Hierarchical timing in varieties of Kuwaiti Arabic

Saleh Ghadanfari

A thesis submitted in fulfilment of the requirements for the degree of Doctor in speech and language sciences

Newcastle university

School of education communication and language sciences

October, 2022

The work embodied in the thesis is my own work. No material in this thesis has been submitted for any other degree or professional qualification and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made.

Saleh Ghadanfari

Abstract

We examine the differences in speech timing between Hadari and Bedouin Kuwaiti Arabic dialects using the speech cycling task. In this task, wherein speakers repeat phrases in time with metronome beats, stressed vowel onsets tend to lie at harmonic phases within the Phrase Repetition Cycle (PRC), e.g., 1/2, reflecting coordination between prosodic units. Hadari has stronger stress contrast, with greater unstressed syllable reduction than Bedouin, which may afford closer alignment to harmonic phases.

Six trochaic and six iambic sentences were read aloud by 22 Hadari and 23 Bedouin speakers at three metronome trial rates: slow, medium, and fast. The phases of the final (external) and medial (internal) stressed syllables – heavy, CVV(C), or light, CVC – relative to the PRC were analysed.

In external and internal phases, both dialects tended to align heavy syllables closer to 1/2 phase than light syllables; however, Hadari aligned light syllables earlier in the PRC than Bedouin, which syllable duration analysis suggested may be due to greater shortening of preceding unstressed syllables in Hadari.

According to spectral balance measures, Hadari contrasted three degrees of positional prominence: initial vs medial vs final, while Bedouin contrasted phrase-initial vs non-phrase-initial, suggesting different metrical structures of stress beats between dialects.

We investigated the timing interaction between syllables and feet in the speech cycling data. We analysed the amplitude envelope modulation rates at foot (2-4 Hz) and syllable (4.5-12 Hz) levels, and found no effect of stress feet on the timing of syllables. The amplitude envelope's power spectrum showed higher peaks for stressed syllables than stress feet, with higher power ratio between stress and syllable levels in Hadari than in Bedouin, which we interpreted as higher temporal stress contrast in Hadari.

This work suggests that temporal coordination is an *affordance* of temporal acoustic cues to the interaction between rhythmic time scales.

Acknowledgments

I am very fortunate to have had the chance to work with Laurence White and to learn from his expertise in speech timing and prosody. Laurence provided insightful commentary and discussion to my work. He has contributed a greater deal to my growth as a researcher, and for that, I thank him.

I am grateful to Jalal Al-Tamimi for always encouraging me to learn new techniques in speech analysis. I am indebted to him for sharing his knowledge and expertise.

I would like to thank Ghada Khattab, who was very patient with my progress in my first Ph.D. year.

Thanks to my examiners, Rachel Smith, and Daniel Duncan, for the interesting discussion during my viva and for the helpful comments on my thesis.

I am grateful to Kuwait university for their generous scholarship and funding for my Ph.D. program.

Finally, I would like to thank friends and family members who supported me during my Ph.D. program.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction and literature review

Linguistic structure is hierarchically organized, such that the linguistic constituents have a nesting relation amongst them. For instance, syllables are grouped within a higher-level prosodic constituent, such as the foot, and feet are grouped within phonological words. The hierarchical structure of speech also has a metrical property in terms of relative prominence. For example, within a stress foot, a certain syllable may be more prominent than others, and this prominent syllable constitutes the head of the prosodic foot. Feet can also differ in relative prominence: the foot with the most prominent syllable constitutes the head of the phonological word (Liberman, 1975; Selkirk, 1986; Nespor & Vogel, 1986). This hierarchical metrical structure of speech is said to constrain the timing of speech units. In the speech cycling experimental paradigm (Cummins & Port, 1998; Tajima, 1999), it has been shown that when speakers repeat a phrase with metronomes, vowel onsets of stressed syllables are constrained to lie at simple phases within a limit cycle. These phases are considered metrically important points that divide the cycle into simple integer ratios, such as 1/3, 1/2, and 2/3. The constraints on vowel onsets of stressed syllables to lie at simple phases are considered as evidence for the hierarchical nesting of lower prosodic units, stressed syllables, within a higher prosodic unit, the phrase repetition cycle. Earlier speech cycling experiments showed that the degree of stress contrast that a language has influences the temporal organization of stressed vowels onsets in speech cycling. For instance, Cummins (2002) showed that English speakers were able to align stressed vowel onsets close to simple phases, whereas Italian and Spanish speakers' alignment was not as close as English to simple phases. The greater stress contrast in English than in Spanish and Italian may afford the closer alignment of stressed vowel onsets in English to metrically important points in the repetition cycle.

In this dissertation, we explore dialectal differences in speech cycling between Hadari and Bedouin Kuwaiti Arabic dialects, two understudied Arabic dialects, regarding their prosodic characteristics. Hadari dialect has a greater degree of unstressed syllable reduction and allows for more consonantal clustering than Bedouin (cf. Holes, 2006; Abu Haider, 2006; Ingham, 1994). Thus, we predict that the greater stress contrast in Hadari would afford closer alignment of vowel onsets of stressed syllables to simple phases in speech cycling than in Bedouin. We will also analyse syllable duration in Hadari and in Bedouin, to explain the

potential variable phase alignment between the two dialects. Also, as speech is metrically organized, we will examine metrical structure differences between Hadari and Bedouin, through an analysis of spectral balance (Sluijter & van Hueven, 1996).

The hierarchical organization of prosodic units has been interpreted by some authors as implying an interaction between different prosodic units in speech timing (e.g., O'Dell & Nieminen, 1999). In another analysis of our speech cycling corpus, we will investigate the potential role of the stress foot in the timing of syllables between Hadari and Bedouin dialects. This will be examined through an analysis of a range of amplitude envelope statistics (Tilsen & Arvaniti, 2013). The amplitude envelope contains frequency modulations at the stress foot level, ~2 Hz - ~4 Hz, and the syllable level, ~5 Hz - ~ 12 Hz. By analysing metrics that quantify the interaction between stress foot level and syllable level energy fluctuations, we explore the possible interaction between stress feet and syllable timing.

Temporal coordination patterns in Hadari and Bedouin in speech cycling will be important in predicting potential patterns of *entrainment* between Hadari and Bedouin interlocutors in naturalistic conversational situations (Włodarczak et al., 2012a,b; Wagner, 2019).

All of the analyses in this work were based on a single set of speech cycling data collected from 22 Hadari and 23 Bedouin speakers, with 12 different phrases and three different metronome periods. Each of the four distinct analyses was represented and discussed in separate chapters.

In this chapter, we will provide an overview on models of speech timing and timing variation across languages.

In section 1.1, we will show how the concept of regular occurrences of rhythmic events led to the notion of isochrony as an explanation for speech timing variation across languages.

Section 1.2 reviews experimental work that examined isochrony in speech timing, which showed no supporting evidence for isochrony in speech.

In section 1.3, we will review the notion of contrastive rhythm (cf. Nolan & Jeon, 2014), which refers to the cross-language variation in the degree of temporal contrast between strong

and weak syllables, as a plausible account for speech timing variation between languages. Also, we will discuss the development of temporal rhythm metrics, which capture language variation in syllable complexity and vowel reduction and their relation to temporal stress contrast. Methodological caveats in using these metrics to capture temporal stress contrast between languages will also be discussed.

As this paper focuses on the prosody of Arabic dialects, particularly Kuwaiti Arabic dialects, we will review in section 1.4 the phonological structural aspects across Arabic dialects, and their effects on the degree of temporal stress contrast.

Section 1.5 discusses theoretical models on the relation between timing variation and the hierarchical structure of speech, contrasting between the coupled oscillators model (O'Dell & Nieminen, 1999) and the locus and domain view (White, 2002, 2014). We will attempt to provide compatible aspects between the two views. In particular, we will show how *temporal coordination* between different rhythmic time scales may be mediated by structural aspects of the speech signal, that refer to local prominence structures. Other important terms in temporal coordination, such as phase relations and perceptual centres will be reviewed.

In section 1.6, we will provide an overview of the speech cycling task and its use in reflecting cross-linguistic variation in temporal coordination based on the degree of temporal stress contrast.

1.1 General background on the concept of rhythm

Rhythm is a temporal phenomenon. It refers to the regular occurrence of events in time (Eriksson, 1991, p. 6-7). Rhythmic instances can vary in nature. They can either be auditory in nature, such as the perception of regularly occurring acoustic events, or visual, such as the regular movement of a pendulum or a bird's wings, which can visually be perceived as rhythmic. Auditory rhythmic processes have certain distinguishable characteristics. First, auditory sequences are perceived with reference to a privileged point recurring in time, which can be the onset of the event. Other rhythmic instances do not necessarily have this feature; the movement of a pendulum can be perceived as rhythmic without referring to a certain point in time. Second, auditory processing tends to structure the series of rhythmic events into groups, at least at certain rates of occurrences. At certain rates, usually normal rates,

rhythmic events might be perceived as individual discrete events recurring regularly. However, at other rates, perhaps faster rates, the series of events is perceived as a structure of discrete groups that recur regularly. It is a ubiquitous feature of the auditory grouping process that some events within rhythmic groups are perceived as more salient than others (Fraisse, 1956; Eriksson, 1991, p. 9; White, 2002, p. 55). For example, the first occurring event within a group can be more prominent than others, leading the successive occurrences of rhythmic groups. The tendency of the auditory processing to posit discrete groups (structures) of events might have to do with the human short-term memory, which creates a "temporal field" in which events are structured in a series of groups (Fraisse, 1978, 1982). This "temporal field" of immediate rhythmic perception is called by psychologists "the psychological present" (Fraisse, 1978). It is useful to quote Woodrow's (1951, p. 1232) statement that describes the grouping of rhythmic events by the auditory process:

"By rhythm, in the psychological sense, is meant the perception of a series of stimuli as a series of groups. The successive groups are ordinarily of similar pattern and experienced as repetitive. Each group is perceived as a whole and therefore has a length lying within the psychological present."

The tendency of the auditory system to group rhythmic events and to assign higher prominence levels for certain events within a group seem to have influenced the human production system of acoustic events (Allen, 1975). For instance, musical rhythms, such as waltz or tango rhythms, are produced in a regular fashion with a succession of groups of tones that have a leading, salient head tone.

The idea of prominence-headed groups that recur regularly was the bases for accounting for timing variation in speech by several phoneticians (for example, Classe, 1939; Allen, 1975; Lehiste, 1977). In speech, certain syllables are more prominent than others, for instance, stressed vs unstressed syllables. The rhythmic foot, which plays an important role in metrical theory, is said to be a timing unit that recurs regularly in an utterance of speech. There are different structures of the metrical foot that depends on headedness and the number of syllables (cf. Hayes, 1985). For instance, trochaic feet are left-headed and comprised of two syllables, while dactyls are left-headed and consist of three syllables. Iambs are right-headed and contain two syllables while anapaests are right-headed and contain three syllables. However, a rhythmical foot that spans a stressed syllable to the following stressed syllable,

including all unstressed syllables in between, has been considered in speech research as the unit for regularity in an utterance of speech, and it is also called inter-stress interval (Lehiste, 1977; Dauer, 1983).

The notion of regular occurrences of rhythmical feet is known as isochrony: the idea that rhythmical feet have equal durations in an utterance of speech at a constant rate. Below, we will review the historical development of isochrony and the experimental work that surrounded it.

1.2 Historical overview of isochrony concept

The observation of isochronous inter-stress intervals was first made by the phonetician Joshua Steele (1779). It was based on an impression that inter-stress intervals within English utterances, notated similarly to musical bars, such as 2/4 and 3/4, occurred at equal durational intervals. Later in the 20th century, the work of Lloyd James (1929) implied another timing unit that might be produced as isochronous, that is, the syllable. Lloyd James (1929) used terms that refer to equal syllable durations in describing the rhythm of languages other than English, such as French and Indian. In his work in 1940, he became the first (as far as we are aware) to use the well-known terms "Morse-code rhythm" and "Machine-gun rhythm" in describing the rhythms of different languages. He considers English and Arabic as examples of languages with "Morse-code rhythm", and French and Telugu as examples of "Machine-gun rhythm".

