Science*, *29*, 531–557. DOI: https://doi.org/10.1207/s15516709cog0000_12

Peper, C., de Boer, B., de Poel, H., & Beek, P. (2008). Interlimb coupling strength scales with movement amplitude. *Neuroscience Letters*, *437*, 10–14. DOI: https://doi.org/10.1016/j.neulet.2008.03.066

Pike, K. (1945). *The intonation of American English*. University of Michigan Press.

Polak, R. (2010). Rhythmic feel as meter: Non-isochronous beat subdivision in jembe music from Mali. *Music Theory Online*, *16*(4). DOI: https://doi.org/10.30535/mto.16.4.4

Polak, R., & London, J. (2014). Timing and meter in Mande drumming from Mali. *Music Theory Online*, 20(1). DOI: https://doi.org/10.30535/mto.20.1.1

Prieto, P., Vanrell, M. del M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, *54*(6), 681–702. DOI: https://doi.org/10.1016/j.specom.2011.12.001

Rathcke, T., Lin, C.-Y., Falk, S., & Dalla Bella, S. (2021). Tapping into linguistic rhythm. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *12*(1), 11. DOI: https://doi.org/10.5334/labphon.248

Remijsen, B. (2002). Lexically contrastive stress accent and lexical tone in Ma'ya. Laboratory Phonology 7, ed . Carlos Gussenhoven & Natasha Warner, pp . 585–614. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110197105.2.585

Remijsen, B. & van Heuven, V. (2005). Stress, tone and discourse prominence in the Curaçao dialect of Papiamentu. *Phonology, 22*, 205–35. DOI: https://doi.org/10.1017/S0952675705000540

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology, 28*(2), 161–169.e5. DOI: https://doi.org/10.1016/j.cub.2017.11.033

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The Speech Focus Position Effect on Jaw–Finger Coordination in a Pointing Task. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1507–1521. DOI: https://doi.org/10.1044/1092-4388(2008/07-0173)

Rosenfelder, I., Fruehwald, J., Keelan Evanini, Seyfarth, S., Gorman, K., Prichard, H., & Jiahong Yuan. (2015). *Fave: Speaker Fix*. Zenodo. DOI: https://doi.org/10.5281/ZENODO.22281

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of Prosody and Position on the Timing of Deictic Gestures. *Journal of Speech, Language, and Hearing Research*, *56*(2), 458–470. DOI: https://doi.org/10.1044/1092-4388(2012/11-0283)

Ryu, Y. U., & Buchanan, J. J. (2004). Amplitude scaling in a bimanual circle-drawing task: Pattern switching and end-effector variability. *Journal of Motor Behavior*, *36*(3), 265–279. DOI: https://doi.org/10.3200/JMBR.36.3.265-279

Saltzman, E., Nam, H., Krivokapić, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the Speech Prosody 2008 Conference*, eds P. A. Barbosa, S. Madureira, & C. Reis, (Campinas: International Speech Communications Association), 175–184.

Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 227–247. DOI: https://doi.org/10.1037/0096-1523.16.2.227

Schöner, G. (1994). From Interlimb Coordination to Trajectory Formation. In *Interlimb Coordination* (pp. 339–368). Elsevier. DOI: https://doi.org/10.1016/B978-0-12-679270-6.50022-X

Schöner, G., Haken, H., & Kelso, J. A. S. (1986). A stochastic theory of phase transitions in human hand movement. *Biological Cybernetics, 53*(4), 247–257. DOI: https://doi.org/10.1007/BF00336995

Schwartz, M., Amazeen, E. L., & Turvey, M. T. (1995). Superimposition in interlimb coordination. *Human Movement Science*, 14, 681–694. DOI: https://doi.org/10.1016/0167-9457(95)00033-X

Scott, S. K. (1993). *P-centres in Speech: An Acoustic Analysis* (Doctoral dissertation). University College London.