Pike (1946) introduced the terms "stress-timing" and "syllable-timing" to characterise the rhythms of different languages: those with a tendency to produce regular inter-stress intervals, such as English and Arabic, are "stress-timed" languages, and those that tend to produce regular syllables such as French and Spanish are "syllable-timed" languages. Abercrombie (1967) introduced the "rhythm class typology", which states that spoken languages fall either under the "stress-timing" group, such as English, Arabic, and Thai or under the "syllable-timing" group, such as Spanish and Italian. Ladefoged (1976) suggested isochronous moras for Japanese. Moras are subsyllabic units and are sensitive to the syllable rhyme: syllables that have long rhymes, such as long vowels or a vowel and a coda consonant, are said to be bimoraic, while those with short rhymes are monomoraic. The suggestion of moras as timing units in Japanese comes from the role that moras play in

Japanese poetry; just as the number of syllables is important for Spanish poetry and feet structure for English poetry, so as moras for Japanese poetry.

In stress-timed languages, Pike (1946) suggested that as the number of unstressed syllables increases in an inter-stress interval, the timing mechanism involved in keeping durations of inter-stress intervals equal is the shortening of unstressed syllables. Other suggested mechanisms involve the shortening of stressed syllables (Jones, 1942). Abercrombie (1967) asserted that the tendency to produce equal inter-stress intervals in stress-timed languages and equal syllables in syllable-timed languages originates from specific respiratory processes referred to as chest pulses and stress pulses. Chest pulses are puffs of air that are resulted from muscle contraction and relaxation in the lungs and are responsible for producing syllables. Stress pulses are intense chest pulses, that result in more prominent syllables, specifically stressed syllables. In syllable-timed languages, chest pulses are produced in an isochronous fashion, while in stress-timed languages, stress-pulses are produced isochronously. In the next section, we will review experimental work that attempted to test isochrony in inter-stress intervals, syllables, and moras.

1.2.1 Experimental work on isochrony

The work of Classe (1939) was one of the first attempts to test isochrony in inter-stress intervals experimentally. He used the kymograph to measure durational patterns in multiple English phrases read by 13 speakers. He finds that speech was rather irregular due to variation in the phonetic structure of stress groups, such as the number of unstressed syllables, and due to variation in the grammatical structure. Classe, however, maintained that isochrony is an underlying principle of speech production in English.

Shen and Peterson (1962) examined inter-stress durations in English prose. They investigated two types of inter-stress intervals; in the first type, inter-stress intervals were headed by primary stressed syllables, and in the second type, inter-stress intervals were headed by secondary stressed syllables. They found that durations of inter-stress intervals of both types were variable and concluded that isochrony is not supported. Similarly, Bolinger (1965) measured the duration of pitch accented inter-stress intervals in read English sentences. He found significant durational variation between inter-stress intervals. He concluded that the

number of syllables in an inter-stress interval is the main factor affecting the duration of an inter-stress interval.

The effect of the number of syllables within inter-stress intervals on inter-stress interval duration was examined in Lea (1974) and Faure et al. (1980). Lea (1974) explored the effect of the number of unstressed syllables intervening between the onset of stressed syllables on inter-stress durations. The major finding was that the duration of inter-stress intervals increased linearly as a function of the number of unstressed syllables intervening between two stressed syllables. Faure et al. (1980) examined the effect of the number of syllables, including stressed syllables, on inter-stress intervals, in the reading of two English speakers. Similar to Lea's findings, the duration of inter-stress intervals increased with the number of syllables within the interval, and thus, isochrony was not supported.

Several studies explored the possible regular durations of syllables. Pointon (1980) reviewed multiple studies that examined syllabic duration in Spanish as a function of syllable structure, stress, and phrasal position. There was noticeable variation in syllable duration due to the aforementioned factors, and Pointon (1980) concluded that the traditional classification of Spanish as a syllable-timed language is not supported. He further concluded that syllable duration is driven by segmental durations, which he terms "segment-timed". Manrique and Signorini (1983) studied syllabic duration in Argentinian Spanish and tested the effect of stress and phrasal position. They assert that syllable-timing in Argentinian Spanish needs to be rejected, as there was considerable variation in syllable duration due to stress and phrasal position. Delattre (1966) reported that the ratio of the duration of stressed to unstressed syllables in syllable-timed French and Spanish are 1.7:1 and 1.3:1, respectively, collapsed over different positions and structures. Different durational ratios of stressed to unstressed syllables refute the idea of equal syllable length in syllable-timed languages.

As for mora-timing in Japanese, Beckman (1982) tested the hypothesis that segment length in Japanese will show temporal compensation for intrinsic durations of adjacent segments in order to keep the length of segments, i.e., moras, equal. However, her findings did not support mora-timing, as no compensatory effects were found. Also, a review by Warner and Arai (1991) on studies of mora-timing in Japanese concluded that mora isochrony could not be supported.

Roach (1982) and Dauer (1983) compared syllabic and inter-stress intervals isochrony between different rhythm classes. Roach (1982) tested Abercrombie's (1967) claims that: (a) in stress-timed languages syllable durations tend to be more variable in an inter-stress group than in syllable-timed languages, (b) stress-pulses in syllable-timed languages are unevenly separated, and (c) in syllable-timed languages the duration of inter-stress intervals tend to positively grow in proportion to the number of syllables in an inter-stress interval, while this tendency might be weak or absent in stress-timed languages. He made recordings of natural speech of a single speaker from three stress-timed languages (English, Russian and Arabic) and a single speaker from three syllable-timed languages (French, Yoruba, and Telugu). In answer to the first claim, Roach (1982) measured the variance of syllables in stress-timed and syllable-timed languages. There were no significant differences between languages in syllables' duration variability, thus, no support for Abercrombie's first claim was found. As for the second claim, Roach measured the deviance from a hypothetical regular inter-stress interval. The duration of each intonational group was measured and divided by the number of inter-stress intervals it contains to obtain a hypothetical measure of a regular inter-stress interval. The percentage of deviation of every actual inter-stress interval from a regular inter-stress was calculated, and then the variance of percentage deviance was calculated for every language. Contrary to the prediction, there was greater variance for stress-timed languages than syllable-timed languages. In testing the third claim, Roach investigated the correlation between the number of syllables in inter-stress intervals and the duration of inter-stress intervals. No significant differences between language classes (syllable-timed vs stress-timed) were found. According to these findings, Roach suggested that languages might sound either syllable-timed or stress-timed based on structural differences, such as syllable complexity and vowel reduction in unstressed syllables.

Dauer (1983) compared inter-stress interval durations in various languages classified as stress-timed or syllable-timed. In Dauer's study, stress-timed languages were English and Thai, and syllable-timed languages were Spanish, Italian, and Greek. From measurements taken from prose text reading, Dauer found no significant differences in mean inter-stress interval durations or in the standard deviation of inter-stress interval durations between all five languages. She also found that the effect of the number of syllables in an inter-stress interval was similar in all languages, such that the duration of inter-stress intervals grew linearly as a function of the number of syllables. Similar to Roach, Dauer concluded that rhythmic differences between languages reside in structural differences such as syllabic

complexity and vowel reduction and their effect on stress assignment. For instance, most of the syllables in Spanish are open syllables (CV), while in English, most syllables are of CVC structure, with up to three consonants in the syllable's onset and four in the coda allowed. Also, in English and other stress-timed languages such as Arabic and Thai, stress assignment is more sensitive to syllable weight, which is related to coda complexity, than Spanish. Furthermore, vowel reduction in unstressed syllables in Spanish is minimal, while in a number of stress-timed languages such as English and Russian, most unstressed syllables contain a reduced vowel. These factors, syllabic complexity and vowel reduction in unstressed syllables, contribute to higher temporal contrast between stressed and unstressed syllables and would play an important role in the percept of cross-linguistic rhythmic differences.

1.2.2 Isochrony as a perceptual phenomenon

Despite the lack of evidence that isochrony is present in the acoustic signal of speech, the idea that prominent speech constituents occur regularly was so appealing that several researchers believed that isochrony is an underlying property of speech (e.g., Classe, 1939). Lehiste (1977) argued that stresses are heard regularly more than they are produced since in production there are multiple constraints on phonemic and syllabic durations that make it difficult to produce regular stress intervals. Lehiste (1977) draws evidence for perceived isochrony from her earlier perceptual experiment (Lehiste, 1972). In that perception study, utterances composed of four inter-stress intervals were played to English participants. The participants failed to identify which of the inter-stress intervals were the longest and which of them were the shortest. Failure to discriminate between different inter-stress interval lengths constitutes evidence that deviation from isochrony is not perceived. Furthermore, in another experiment, Lehiste (1972) synthesized non-speech intervals (clicks separated with noise) and found that participants could distinguish between long and short intervals. The latter finding shows that isochrony is exclusive to speech. However, interpreting listeners' inability to distinguish between long and short inter-stress intervals, as in Lehitse's (1972) experiment, as reflecting underlying isochrony is problematic as another contradictory explanation is possible. Listeners' underestimation of inter-stress interval lengths may indicate that inter-stress intervals are not relevant units of speech, and therefore, listeners would not pay attention to their durations (White, 2002, p. 60).

A similar assertion of perceived isochrony is made by Darwin and Donovan (1980). In their experiment, participants were presented with speech stimuli and non-speech stimuli, in which the durational intervals were unevenly distributed. They were asked to tap out the rhythmic pattern of both types of stimuli. The findings were that participants produced regular inter-tap intervals to the speech stimuli but not to the non-speech. This indicates that participants perceive isochrony even when it is not present in the physical aspect of speech. Also, the fact that regular inter-taps were found only in the speech stimuli but not in non-speech indicate that isochrony is special to speech.

However, findings from Scott et al. (1985) point to the fact that isochronous inter-taps of speech stimuli, as in Darwin and Donovan's work, might be an artefact of the experimental materials, which has nothing to do with an underlying isochronous structure of speech stimuli. They had English and French participants tap their fingers to English and French stimuli, as well as to non-speech. The English sentences had four stresses that participants needed to tap out. As for the French sentences, although French does not have lexical stress that would impose a certain metrical structure (alternating strong and weak syllables), the French syllables that participants were asked to tap out were separated by roughly the same number of syllables separating stresses in the English sentences. The non-speech stimuli were a series of noise bursts whose rhythm was identical to the test speech sentences.

Their general hypothesis was that if isochrony was an underlying principle of speech, participants of both languages should respond to the English stimuli, a stress-timed language, with regular inter-tap intervals, while regular inter-tap intervals should disappear when responding to the French stimuli, a syllable-timed language.

Their principal findings were that English and French listeners regularized both the English and French stimuli, suggesting that regular inter-tap intervals do not reflect any underlying isochronous structure of stress-timed English and syllable-timed French. Listeners from both languages did not produce regular inter-taps for the none-speech stimuli. Scott et al. (1985) take the latter finding as indicating that responding to the speech stimuli with regular inter-taps but to the non-speech with irregular inter-taps is related to the complexity of the stimuli. Speech stimuli are more complex than non-speech, and listeners might tend to respond to the complex stimuli with regular inter-taps. To further support this interpretation, Scott et al. conducted another experiment examining the responses of English listeners to speech stimuli

and to two sets of non-speech. The first set of non-speech stimuli was similar to that in the first experiment, which was made of noise bursts. The second set of the non-speech was made by degrading the segmental acoustic information of the speech utterances outside the target stressed syllables. The degraded speech sample was unintelligible but was similar to the speech sample in the acoustic complexity of the target syllable. The hypothesis was that if listeners produced regular inter-taps for the speech stimuli and the unintelligible speech stimuli but not for the non-speech, this would that the complexity of the stimuli is the main drive for producing regular inter-taps. If, on the other hand, listeners responded to the unintelligible speech and the non-speech the same way, i.e., with irregular intervals, this would mean that producing regular inter-taps is special to language.

Listeners produced regular inter-taps for the speech and the unintelligible speech stimuli but irregular inter-taps for the non-speech, which supports the hypothesis that the complexity of the stimuli is the main reason participants produce regular inter-taps. Scott et al. (1985) conclude by saying:

"The results of this experiment show that the phenomenon of regularization is not even specific to speech, but extends to other unintelligible noises with some speech-like properties. They raise the possibility that the subjects are not actually doing anything very interesting at all- that they are simply exhibiting a response bias toward evenly spaced taps when the task becomes difficult." (Scott et al., 1985, p. 161)

Indeed, some findings in Darwin and Donovan (1980) point towards the role of complexity in the stimuli in producing regular responses. Darwin and Donovan examined the listeners' tapping to two sentences, a sentence with one tone group and another with two tones group. The sentence with one tone group was: "Tell the terrified town that TALE" with an intonational boundary occurring with the word "TALE", and the sentence with two tone groups was: "Tell the TOWNFOLK/ a tale of TERROR" with intonational boundaries occurring at the words TOWNFALK and TERROR. Listeners produced regular rhythm for the sentence with one tone group but not for the two tones groups. In his review of Darwin and Donovan findings, Eriksson (1991, p. 59) conjectured that regular taps to the one-tone group sentence could be because of the higher complexity of this sentence. In the one-tone group sentence, listeners would have to tap out a sequence of four syllables in a single group, but in the two-tone group sentence, listeners needed to tap out a sequence of two syllables per

group, which seemed an easier task. Thus, listeners could follow the temporal pattern of the two-tone group sentence more easily, thus producing irregular taps, than in the more complex one-tone group sentence, which led to regular taps. Bell and Fowler (1984) provided a similar explanation to Eriksson's as they reflected on Darwin and Donovan's findings, asserting that responding to speech stimuli with irregular taps, as in the two-tone group sentence, may reflect listeners' ability to follow the variable temporal patterns of the speech stimuli.

In all, the review of perceptual isochrony studies does not support an underlying isochronous structure of speech. Listeners' performance in tapping tasks seems to be driven by the complexity of speech structures in terms of variability in segmental and prosodic temporal patterns. This conclusion is in line with Roach's (1982) and Dauer's (1983) proposals that rhythmic variation in speech is mainly attributed to the degree of variability in phonotactics, and in temporal contrast between strong and weak syllables in the speech signal. In the next section, we will review the role of phonotactics and stress-related temporal variation in accounting for cross-linguistic rhythmic differences.

1.3 Contrastive rhythm and temporal rhythm metrics

Since Roach (1982) and Dauer (1983) proposed that rhythmic structure among languages differs based on language-specific phonotactics and degree of vowel reduction and their relation to stress assignment, research has focused on contrastive rhythm in terms of the degree of temporal contrast between strong and weak syllables (cf. Nolan & Jeon, 2014). Temporal stress contrast has important implications for speech processing, in particular, segmentation of words in forgoing speech stream (Cutler & Norris, 1989).

Research has focused on variability in vocalic and consonantal intervals to quantify the degree of temporal stress contrast between and within languages. Support for investigating vocalic and consonantal variability in studying rhythmic language typology comes from language acquisition studies in infants. Mehler et al. (1996) reviewed research on young infants' ability to discriminate their first language prosodic structure from other languages. They proposed that young infants are sensitive to the temporal contrast in vowels' and consonants' durations in determining their first language prosodic structure.

Therefore, Ramus et al. (1999) proposed metrics that quantify vocalic and consonantal variation in the speech signal of languages that belong to different rhythm classes. They proposed the standard deviation of vocalic and consonantal intervals, ΔV and ΔC, respectively, and %V, which calculates the percentage of the utterance duration taken up by vocalic intervals. Since stress-timed languages demonstrate more variability in syllable structure and have longer vowels than syllable-timed languages, it was expected that stress-timed languages would show greater variability in consonantal and vocalic interval durations than syllable-timed languages. Also, because stress-timed languages exhibit a greater degree of unstressed vowel reduction than syllable-timed languages, vocalic intervals in the former will take up a smaller proportion of the utterance duration.

The standard deviation of consonantal intervals (ΔC) and the proportion of the signal that is vocalic (%V) were the most effective metrics that showed language-specific rhythmic differences. Ramus et al.'s results, illustrated in Figure 1.1, show that prototypical stress-timed languages, like English and Dutch, and prototypical syllable-timed languages, like Spanish and Italian, occupied distinct clusters in the rhythmic space. The results also show a third distinct rhythmic cluster of Japanese, a mora-timed language. Languages like Polish and Catalan, which are said to have mixed rhythms, were categorically grouped with stress-timed languages and syllable-timed languages, respectively. The distinct grouping of languages into three rhythm classes supports the rhythm class hypothesis (Abercrombie, 1967), and the difference in phonological properties is what underlies the classification of languages into distinct rhythm classes.

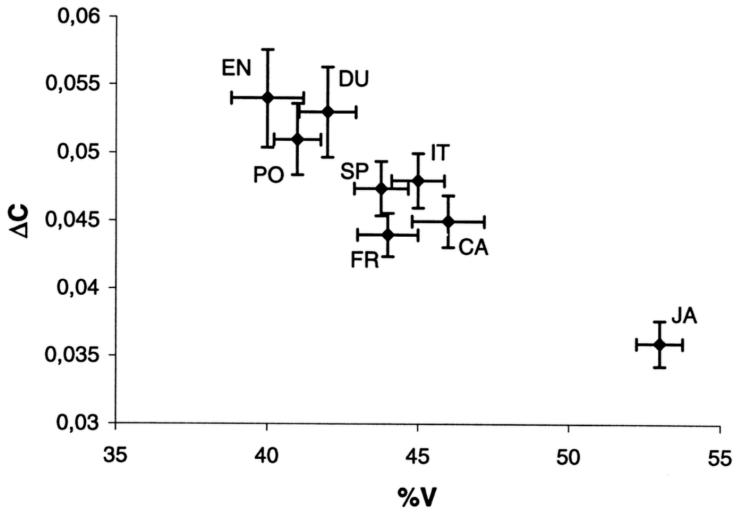


Figure 1.1: Metric score of ΔC and V% measurements. Source: Ramus et al. (1999, p. 273).

Metrics proposed by Ramus et al. (1999) quantify temporal variation over the whole spoken utterances. Low et al. (2000) and Grabe and Low (2002) proposed the pairwise variability indices (PVI) in order to capture temporal contrast (mainly due to lexical stress) between alternating, sequential consonantal, and vocalic intervals. There are two sets of PVI indices. The first, proposed by Low et al. (2000), is the normalized pairwise variability index (nPVI). The nPVI computes the absolute difference between sequential vocalic intervals and divides the difference by the mean duration of the same intervals. Dividing by the mean is to normalize for speech rate variation. The formula for the nPVI measure is provided below:

$$100 \times \left(\sum_{k=1}^{m-1} |(d_k - d_{k+1}) / ((d_k + d_{k+1})/2)| \right) / (m-1)$$

Low et al. (2000) compared the scores of nPVI of stress-timed British English and syllable-timed Singapore English. The latter dialect had a lower nPVI score than the former, reflecting the lower temporal stress contrast between successive vocalic intervals in syllable-timed Singapore English compared to stress-timed British English.

Grabe and Low (2002) proposed the raw pairwise variability (rPVI) for consonantal intervals. It computes the absolute difference between consecutive consonantal intervals without normalizing for rate variation. The formula for computing the rPVI is provided below:

$$\left(\sum_{k=1}^{m-1} |(d_k - d_{k+1})| \right) / (m-1)$$

The rationale for not normalizing for consonantal intervals is that the mean duration of consonantal intervals varies significantly due to language-specific syllable structure (cf. Dellwo & Wagner, 2003). Therefore, normalizing for rate variation by dividing by the mean would result in the loss of significant language-specific variation in syllable structure (Grabe & Low 2002, p. 5-6).

Grabe and Low (2002) used the rPVI and nPVI measures to account for rhythmic differences between several languages that have different rhythmic structures. Data were elicited from one speaker per language reading a passage of text. Scores of nPVI successfully

discriminated between stress-timed English, German, and Dutch and syllable-timed French and Spanish. Scores of rPVI did not show the same discriminatory efficacy as nPVI, as the aforementioned languages had similar rPVI scores. However, rPVI score showed a clear distinction between Polish, a language with mixed rhythm (higher instances of consonantal clustering but without vowel reduction), and Estonian, an unclassified language, as Polish had a very high rPVI score and Estonian had a very low rPVI. Mora-timed Japanese did not occupy a distinct rhythm space, contrary to findings from Ramus et al. (1999), as it was classified with syllable-timed languages based on its low nPVI score, reflecting the weak contrast between strong and weak syllables in the language. Importantly, the data, in general, did not reflect distinct categorization of languages into separate rhythm classes. This was clear for unclassified languages such as Romanian, Greek, and Welsh as they were in the middle of the rhythm space between stress-timed and syllable-timed languages. Based on the latter finding, Grabe and Low (2002) concluded that a distinct classification of languages is not supported, contrary to Ramus et al. (1999). Rather, a continuum of rhythms has to be acknowledged. This means that a language can exhibit features of stress-timing in one dimension but syllable-timing in another. For example, Estonian was found to have very low rPVI (a feature of syllable-timing) but had nPVI scores higher than syllable-timed languages.

Grabe and Low (2002) also compared scores of %V and ΔC, which showed clear differences between stress-timed and syllable-timed languages in Ramus et al. (1999). Scores in Grabe and Low (2002) classified the languages tested in Ramus et al. (1999) differently. For example, Catalan was classified with syllable-timed languages in Ramus et al.'s study due to the high %V score, while in Grabe and Low's (2002) study, Catalan had a very low %V and close to stress-timed English. Japanese, which occupied a distinct rhythm class in Ramus et al.'s (1999) study, with a high %V, had scores similar to stress-timed Dutch in Grabe and Low's (2002) study. Ramus (2002) alluded that the reason for such discrepancies between Grabe and Low's (2002) study and Ramus et al.'s (1999) study is in the way in which speech rate variation, which refers to the number of syllables per second, was accounted for in both studies. In Ramus et al. (1999), speech rate was controlled for by having a similar number of syllables in each sentence (15 to 19) and similar sentence durations (around 3 seconds). On the other hand, Grabe and Low (2002) used the normalized pairwise variability index for vocalic intervals (nPVI). Ramus (2002) tested the nPVI and rPVI on Ramus et al. (1999) data and found striking similarities with %V and ΔC metrics in language classification. Indeed, nPVI and rPVI showed clearer categorization of stress-timed, syllable-timed and mora-timed

than %V and ΔC. Thus, Ramus (2002) suggested that normalizing for speech rate in variance-based metrics such as ΔV and ΔC and in %V is a useful procedure to control for speakers' idiosyncrasies and to quantify cross-linguistic rhythmic variation. However, Ramus (2002) questioned the validity of Grabe and Low's (2002) findings, despite controlling for rate variation with the nPVI metric since they only had one speaker per language. It is possible that with one speaker per language, data in Grabe and Low (2002) might reflect speakers' idiosyncrasies as much as language-specific rhythmic variation. Thus, according to Ramus (2002), the assertion that a continuum of rhythms needs to be acknowledged based on Grabe and Low's (2002) findings cannot be supported, with one speaker per language.

The effect of speech rate on different rhythm metrics was demonstrated in Barry et al. (2003) and Dellwo and Wagner (2003) as ΔV and ΔC, but not %V, varied noticeably with speech rate. Thus, following Ramus's (2002) suggestion, Dellwo and Wagner (2003) devised the variation coefficient of ΔC (VarcoC) to normalize for the effect of rate on the variability of consonantal intervals. It is computed by dividing ΔC by the mean duration of consonantal intervals and multiplying by 100. They showed that VarcoC provided clearer categorisation of stress-timed English and German and syllable-timed French at all rates than ΔC.