Shattuck-Hufnagel, S., & Ren, A. (2018). The Prosodic Characteristics of Non-referential Co-speech Gestures in a Sample of Academic-Lecture-Style Speech. *Frontiers in Psychology*, *9*, 1514. DOI: https://doi.org/10.3389/fpsyg.2018.01514

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471–2485. DOI: https://doi.org/10.1121/1.417955

Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America*, *101*(1), 503–513. DOI: https://doi.org/10.1121/1.417994

Smith, J. L. (2002). Phonological augmentation in prominent positions. *Doctoral Dissertations Available from Proquest*. AAI3056278. https://scholarworks.umass.edu/dissertations/AAI3056278

Smith, J. L. (2004). *Phonological Augmentation in Prominent Positions* (1st ed.). Routledge. DOI: https://doi.org/10.4324/9780203506394

Soley, G., & Hannon, E. E. (2010). Infants prefer the musical meter of their own culture: A cross-cultural comparison. *Developmental Psychology*, *46*(1), 286–292. DOI: https://doi.org/10.1037/a0017555

Sonoda, Y. (1987). Effect of speaking rate on articulatory dynamics and motor event. *Journal of Phonetics*, *15*(2), 145–156. DOI: https://doi.org/10.1016/S0095-4470(19)30554-6

Stupacher, J., Wood, G., & Witte, M. (2017). Neural Entrainment to Polyrhythms: A Comparison of Musicians and Non-musicians. *Frontiers in Neuroscience*, *11*. DOI: https://doi.org/10.3389/fnins.2017.00208

Tajima, K. (1998). *Speech rhythm in English and Japanese: Experiments in Speech Cycling* (Doctoral dissertation). Indiana University.

Tajima, K., & Port, R. F. (2003). Speech rhythm in English and Japanese. In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI* (pp. 317–334). Cambridge University Press.

Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., & Zion Golumbic, E. (2017). Neural Entrainment to the Beat: The "Missing-Pulse" Phenomenon. *The Journal of Neuroscience*, *37*(26), 6331–6341. DOI: https://doi.org/10.1523/JNEUROSCI.2500-16.2017

Tallman, A. J. R., & Elías-Ulloa, J. (2020). The acoustic correlates of stress and tone in Chácobo (Pano): A production study. *The Journal of the Acoustical Society of America*, *147*(4), 3028–3042. DOI: https://doi.org/10.1121/10.0001014

Temperley, N., & Temperley, D. (2012). Stress-meter alignment in French vocal music. *Journal of the Acoustical Society of America*, *134*(1), 520–527. DOI: https://doi.org/10.1121/1.4807566

Tilsen, S. (2016). *A shared control parameter for F0 and intensity*, 1066–1070. DOI: https://doi.org/10.21437/SpeechProsody.2016-219

Tilsen, S. (2009). Multitimescale Dynamical Interactions Between Speech Rhythm and Gesture. *Cognitive Science*, *33*(5), 839–879. DOI: https://doi.org/10.1111/j.1551-6709.2009.01037.x

Titze, I. R. (2000). *Principles of voice production*. Iowa City, IA: National Center for Voice and Speech.

Titze, I. R., & Sundberg, J. (1992). Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America, 91*, 2936–2946. DOI: https://doi.org/10.1121/1.402929

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451. DOI: https://doi.org/10.1121/1.429414

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440. DOI: https://doi.org/10.1006/jpho.2000.0123

Van Summers, W. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, *82*(3), 847–863. DOI: https://doi.org/10.1121/1.395284

Vogel, I., Athanasopoulou, A., & Pincus, N. (2016). Prominence, Contrast, and the Functional Load Hypothesis: An Acoustic Investigation. In J. Heinz, R. Goedemans & H. van der Hulst (Eds.), *Dimensions of Phonological Stress* (pp. 123–167). Cambridge University Press. DOI: https://doi.org/10.1017/9781316212745.006

von Holst, E. (1973). *The Behavioural Physiology of Animals and Man: The Collected Papers of Eric von Holst*. Coral Gables, FL: University of Miami Press.

Wei, C. (2006). Not Crazy, Just Talking On The Phone: Gestures And Mobile Phone Conversations. *2006 IEEE International Professional Communication Conference*, 299–307. DOI: https://doi.org/10.1109/IPCC.2006.320363

Yu, A. C. L. (2010). Tonal effects on perceived vowel duration. *Laborabory Phonology 10*, 151–168.

Zhang, Z. (2016). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of the Acoustical Society of America*, *139*, 1493–1507. DOI: https://doi.org/10.1121/1.4944754

Zvonik, E., & Cummins, F. (2002). Pause duration and variability in read texts. In *Proceedings of the International Congress of Spoken Language Processing*, Denver, CO, pp. 1109. DOI: https://doi.org/10.21437/ICSLP.2002-367