White and Mattys (2007a) tested the efficacy of the metrics mentioned so far (%V, ΔV, ΔC, n-PVI, r-PVI, VarcoC) as well a rate normalized metric for ΔV (VarcoV) in discriminating between stress-timed English and Dutch and syllable-timed Spanish and French. The number of speakers used in their study (six speakers per language) was large enough to demonstrate cross-linguistic rhythmic differences. Speech rate was not controlled for in the read sentences they used; thus, White and Mattys also tested the effect of speech rate variation on metric scores. They found that the most effective metrics that discriminated between stress-timed and syllable-timed languages were %V, n-PVI, and VarcoV. Furthermore, %V and VarcoV showed within class differences. French had higher %V and VarcoV scores than Spanish, and English had lower %V scores than Dutch, reflecting that Dutch utilises vowel reduction less than English (Swan & Smith, 2001). VarcoC did not show significant differences between the languages in White and Mattys (2007a), although it was reported in Dellwo and Wagner (2003) to discriminate between different languages. Note, however, that Dellwo and Wagner (2003) did not provide significance testing, so we cannot be sure about the statistical reliability of their results. The lack of significance of VarcoC could be seen as supporting

Grabe and Low's (2002) suggestion that normalizing for speech by dividing by mean consonant duration may result in the loss of relevant language-specific variability. In White and Mattys (2007a), the effect of speech rate was significant for non-rate-normalized metrics (ΔV, ΔC and r-PVI). In particular, there was a significant inverse correlation between speech rate and the less effective metrics, such that metric score decreased with an increase in speech rate. This explains why these metrics were less effective in showing language-specific contrastive rhythm patterns. Rate effects, however, were not significant for rate normalized metrics as well as for %V.

Despite the apparent evidence from rhythm metrics patterns of Grabe and Low (2002) and White and Mattys (2007a) undermining the rhythm class hypothesis (see also Loukina et al., 2011; Arvaniti, 2012), some perceptual studies suggest that rhythm class might have a role is language discrimination tasks. Ramus et al. (2003) used delexicalised speech, in which all consonants were transformed into /s/ and all vowels into /a/ with constant f0, to test listeners' ability to discriminate languages with different rhythms. The delexicalised speech preserves only the temporal variation of consonantal and vocalic intervals and degrades other phonemic or intonational information that might affect participants' perception of temporal variation of vocalic and consonantal intervals. Ramus et al. (2003) showed that adult French listeners could only discriminate between languages that belong to distinct rhythm class but not within class (for example, English vs Spanish, but not English vs Dutch), constituting evidence for the role of rhythm class in language identification.

However, White et al. (2012) showed that adult English listeners could distinguish *sasasa* stimuli of languages that belong to different rhythm classes (Southern British English vs Spanish) and of languages that belong to the same rhythm class (e.g., Orlando English vs Welsh English). Listeners' responses were accounted for by various contrastive rhythm metrics that capture alternation between strong and weak syllables, speech rate (syllables per second), and phrase final lengthening. The predictive efficacy of such cues, however, varied. For example, speech rate was found to be the primary durational cue to distinguish between rhythm classes (Southern British English and Spanish). Also, discrimination between and within rhythm classes depended on the available durational cues. In particular, discrimination between rhythm classes was the strongest because two different durational cues were available, contrastive cues and phrase final lengthening, while discrimination fell within class as only contrastive cues were available for Southern British English vs Welsh English and

only final lengthening was available for Welsh English vs Orlando English. White et al. (2012) concluded that discrimination between languages depends on the availability of different durational cues, not on rhythm class, and such discrimination is gradient, not categorical.

There are important points of criticism that can be addressed regarding the use of rhythm metrics in quantifying cross-linguistic temporal stress contrast. Gibbon (2003, 2006) pointed out that since PVI metrics assume strict binary alternation between strong and weak syllables, they do not capture the difference between simple alternation, as in unary rhythm (utterances with monosyllabic words), and complex alternation, as in ternary rhythm (dactylic *sww* or anapaest *wws* words). For example, the alternation pattern is simple in a sentence with unary rhythm: "This one big fat bear swam near Jane's boat", while it is more complex, with *sww* structure in the sentence: "Jonathan Appleby wandered around with a tune on his lips saw Jennifer Middleton playing a xylophone down on the market-place". The perception of these rhythmic patterns may be different, while the PVI fails to capture such differences. Also, Gibbon pointed to another formal problem in the PVI metrics. He noted that PVI metrics do not capture the difference between alternating and exponentially increasing sequences. For example, the PVI score is the same in the alternating sequence (2, 4, 2, 4, 2, 4) and in the exponentially increasing sequence (2, 4, 8, 16, 32, 64), despite differences in standard deviations and variation coefficients between the two sequences.

Prieto et al. (2012) and Wiget et al. (2010) raised the issue that recorded materials could substantially affect metric scores. Specifically, Prieto et al. (2012) constructed sentences made predominantly with CV syllables, and another set of sentences made predominantly with CVC syllables. They asked English, Spanish and Catalan speakers to read those sentences made with different syllabic construction. They showed that, for instance, %V, which can be representative of the syllabic structure in read materials, differed between the CV and CVC sentences, with higher scores in the former for all languages, thus reflecting the predominance of CV syllables. This shows that read linguistic materials can affect metric scores. Wiget et al. (2010) tested the effect of six sentences read by five Southern Standard British English and found a significant variation between sentences in the scores of %V, VarcoV, and nPVI. Wiget et al. alluded that such between-sentence variation can be due to the distribution of strong and weak syllables and referred to Low et al. (2000), who constructed sentences with only strong syllables and found that they had lower nPVI scores

than sentences containing alternations between strong and weak syllables. In order to monitor alternations between strong and weak syllables, Wiget et al. (2010) devised the contrast regularity index (CRI). This measure assigns a maximum value of 1 for sentences with regular alternations between strong and weak syllables, and lower values for sentences containing strong-strong or weak-weak sequences. They showed that sentences with high CRI values, i.e., have regular alternations between strong and weak syllables, had high nPVI scores while those with low CRI values had low nPVI scores. Global measures, such as VarcoV, did not show such a consistent correlation with CRI, as sentences with high CRI value had low VarcoV values. Wiget et al. asserted that this is a strength point for the nPVI over global measures, as the former is sensitive to the metrical structure, i.e., alternations of strong and weak syllables, in the linguistic material.

In line with the concerns raised by Prieto et al. and Wiget et al., Arvaniti (2009) constructed sentences that mimic a syllable-timed pattern, with more CV syllables and fewer consonantal clusters, and sentences with a stress-timed pattern with more syllables with consonant clusters, and regular alternations between strong and weak syllables. Arvaniti examined the effect of the materials on metric scores by comparing the controlled materials to the uncontrolled materials. She hypothesizes that if metric scores reflect distinct rhythmic classes, the uncontrolled material of, for example, syllable-timed Spanish should cluster with the controlled syllable-timed materials. Instead, she found a significant effect of materials on metric scores, which clustered languages with unexpected categories. For instance, controlled stress-timed materials produced by Spanish speakers clustered with uncontrolled English materials, and controlled syllable-timed materials read by English speakers clustered with uncontrolled Spanish. Thus, the interpretation of metric scores as reflecting language-specific stable rhythmic characteristics can be problematic when the type of material used in a study is not taken into consideration. Therefore, when designing speech corpora to examine languages' rhythmic properties, we should use a large amount of material, or materials should be constructed to reflect specific rhythmic properties of the language of interest (Wiget et al., 2010).

Wiget et al. (2010) also pointed to two important sources of variation in metric scores: inter-speaker variability and segmentation procedures. Clearly, speakers within languages vary in their realisation of temporal contrast between strong and weak syllables. Such variability may endanger any account for cross-language typology, especially if the magnitude of the

difference between speakers is larger than that between languages. Articulation rate is one factor that may lead to significant inter-speaker variability; hence, rate-normalized metrics were suggested to handle such variability. Wiget et al. found that the mean difference between British English speakers in rate-normalized metrics, VarcoV, and nPVI (as well as in %V) relative to the mean difference between English and Spanish, as reported in White and Mattys (2007a), was small. Hence, Wiget et al. recommended using rate-normalized metrics (and %V) in cross-language comparison since they account for inter-speaker variability. On the other hand, non-rate-normalized metrics, such as, ΔV, ΔC and r-PVI, show significant inter-speaker variability, thus, they may be problematic in quantifying cross-language rhythmic differences. Also, between-speaker variability highlights the fact that studies that report cross-language differences based on a single speaker are not reliable, and a sufficient number of speakers should be included.

As for segmentation procedures, different criteria for segmenting the speech signal into consonants and vowels by different measurers might significantly affect the metric scores. Wiget et al. (2010) tested the effect of different measurers on metric scores in their corpus and found small effects of different measurers on the metric scores. Wiget et al. also used an automatic segmentation procedure that is based on statistical properties of the speech signal to avoid potential inconsistencies among different measurers. They found that the automatic phone alignment performed well in that metric scores based on it were consistent with those based on human measurers. Thus, they suggested the use of automatic segmentation procedures for consistency and to save time.

Another important point concerns the relation between cross-language temporal differences and speech rate. Cross-language differences based on speech rate stem from language-specific phonotactics. For example, Spanish utterances are predominantly made up of open CV syllables, resulting in Spanish speakers having an articulation rate faster than English which has complex syllabic structures. As mentioned earlier, White et al. (2012) suggested that speech rate was a stronger acoustic cue for discriminating between stress-timed English and syllable-timed Spanish than contrastive rhythm metrics. Similar results can be found in Dellwo (2010), who showed that German and French listeners relied on speech rate variation more than contrastive rhythm metrics in discriminating between delexicalized German and French stimuli. Thus, while rate-normalized metrics are reliable for discriminating between

languages (White & Mattys, 2007a), speech rate also needs to be reported since it is an important cue for language discrimination.

Arvaniti (2012) found that some metric scores, even the rate-normalized ones, were not significant in differentiating between languages. Perhaps, for some elicitation methods, speech rate would be more relevant in showing language differences. Unfortunately, rate differences were not reported.

While rhythm metrics are meant to capture the relative temporal contrast between strong and weak syllables, there are other timing patterns that can affect metric scores. For example, the use of duration interacts with several timing aspects, such as preserving the contrastive length of segments, demarcating phrasal boundaries by segmental lengthening, and lengthening associated with phrasal stress. Therefore, it is difficult to be certain about the source of durational variability that rhythm metrics convey. Thus, we must admit that rhythm metrics provide a broad approximation of durational variability that is due to multiple factors. (Arvaniti, 2012; Wiget et al., 2010).

Also, rhythm metrics seek only to capture the temporal contrast between strong and weak syllables, while the percept of such contrast varies with multiple non-temporal cues, such as fundamental frequency, overall intensity, and spectral properties of vowels. Furthermore, languages vary in the weighting of different cues in the perception of the contrast between strong and weak syllables. Cumming (2011a,b) attempted to capture language-specific dependency on certain acoustic cues in perceiving the contrast between strong and weak syllables and provided a PVI measure that integrates different weighted acoustic cues. In particular, Cumming (2011a) found that listeners of Swiss German, Swiss French, and French weighted duration and f0 excursion differently. Swiss German listeners were more sensitive to duration manipulation in judging syllables' prominence than f0 manipulation, suggesting that duration is a more important cue in prominence perception. On the other hand, listeners of Swiss French and French were sensitive to duration changes and f0 changes to a similar degree. The perceptual weight of duration and f0 excursion was obtained from the standardised b coefficients of a logistic regression analysis. Cumming (2011b) conducted a production study of the aforementioned languages and computed normalized PVI of duration and f0 excursion of phonologically defined syllables. The PVI of each measure was multiplied by its perceptual weight found in Cumming (2011a) and then integrated into a

single measure. Based on the weighted, integrated PVI, all three languages had similar values. This finding shows that when multidimensionality of the percept of rhythm is taken into account, languages may not appear very distinct. Thus, it is important when providing an account for rhythm perception to take into consideration the multidimensional nature of rhythm and not to focus only on timing.

1.4 Acoustic correlates of speech rhythm in Arabic varieties

The aim of this section is to demonstrate the variability in Arabic dialects in temporal stress contrast. Arabic dialects have variable aspects in syllable structure, and vowel reduction that interact with stress.

Arabic, in general, is a quantity-sensitive language (Watson, 2011a). Stress falls on the super heavy syllable in the word, CVVC and CVCC, and if there is no super heavy syllable, stress falls on the heavy CVV. The syllable CVC is also considered heavy; however, dialects differ in assigning stress to it based on its position in the word. For example, Egyptian treats the final consonant as extrametrical and assigns stress to the initial CV syllable in forms like `CV.CV(C). Egyptian assigns stress to CVC only when non-final (Watson, 2011a). Other dialects, such as Gulf dialects, stress CVC even when it is word-final in forms like CV.`CVC (Watson, 2011b,c). When CVV and CVC syllables are in the same word, all Gulf dialects stress CVV syllable, thus indicating that CVV is heavier than CVC. These are the general rules for stress assignment in Arabic dialects. However, there are restrictions that some Arabic dialects posit on super heavy syllables, CVVC and CVCC, that would result in variable degrees of temporal stress contrast in surface timing. For instance, Egyptian Arabic only allows for CVVC and CVCC syllables to surface in word-final position. In word-medial position, however, Egyptian exhibits two different phonological processes. For word-medial CVVC, a process of vowel shortening is applied, resulting in heavy CVC syllables:

ki.`taab “a book”
ki.taab + ha → [ki.`tab.ha] “her book”

The super heavy CVCC in Egyptian is resolved through epenthesis, creating a new syllable, with the final consonant of the super heavy syllable syllabified as the onset of the new syllable.

bint “a daughter”
bint + na → [`bin.ti.na] “our daughter”

On the other hand, north African dialects such as Moroccan and Tunisian, and Levantine dialects such as Lebanese and Jordanian allow for both types of super heavy syllables, word-medially:

ki.`taab “a book”
ki.taab + ha → [ki.`taab.ha] “her book”

bint “a daughter”
bint + na → [`bint.na] “our daughter”

However, North African dialects allow for more instances of CVCC syllables than Levantine dialects, as the latter group might apply optional epenthesis when CVCC is word-medial (Farwaneh, 1995). Also, north African dialects, especially Moroccan, allow for more clustering of consonants in the syllable onset, up to three consonants (Dell & Elmedlaoui, 2002) (however, this might influence speech rate variation rather than stress contrast).

A survey conducted by Hamdi et al. (2005) from readings of Moroccan, Tunisian and Lebanese speakers provided in the *Araber* corpus (Barkat et al., 2004) supports the reported description of these dialects. Moroccan Arabic exhibited the highest occurrences of syllables with onset clusters (up to three consonants) and with coda clusters (up to two consonants). Tunisian came second, with complex onsets and codas that contained up to two consonants. Lebanese came last.

Another feature that is special to north African dialects is that it is known to have a shorter realisation of vowels than most Arabic dialects (Hamdi et al., 2004). Indeed, phonological vowel length, which is present in most Arabic dialects is claimed to be lost in Moroccan Arabic (Bruggeman, 2018).

Studies on temporal rhythm metrics have shown that structural differences between dialects lead to variation in temporal stress contrast. For example, White and Mattys (2007b) examined metric scores of several British English accents. They showed that southern

standard British English has significantly lower %V and higher VarcoV scores than Welsh Valleys English and Bristolian English, both reported to have lower temporal contrast between strong and weak syllables than southern British English. Despite such differences between the accents of British English, they were distinct from syllable-timed Spanish. Also, other examples within language variation suggest a more complex classification. For example, Frota and Vígaro (2001) showed that European Portuguese had higher variability in consonantal interval duration but lower vocalic interval duration than Brazilian Portuguese.

Hamdi et al. (2004) examined the temporal variation in north African (western) Arabic dialects (Moroccan, Tunisian and Algerian), eastern Arabic dialects (Lebanese, Jordanian and Egyptian). Temporal variation in other languages reported to have different rhythmic classifications were also examined, particularly stress-timed English, syllable-timed French, and Catalan with mixed rhythm. They used %V and ΔC to quantify the temporal variation. Speakers from the different languages were presented with sentences in French from the story: "*The north Wind and the Sun*". The speakers, who apparently were fluent in French, translated the sentences into their own language and dialect, and recordings were taken from the translated materials. Results showed a clear classification of western and eastern dialects, with statistically significant differences between but not within the dialectal groups. Figure 1.2 summarises the findings from Hamdi et al.'s (2004) study.

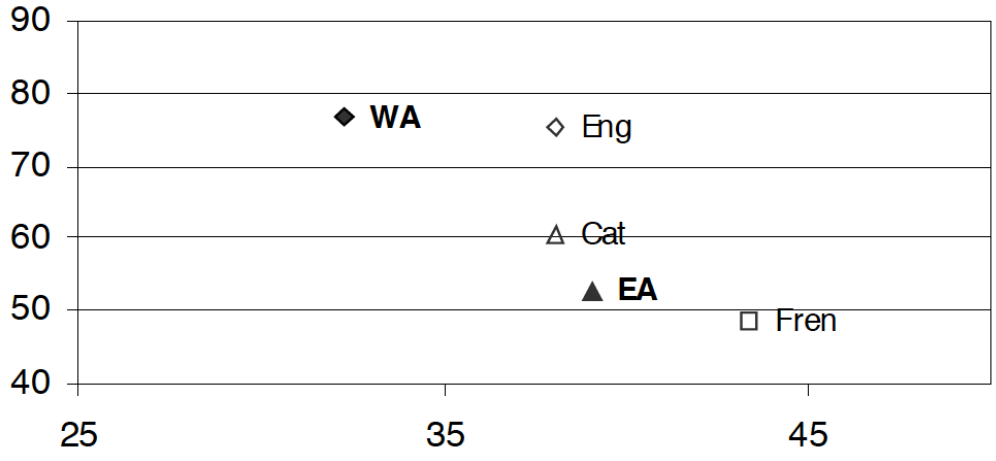


Figure 1.2: Classification of Western and Eastern Arabic dialects based on %V (x-axis) and ΔC (y-axis) along with English, French and Catalan scores. Source: Hamdi et al. (2004, p. 4).

Western dialects had higher ΔC scores than eastern dialects, while the former had lower %V scores than the latter. These findings accord with the previous description of the dialects, that western dialects allow for more consonantal clustering and have a shorter realisation of vowels than eastern dialects. Given this variation based on vocalic and consonantal intervals durations, the classification of western and eastern Arabic dialects along the stress-timing and syllable-timing continuum is complex. The comparison to other languages showed that western dialects were classified with stress-timed English in ΔC scores, while eastern dialects were more similar to syllable-timed French. As for %V, western dialects had lower scores than English, while eastern dialects were similar to English.

Biadsy and Hirschberg (2009) studied durational rhythm metrics, %V, ΔV and ΔC of Gulf Arabic speakers along with other Arabic dialects, such as Egyptian and Levantine, from a corpus of conversational speech. The study showed significant differences between dialects. Gulf Arabic and Levantine had higher ΔC and lower %V scores than Egyptian Arabic, indicating more complex syllable structure in the former dialects than the latter. Egyptian, however, had higher ΔV scores than Gulf and Levantine, suggesting more variability in vocalic intervals in Egyptian than Gulf and Levantine. Perhaps, higher ΔV occurs in Egyptian because the dialect allows for CVV syllables and applies vowel reduction in CVVC syllables, resulting in higher variation in vocalic intervals duration.

We should note, however, that non-rate-normalized metrics, ΔV and ΔC that were used in Hamdi et al. (2004) and Biadsy and Hirschberg (2009) are sensitive to inter-speaker variability, thus, they may not be useful for cross-dialect rhythmic comparisons. Also, as discussed in section 1.3, while rhythm metrics are meant to capture temporal variation in stress contrast, they are confounded with other temporal phenomena, such phrase final lengthening and phrasal stress. To provide a clearer picture of the role of duration as a cue to stress in Arabic dialects, we will review several studies that capture timing variation through direct measures. Some of the studies also report boundary-related and phrasal stress lengthening and other cues to prominence – those findings will also be reviewed.

Bouchhioua (2008) studied duration as a correlate of lexical and phrasal stress in Tunisian Arabic. Phrasally-accented words were embedded in a carrier phrase of the form: Say [target] again. To test for the effect of lexical stress without being confounded with phrasal stress,

target words were placed in contexts in which they do not carry phrasal stress. In particular, target words were included in a carrier (reference) sentence; then target words appeared in another *target* sentences. This ensured that target words were previously given in the carrier sentence, so that they do not get further emphasised by speakers in the target sentence. Examples of materials that avoid phrasal stress in target words are given below:

(carrier sentence)
[target] kelme sɛhlɛ
[target] word easy
gloss: "[target] is an easy word

(target sentences)
qul [target] ʒuʒ mərːat
say [target] two times
gloss: "say target twice"

ʕawd [target] ʒuʒ mərːat
repeat [taget] two times
gloss: "repeat [target] twice"

In the example above, as the target word was given in the carrier sentence, it is less likely that it carries phrasal stress in the target sentences, which seem to have phrasal stress on different new words.

It was found that duration did not distinguish between lexically stressed and unstressed syllables. However, other acoustic cues, spectral balance, f0, and vowel quality, reliably distinguished between stressed and unstressed syllables in unfocused words.

Focus condition (+Focus and –Focus) had a significant effect on the duration of stressed and unstressed syllables, such that stressed and unstressed syllables in focused words were longer than their counterparts in unfocused words. The effect of focus on the duration of stressed syllables was dependent on the stressed syllable's position in the word. Specifically, word-final stressed syllables in focused words were 34 % longer than word-initial stressed syllables in unfocused words. No such interaction was found for unstressed syllables in focused words.

Overall intensity, spectral balance, f0, and vowel quality were reliable cues for focus in Tunisian Arabic.

Bruggeman (2018) investigated acoustic correlates to lexical stress in Moroccan Arabic. The effect of lexical stress not bing confounded with phrasal stress was accounted for in a similar way to that in Bouchhioua (2008), as target words were given in carrier sentences prior to the target sentences to ensure that they do not carry new information, thus do not get further emphasized.

The target syllable was always light (CV) in penultimate position, and it was either stressed, in words of the form `CV.CV or unstressed, in words of the form CV.`CVC. The dependent variable was vowel duration.

It was found that the duration of vowels was not a reliable correlate of stress. Indeed, unstressed vowels were longer than stressed syllables, albeit with a very small difference of 3 ms. Bruggeman also computed f0, spectral energy centre of gravity, and F1 and F2 values of stressed and unstressed vowels. None of the aforementioned cues reliably distinguished between stressed and unstressed vowels. Bruggeman (2018) concluded that there is no evidence for the presence of lexical stress in Moroccan Arabic.

de Jong and Zawaydeh (1999) investigated timing variation due to lexical stress as well as boundary-related lengthening in Jordanian Arabic. The test words occurred either in phrase-final position in statement, question, and isolation forms or in non-phrase-final position in statement and question forms. The test words comprised four syllables, and stress was either penultimate, e.g., fada**bak**ha “he danced”, or antepenultimate, e.g., fa**ba**rada “he got cold”.

Starting with boundary-related lengthening, word-final vowels, which were never stressed, were significantly longer than non-word-final vowels, with a difference of around 60 ms. Word-final vowels in open CV syllables were longer than their counterparts in closed CVC syllables. This reflects the progressive nature of final lengthening (Beckman & Edwards, 1990); since vowels in CV syllables are immediately adjacent to the phrase boundary, they receive more lengthening effect than vowels in CVC syllables in which a consonant intervenes between the phrase boundary and the vowel. Consistent with the progressive effects of final lengthening is the asymmetry between penultimate and antepenultimate

syllables. Stressed and unstressed vowels in penultimate position were longer than their counterparts in antepenult position. Note, however, it is unknown whether boundary-related lengthening is due to the word or the phrase level, as durations of vowels in phrase-final and non-phrase-final words were collapsed.

As for stress-related effects on duration, stressed vowels were longer than unstressed vowels. However, the difference was small, at about 10 ms only. We also need to note that de Jong and Zawaydeh (1999) did not control for the possible confound of phrasal accent on the test words. Thus, the reported difference between stressed and unstressed syllables might be confounded with phrasal stress.

The effect of vowel quality was also investigated. Stressed vowels compared to unstressed vowels (all low /a/ vowels) had significantly higher F1 values.

de Jong and Zawaydeh (2002) investigated the effect of stress and focus on vowel quantity in Jordanian Arabic. Target words appeared in contexts that either carried phrasal accent or not, thus disentangling the effect of phrasal stress from lexical stress. Stressed syllables contained either a long or a short vowel. There was a significant main effect for stress (collapsed over long and short vowels) with a 20 ms difference between stressed and unstressed syllables (25%), which is larger than that in de Jong and Zawaydeh (1999). Stressed and unstressed syllables in focused words were not different from their counterparts in unfocused words. Quantity had a significant main effect: long vowels were 70 ms longer than short vowels. There was a significant two-way interaction between stress and quantity. The durational difference between stressed long and unstressed long vowels was larger (around 28 ms, 19%) than the difference between stressed short and unstressed short vowels (around 10 ms, 16%). It is also useful to report the difference between stressed long and unstressed short vowels: it was 98 ms (66.2%).[1] There is no point in reporting the difference between stressed short and unstressed long vowels since they can't be in the same word in Arabic. There was no interaction between focus and quantity.

[1] The differences between stressed and unstressed vowels were obtained from visual inspection of the figures the authors provided, as exact differences in milliseconds or proportions were not given.

As for vowel quality effects, F1 increase, but not F2, was a reliable predictor of stress contrast and focus.

Vogel et al. (2017) also examined acoustic correlates of stress and focus in Jordanian Arabic, along with final lengthening at the word level. The effect of focus was controlled for, similar to the way used in Bouchhioua (2008). Target words were in non-phrase-final position to examine the effect of final lengthening at the word level. Test vowels, stressed and unstressed with long and short vowels, were in word-initial position in trisyllabic words.

Several acoustic properties were investigated, including mean and standard deviation of f0, overall intensity, duration, and vowel centralisation (taken as the Euclidean distance in the vowel space based on F1 and F2 values). Binary logistic regression was used to examine the predictive power of each acoustic cue of the difference between stressed and unstressed vowels and between stressed vowels in focused and non-focused words. In non-focused words, mean f0 was the strongest acoustic property to distinguish between stressed and unstressed vowels, followed by intensity and duration. The standard deviation of f0 and vowel centralisation did not have a significant effect on the classification of stressed and unstressed vowels. In focus condition, mean f0 was the strongest property in discriminating between stressed syllables in focused and unfocused words, followed by duration and intensity.

Differences between stressed and unstressed vowels duration in non-focused words were somewhat smaller than those reported in de Jong and Zawaydeh (2002). Stressed short vowels were only 2 ms longer than unstressed vowels (4%), and stressed long vowels were longer than unstressed long vowels by 14 ms (12%). The difference between stressed long vowels and unstressed short vowels was substantial at 63 ms (56%), albeit smaller than that reported in de Jong and Zawaydeh (2002). In focused and non-focused words, the difference between stressed long vowels was 12 ms (9%), between stressed short vowels 7 ms (12%), and between stressed long vowels in focused words, and stressed short vowels, in non-focused words was 66 ms (53%).

Durational effects on word-final lengthening (in non-focused words) were observable, with a mean of 12 ms (15%) for final CV unstressed vowels compared to non-final CV vowels. When a stressed long vowel preceded final unstressed CV syllables, final lengthening was

substantial, with a difference of around 61 ms (57%). Vogel et al. (2017) attributed the substantial lengthening in the latter case to the durational effects of the adjacent stressed syllable.

Almbark et al. (2014) compared acoustic correlates of stress and accent in Egyptian and Jordanian Arabic. The target accented words appeared in a contrastive context, while the target unaccented words appeared in a context where focus was assigned to a different word. Target syllables were word-initial. Duration, f0, and intensity were reliable cues to stress but not spectral balance or F1 and F2 values. Phrasal accent had a significant effect on duration and intensity of target syllables but not on f0, spectral balance, and F1 and F2 values. Almbark et al. (2014) did not provide raw values of durations of target syllables for the two dialects, as durations were normalised by dividing by word duration, but they report that differences between stressed and unstressed syllables were more pronounced in Egyptian than in Jordanian Arabic.

Chahal (2001) investigated whether there are reliable acoustic correlates that can distinguish three levels of prominence in Lebanese Arabic, namely, lexical stress, broad and narrow focus. Chahal found that duration, f0, and intensity reliably distinguished between the three levels of prominence. However, she did not report differences between unaccented lexically stressed syllables and unstressed syllables.

Kelly (2021) investigated phrase-final lengthening in Lebanese Arabic in disyllabic words with initial stress. To control for the effect of lengthening due to accent, target words, in phrase-medial and phrase-final positions, were in contrastive focus, thus both carried an accent. There was an interaction between vowel length and phrasal position such that lengthening effects were larger for stressed long vowels than stressed short vowels. Stressed long vowels in phrase-final position were longer than stressed long vowels in phrase-medial position by 8.4 %, while stressed short vowels in phrase-final position were longer than their counterparts in phrase-medial position by 0.6 % only. Unstressed vowels, which were only short, were longer in the phrase-final position than in the phrase-medial position. There was also a stress-adjacency effect on unstressed syllables. Unstressed syllables in the phrase-final position were longer than in the phrase-medial position by 14% when the preceding syllable was stressed long, and they were longer by 11% when the preceding syllable was stressed short.

Several studies investigated the effects of broad and narrow focus on stressed syllables' duration and fundamental frequency in Gulf Arabic dialects, for example Yeou et al. (2007) for Kuwaiti Arabic; Alzaidi (2014) for Hijazi Arabic; Almalki (2020) for urban Najdi Arabic. These studies reported significant durational and tonal contributions in marking different focus types in different Gulf dialects. Unfortunately, these studies did not report the temporal contrast degrees between stressed and unstressed syllables, which is one of our main interests in dialectal typology.

1.4.1 Hadari and Bedouin Kuwaiti dialects

Kuwait lies at the tip of the Persian Gulf (also known as the Arabian Gulf), and Saudi Arabia boards it from the south and Iraq from the north. Kuwait also has maritime borders with Iran (Figure 1.3). Because of its location in Arabian Gulf, Kuwait prospered in the 18th century, becoming a port for goods transportation from different parts of Arabian Gulf countries and India. Therefore, many have migrated from neighbouring countries, such as Saudi Arabia, Iraq, and Iran, looking for better financial conditions. This migration extended until the mid of 20th century, which witnessed oil discovery and resulted in substantial financial and economic growth (Crystal, 1990; Sagher, 2004). Due to these migrations, Kuwait has different dialects that stem from different origins of migration; however, there are two main dialectal groups in Kuwait, the Bedouin (rural) and the Hadari (sedentary) dialects, which their prosodic timing characteristics are the focus of this study.

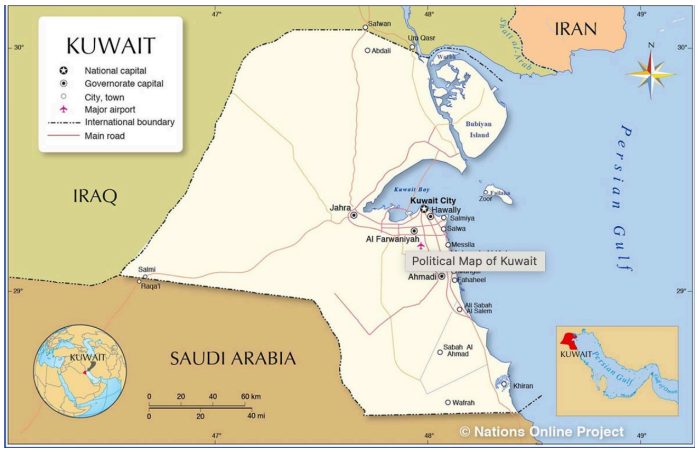


Figure 1. 3: A map of Kuwait. Kuwait boards Saudi Arabia from the south and Iraq from the north. Kuwait also has maritime borders with Iran. Retrieved from: https://www.nationsonline.org/oneworld/map/kuwait-map.htm.

Bedouin Arabic is the dialect of Arabic speakers who consider themselves to have tribal origins. Bedouins used to live in the desert of the Arabian Gulf countries but moved to cities for better living conditions. Bedouin speakers in Kuwait have origins that belong to Najd in Saudi Arabia, which lies between Hijaz and eastern Arabia, which includes the eastern part of Saudi Arabia (ħasa), Kuwait and Bahrain. Several tribes constitute the Bedouin community in Kuwait. Some of these tribes are Al-Enzi, Al-Shammari, Al-Autaibi, Al-Azmi and Al-Harbi (see Ingham,1994 for listing more tribes that belong to different parts of Najd). Despite living in the cities of Kuwait, a lot of Bedouin speakers preserved their own dialect, since dialect represents an important part of their social and demographic identity. The other dialectal group in Kuwait is Hadari (sedentary) which also includes speakers that migrated from Najd, specifically, from the central part of Najd, but are described to be Hadari even before migration. Other groups that constitute the Hadari community migrated from Iran, the eastern part of Saudi Arabia (ħasa), Bahrain and Iraq (Holes, 2006).

The Hadari dialect is considered more prestigious than the Bedouin dialect (Rosenhouse, 2006). However, even within Hadari dialects, there are stylistic variations, especially at the phonemic level, some of which may be considered more prestigious than others (for an in-depth sociolinguistic study of dialectal variation in Kuwait, see Taqi, 2010).

Bedouin dialects, in general, are said to be more conservative in their phonological properties than Hadari dialects, because the former preserve features from classical Arabic, that have been subjected to phonological changes in the Hadari dialects (Abu Haider, 2006). Based on the phonological differences that each dialect shows, rhythmic differences, i.e., different degrees of temporal stress contrasts, are expected. Below, differences in the syllable structure between the two dialects are listed.

1.4.1.1 Differences in syllable structure and vowel reduction in Hadari and Bedouin dialects

As discussed above, dialectal differences in syllable structure and in vowel reduction in unstressed syllables have consequences for the temporal contrast between stressed and unstressed syllables. Several differences in the syllable structure between Bedouin Kuwaiti Arabic and Hadari Kuwaiti Arabic might have effects on the temporal stress contrast. These include greater permissibility in Hadari than in Bedouin of a more complex structure of

stressed syllables, especially with complex coda clusters, and greater tolerance in Hadari than in Bedouin of super long stressed syllables in word-medial position. Also, Hadari has greater unstressed syllable reduction than Bedouin.

1.4.1.1.1 Permissibility of consonantal clustering

In the word-initial position, both dialects allow consonant clusters. The maximum number of consonants in a cluster is two (cf. Ingham, 1994; Farwaneh, 1995). Below are examples of words that have initial consonant clusters:

(1)
sla:l "baskets"
tla:l "hills"
ħma:r "a donkey"

Coda consonant clusters are allowed for both dialects in word-final position:

(2)
galb "heart"
darb " road"
karʃ "belly"

The dialects differ from each other, however, when the words that contain final clusters are suffixed:

Hadari:
(3)
galb+na → 'galb.na "our heart"
darb + na → 'darb.na "our road"

Bedouin:
(4)
galb + na → 'gal.ba.na "our heart"
darb + na → 'dar.ba.na "our road"

Thus, Hadari allows for coda clusters when the word is suffixed, while Bedouin tends to break the clusters through epenthesis.

1.4.1.1.2 Super long syllables

In word-final position, both dialects allow for super long syllables, i.e., syllables that contain a long vowel followed by a coda:

(5)
beet "a house"
daar "a room"

However, word-medially, when words containing a super long syllable are suffixed, Bedouin inserts an epenthetic vowel, usually a low vowel "a" (cf. Ingham, 1994, p. 17), resulting in the syllabification of the stem coda into a new syllable:

(6)
beet + na → 'bee.ta.na "our house"
daar + na → 'daa.ra.na "our room"

On the other hand, Hadari allows for super long syllables word-medially:
(7)
beet + na → 'beet.na "our house"
daar + na → 'daar.na "our room"

1.4.1.1.3 Short vowels in open and closed syllables

The type of vowels produced in open (CV) unstressed syllables differ between the two dialects, where there is a tendency in Hadari to produce a centralised vowel, whereas Bedouin produces full low vowel "a" (cf. Ingham, 1994, p. 18).

The examples below are for Bedouin and specifically for words in which unstressed CV syllables are followed by a closed (CVC) syllable:

(8)
xa`lat “he mixed”
ga`lab “he overturned”
da`xal “he entered”
za`ʕal “he got upset”

On the other hand, in Hadari, the low vowel is usually centralised (Holes, 2006):

(11)
xə`lat “he mixed”
gə`lab “he overturned”
də`xal “he entered”
zə`ʕal “he got upset”

The fact that Hadari allows for (a) more complex and longer stressed syllables, especially word-medially, than Bedouin, and (b) Hadari exhibits vowel reduction in unstressed syllables more than Bedouin makes it plausible that Hadari might exhibit more temporal stress contrast than Bedouin.

1.5 Models of speech timing

Speech timing must be considered in contexts that refer to the communication of the linguistic structure. Thus, models of speech timing must capture the mechanisms by which speakers manipulate timing in order to communicate the linguistic structure to listeners (White, 2014). The linguistic structure is hierarchical in nature (Selkirk, 1986; Nespor & Vogel, 1986), with several prosodic constituents that have a nesting relationship amongst them; syllables are nested within higher prosodic units such as feet, and feet within phonological words and words within phrases. There is also a metrical organizational property in the prosodic hierarchy, that refers to the relative prominence of prosodic constituents. Thus, for example, within a foot, a certain syllable can be more prominent than others, and this prominent syllable constitutes the head of the foot. Feet themselves also have a relation in relative prominence; within a word, the foot that contains the most prominent syllable is considered the head of the word. The word that nests the most prominent foot is considered the most prominent word within the phrase. Relative prominence is not the only

aspect of the prosodic hierarchy. Final words have a special prosodic status as they are correlated with phonetic lengthening that marks the phrase boundary (Liberman, 1975).

Researchers criticised durational rhythm metrics (e.g., Arvaniti, 2012; Gibbon, 2006, 2009) for not capturing the hierarchical linguistic structure and the means by which speakers communicate the structure to listeners. As discussed earlier, rhythm metrics provide crude measurements for temporal variation due to lexical and phrasal prominence as well as final lengthening and may not capture the temporal mechanisms by which speakers communicate the hierarchical prosodic structure.

Below, we contrast two views of speech timing: the coupled oscillators model (O'Dell & Nieminen, 1999) and the locus and domain approach (White, 2002, 2014). The coupled oscillators model assumes a temporal coordination relation between prosodically-nested constituents, such as syllables and feet, in that they influence the timing of each other. In this conceptualisation of speech timing, the coupled oscillators model captures a structure that refers to the grouping relation between higher-level and lower-level prosodic constituents. On the other hand, the locus and domain approach does not assume any timing effect of a higher unit on its subconstituents. Rather, the prosodic structure affects timing through phonologically defined loci. For example, the locus of lexical stress is the stressed syllable, with greater timing effects localised to the nucleus of the lexically stressed syllable. Below, we will describe in more detail the timing mechanisms in signalling the prosodic structure that each approach provides.

1.5.1 The coupled oscillators model

Several researchers believe that an account for timing should capture the nesting relation between the constituents of the prosodic hierarchy, for example, syllables and feet, or inter-stress intervals. Grouping of stressed and unstressed syllables within inter-stress intervals entails a structure that influences variation in surface timing. Dauer's (1983) observation from her study that the duration of inter-stress intervals is not independent of the number of syllables, nor is it a simple additive function of the number of syllables, hints at a mutual timing effect between an inter-stress interval and syllables within it. In particular, while the number of syllables in an inter-stress causes a growth in its duration (syllable effect), the size of an inter-stress in terms of the number of syllables might affect syllable durations through

compression effects (inter-stress effect), thus reflecting mutual timing effect between the two levels of prosodic structure. Eriksson (1991) investigated this possibility by reanalysing Dauer's data of canonical stress-timed English and Thai and canonical syllable-timed Spanish, Greek and Italian. Eriksson used a linear regression analysis to model the inter-stress duration as a function of the number of syllables in an inter-stress interval. He found that the linear regression slope, which represents the duration added to the inter-stress interval by the number of syllables in it, was similar in all languages at 100 ms. However, an interesting difference was that of the initial value of the durational increase, i.e., the the linear regression intercept. The intercept value clustered around 200 ms in stress-timed languages, English and Thai, and at 100 ms in syllable-timed languages, Spanish, Greek, and Italian. The inter-stress interval duration as a function of the number of syllables can thus be expressed in terms of a simple linear regression equation:

$$(I = a + nb)$$

Where *a* is the intercept, *b* is the slope of the equation, and *n* is a function of the number of syllables in an inter-stress interval. Eriksson asserted that the natural interpretation of intercept value is that it refers to the extra duration added by stressed syllables in the inter-stress. However, Eriksson also observed that the intercept value does not define where the "extra duration" effect comes from in the inter-stress interval. Thus, Eriksson raised the possibility that the different intercept values between languages can be due to variable compression effects from syllables in the inter-stress interval. As such, mutual timing effects between inter-stress intervals and the number of syllables can be expressed through syllabic compression imposed by the size of the inter-stress interval (inter-stress effect), and through duration added by the number of syllables (syllable effect).

O'Dell and Nieminen (1999) suggested capturing the mutual timing effects between inter-stress intervals and syllables by positing two interacting oscillators. Specifically, in their approach, there are two oscillators that represent two levels of the prosodic hierarchy; the syllabic oscillator and the inter-stress oscillator. Every oscillator is associated with its own natural frequency. The syllabic oscillator is faster in frequency than the inter-stress oscillator. When the oscillators are observed in isolation they will produce periodicity at a single prosodic level, either the syllable level, thus syllable-timing, or at the inter-stress level, thus stress-timing. In O'Dell and Nieminen's framework, however, the oscillators are said to

interact with each other by a coupling function. As such, the oscillators settle at a stable frequency pattern, in which the frequency of the faster oscillator is an integer multiple of the frequency of the slower oscillator (Windmann, 2016, p. 70-71). Figure 1.4 shows a schematic representation of a 1:2 ratio of the syllable oscillator to the inter-stress oscillator representing a stable state coupling.

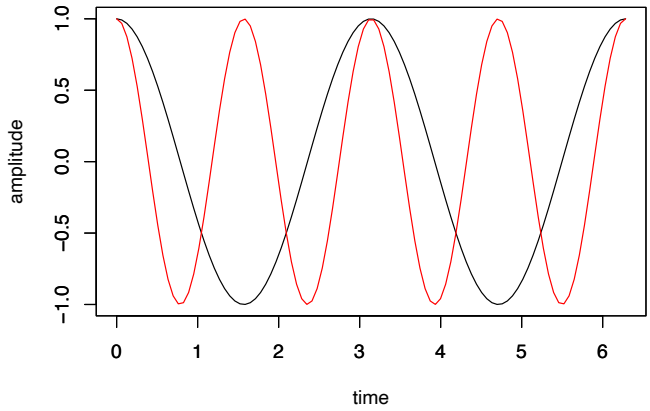


Figure 1.4: Stable state between the syllabic oscillator (in red) and the inter-stress oscillator (black) where the frequency of the syllabic oscillator is an integer multiple of the frequency of the inter-stress oscillator at 1:2 ratios.

According to O'Dell and Nieminen (1999), languages differ in the dominance of either of the oscillators. In stress-timed languages, the inter-stress oscillator is the more dominant, which means that as the number of syllables increases in a stress group, the inter-stress oscillator preserves its natural frequency and becomes less prone to frequency changes than the syllabic oscillator. The opposite is true in syllable-timed languages: as the number of syllables increases in the stress group, the stress oscillator becomes prone to more changes in frequency, while the syllabic oscillator preserves its natural frequency.

O'Dell and Nieminen (1999) modelled the interaction between the two oscillators with the following two equations:

(1)

$\theta_1 = \omega_1 + H(\phi_n)$

(2)

$\theta_2 = \omega_2 - r\mathrm{H}(\phi_n)$

The two equations depict that the natural frequency of the inter-stress interval oscillator (ω_1) and the syllabic oscillator (ω_2) interact with each other by a coupling function H that depends on the average phase difference ϕ and the number of syllables in the stress group n (when phase difference is zero, the two oscillators will be in a stable state as in Figure 1.3). The coupling function is added with different signs to ω_1 and ω_2. The only varying function between the two oscillators is the coupling strength function *r*, which determines the relative dominance of the inter-stress oscillator to the syllabic oscillator. Thus, the actual frequency of the two oscillators θ_1 and θ_2 depends on the coupling function H, which, as explained earlier, would lead the two oscillators to settle at a stable frequency, where the faster oscillator is an integer multiple of the slower oscillator. The varying factor *r* differentiates between languages in terms of the dominance of one of the oscillators in the coupling relation.

O'Dell and Nieminen (1999) provide another crucial part of their model, which is the period of the stress oscillator. It is written as follows in equation 3:

(3)

$$\mathrm{T}_1(\mathrm{n}) = \frac{1}{\omega 1 + \mathrm{H}(\phi \mathrm{n})} = \frac{r}{r\omega 1 + \omega 2} + \frac{1}{r\omega 1 + \omega 2} n$$

Equation (3) depicts that the period of the stress oscillator is a linear function of the number of syllables in the stress group. This is similar to the linear regression equation provided by Eriksson (1991) ($I = a + nb$). Since the stress oscillator is a linear function of the number of syllables, the *r* relative strength function can be empirically estimated as a ratio of the intercept *a*, which reflects stress level effects, to the slope *b*, which reflects the duration added by syllables. Thus, *r* can be written as: $r = a/b$.

If $r > 1$, it depicts the dominance of the stress oscillator, whereas if $r \leq 1$, it is the syllabic oscillator that is the most dominant.

O'Dell and Nieminen (1999) applied the strength parameter to the same data used in Eriksson (1991) and added data from Finnish. The *r* parameter value ($r=a/b$) classified languages the same way in Erickson (1991). Finnish was classified as syllable-timed.

As the coupled oscillators model assumes a coupling relation between the syllable and the inter-stress oscillators, with gradient variation between languages in the coupling strength, the coupled oscillators model differs from isochrony-based models that assume strict surface timing pattern of either syllables or inter-stress. Rather, the relationship between the assumed oscillators is based on hierarchical nesting, which implies mutual timing effects between prosodic units.

A useful example of an interaction between different levels of the prosodic hierarchy in explaining timing variation comes from Asu and Nolan's (2006) study of Estonian and English. In the first experiment, Asu and Nolan (2006) computed the PVI of vocalic and consonantal intervals, as well as of phonological syllables and inter-stress intervals. They found considerable between speaker variability in PVI scores of vocalic and consonantal intervals compared to PVIs of syllables and inter-stress intervals. They suggested that low inter-speaker variability in syllables' and inter-stress intervals' PVIs reflects greater control of speakers in the timing of syllables and inter-stress intervals, which suggests that syllables and inter-stress intervals are important timing units in Estonian. In a second experiment, they compared syllables' and inter-stress intervals' PVIs between English and Estonian. While PVI scores of inter-stress intervals were similar in both languages, PVI scores of the syllable level were significantly different between the two languages: syllabic PVI was higher in English than in Estonian. They suggested that this pattern reflects variable degrees of interaction between inter-stress intervals and syllables between English and Estonian. The higher syllable's PVI in English may be due to greater unstressed syllable reduction, which may be caused by the size of inter-stress intervals. Thus, while syllables and inter-stress intervals are important timing units in English and Estonian, inter-stress intervals in English show greater dominance in the interaction with syllables, than in Estonian.

Bouzon and Hirst (2004) investigated potential timing interaction between different hierarchical constituents of the prosodic structure, e.g., phones within syllables, syllables within feet, and feet within the intonational phrase. They were also interested in comparing the effects of the foot and the natural rhythm unit (Jassem et al., 1984) on syllables duration.

The foot spans two stresses across word boundaries, and includes unstressed syllables within the two stresses. The natural rhythm unit, however, starts with a stressed syllable and includes the following unstressed syllables within the same phonological word. Every other unstressed syllable that does not belong to the natural rhythm unit belongs to a unit called the anacrusis. It is said that the natural rhythm unit tends to have equal durations, as proposed by Jassem et al. (1984), while the anacrusis tends to be articulated as fast as possible and thus tends to increase linearly with the number of syllables within it. Bouzon and Hirst examined two hypotheses. The first is strict isochrony, in which higher-level units would be equal in duration regardless of the number of subconstituents in them, and the second is weak isochrony, in which subconstituents would compress to some degree only for the higher units to show a *tendency* towards isochrony, but not complete isochrony. Bouzon and Hirst tested these hypotheses on a corpus of British English. For the strict isochrony, they found a positive correlation between the size of higher units, in terms of the number of subconstituents, and the duration of the higher units, which refutes the strict isochrony hypothesis. With regard to the weak isochrony hypothesis, the prediction was that there should be a negative correlation between the size of higher units and the duration of their subconstituents, which would provide evidence for a tendency towards isochrony. They found a negative correlation between the complexity of higher units and the duration of subconstituents, as evidenced by the negative slope of their regression analysis. They also found that the natural rhythm unit had stronger compression effects on syllables than the foot, as indicated by the larger slope value of the natural rhythm unit. The durational compression effects of lower-level units imposed by the size of higher-level units may be seen as evidence for the timing interaction between different levels of the prosodic structure.

Similar to Bouzon and Hirst (2004), Kim and Cole (2005) found polysyllabic shortening in English data. Stressed syllable duration was shorter as the size of the stress foot increased, thus reflecting a timing effect of stress foot on syllables. O'Dell and Nieminen (2009) refer to polysyllabic shortening, the inverse relationship between the size of inter-stress intervals, in terms of the number of syllables, and the duration of syllables in the inter-stress, as a reflection of the interaction between syllabic and inter-stress oscillators. They refer to polysyllabic shortening as "rhythmic gradation" as a better description of the oscillatory interaction.

The ability of speakers to speak in synchrony when presented with a text and asked to read it together has also been modelled with interacting oscillators. Cummins (2003) showed that the mean difference between the parallel speech waveform of two interlocutors was small at 60 ms at the beginning of the phrase, and 40 ms for phrase non-initial position, suggesting temporal coordination between speakers in synchronized speech.

Following this finding from Cummins (2003), Krivokapić (2013) studied the temporal coordination patterns in American English and in Indian English when speakers of each dialect are asked to speak in synchrony. American English is classified as "stress-timed", while Indian English as "syllable-timed". Krivokapić hypothesised that American English speakers would show tendencies to "syllable-timing" when they speak in synchrony with Indian speakers. Specifically, American English speakers may exhibit less polysyllabic shortening in the synchronous condition compared to the solo condition. Consequently, the foot duration would increase more as a function of the number of syllables in the synchronous condition. On the other hand, Indian English speakers would converge to a "stress-timing" pattern when speaking in synchrony with American English speakers. The hypothesis was that Indian English speakers might exhibit more polysyllabic shortening in the synchronous condition compared to the solo condition, and foot duration would show less increase as a function of the number of syllables in the synchronous condition. It was found that, in the solo condition, both dialects showed mixed patterns of "stress-timing" and "syllable-timing". Speakers of both dialects exhibited polysyllabic shortening, as the duration of stressed syllables in monosyllabic feet was longer than in multisyllabic feet. Polysyllabic shortening, however, was stronger in American English than in Indian English, which argues for the classification of American English as "stress-timed", or in coupled oscillators terms, reflects the dominance of the stress oscillator over the syllabic oscillator in English. Foot duration increased linearly as a function of the number of syllables in both dialects, whereas Indian English showed more increase in foot duration than American English, which can be considered as a manifestation of the dominance of the syllabic oscillator in the former. In the synchronous condition, one Indian English speaker showed evidence for rhythmic convergence. This speaker exhibited polysyllabic shortening to a greater extent in the synchronous condition than in solo condition and the effect of the number of syllables on foot duration was weaker in the synchronous condition than in the solo condition. American English speakers showed evidence for convergence only at the foot level: the effect of the number of syllables on foot duration was stronger in the synchronous condition than in the

solo condition. Krivokapić (2013) modelled convergence effects as a system of coupled oscillators. In the solo condition, American English exhibited the dominance of the stress oscillator; however, in the synchronous condition, speakers of the dialect converged to the rhythmic properties of Indian English, allowing for the dominance of the syllabic oscillator. The opposite is true for one Indian English speaker who exhibited convergence to American English, thus allowing for the dominance of the stress oscillator.

Several improvements to the coupled oscillators model have been suggested by Saltzman et al. (2008). The need for improvements was based on data simulation that Saltzman et al. conducted based on the coupled oscillators model to account for polysyllabic shortening reported in Kim and Cole's (2005) analysis of English corpus. Their simulation produced syllabic durations that were shorter in tri-syllabic feet than in bi-syllabic feet. However, they note an important drawback in their simulation that produces equal durations of stressed and unstressed syllables, while stressed syllables are supposed to be longer. In order to overcome this, Saltzman et al. invoke a temporal modulation function (μ_T-gesture) that slows down the phase flow of the syllabic oscillator during the period of the stressed syllable within the foot oscillator. The temporal modulation function was given the shape of a half-cosine that varies from 0 to 1. The half-cosine function starts at 0 at the beginning of the stressed syllable and smoothly increases to 1, causing the effect of slowing down the phase flow of the stressed syllable's cycle. At the unstressed syllable's cycle, the value of the temporal modulation function is set at 0, thus does not affect the phase flow of the unstressed syllable's cycle. Slowing down the phase flow of the stressed syllable's oscillatory cycle causes lengthening of the gesture of the stressed syllables, leading to the longer duration of stressed syllables compared to unstressed syllables.

However, a drawback in the modulation function in Saltzman et al.'s (2008) simulation is that it resulted in the compression of unstressed syllables, when results in Kim and Cole (2005) only show compression in stressed syllables in polysyllabic feet. To account for compression in stressed syllables but not unstressed syllables, Saltzman et al. modulated the coupling strength function, which caused the dominance of the foot oscillator, by switching off the coupling force from the foot oscillator to the syllable oscillator during the period of the unstressed syllable's cycles. This results in the foot oscillator not affecting the phase, thus duration, of unstressed syllables in polysyllabic feet.

Similar to the temporal modulation function, Byrd and Saltzman (2003) proposed the prosodic gesture ($\pi - gesture$) to account for edge-related lengthening phenomena, i.e., phrase-initial and phrase-final lengthening. The prosodic gesture is placed at the edge of an utterance, slowing down the time flow of an utterance, and resulting in lengthening effects. The prosodic gesture, which is shaped as a half-cosine, reaches its maximum value at the edge of an utterance. As a result, individual gestures that overlap with the prosodic gesture get lengthened, and the closer the gesture to the prosodic gesture peak, the stronger the lengthening effects. It is noteworthy that the main difference between the prosodic gesture and the temporal modulation gesture is that the latter directly influences the dynamics of the articulatory gestures of syllables, while the former is only meant to slow down the time flow of the utterance at certain phrase positions, resulting in longer constriction time of gestures.

In summary, the coupled oscillators model asserts that the hierarchical grouping relation between stress feet and syllables implies an interaction between the two levels of prosodic structure in surface timing. Hierarchical timing relation between syllables and stress feet is evidenced in the linear relation between the number of syllables within a foot and the foot's duration (Eriksson, 1991). O'Dell and Nieminen (1999) modelled the timing relation between syllables and stress feet as a system of coupled oscillators. Languages vary in the relative coupling strength between the two oscillators. For instance, in English, the stress oscillator is said to be more dominant than the syllable oscillator, excreting a greater influence on the natural frequency of the syllable oscillator. In surface timing, the dominance of the stress oscillator is reflected in the compression effects that the stress foot size imposes on syllables within the foot. On the other hand, less syllabic compression within the foot, as in Spanish, is modelled with the dominance of the syllable oscillator. Improvements to the coupled oscillators model were suggested to reconcile with empirical data and account for prosodic boundary lengthening effects. Saltzman et al. (2008) suggested a temporal modulation gesture to account for stressed syllable lengthening and suggested modulating the coupling force of the stress oscillator to account for stressed syllable polysyllabic shortening. Byrd and Saltzman (2003) suggested a modulation gesture that slows down the time flow at the utterance's boundaries to model phrase-initial and phrase-final lengthening.

In the following section, we will review the locus and domain approach of speech timing and attempt to reflect upon the plausibility of hierarchical timing suggested by the coupled oscillators model based on well-attested local prosodic timing influences across languages.

1.5.2 Locus and domain view of speech timing

As described briefly above, the locus and domain approach posits that the prosodic structure influences speech timing at domain heads, which refer to the prominence structure in the utterance, and domain edges, which refer to word and phrase boundaries. Within domain heads and domain edges, timing effects are localised to structurally-defined segments. For example, in English, lexical stress is marked by lengthening the lexically stressed syllable, with greater lengthening effects localised at the stressed syllable nucleus (Klatt, 1976). Phrasal stress influences the duration of pitch accented words, with a greater lengthening effect at the stressed syllable (Turk & White, 1999). For domain edges, the onset of the initial syllable of the word is lengthened more than the onset of the word-medial syllable, marking the beginning of the word (Oller, 1973). Words adjacent to phrase boundaries are lengthened at the word rhyme, with lengthening effects localised at the stressed syllable rhyme and the word-final syllable rhyme. Syllables between the primary stressed syllable and the word-final syllable are skipped from phrase final lengthening effects (Turk & Shattuck-Hufnagel, 2007). Lengthening at the word level is distinct from that of the phrase level. At the word level, lengthening targets the primary stressed vowel, with greater lengthening effects as the primary stressed vowel is closer to the word edge. In the absence of a phrase boundary, lengthening does not target segments following the primary stressed vowel (White, 2002; White & Turk, 2010).

Domain-edge lengthening effects, such as phrase-final lengthening, appear to be universal (Beckman, 1992) as it is widely attested for many languages (e.g., Lebanese Arabic: Kelly, 2021; Hebrew: Berkovits, 1994; Czech: Dankovičovà, 1997; Dutch: Gussenhoven & Rietveld, 1992). Domain-head effects are also employed widely, with variability in the magnitude of lengthening due to prominence and with some languages lacking some levels of prominence, such as the lack of lexical stress in French and Korean.
These localised timing effects aid listeners in perceiving prominences and words and utterances boundaries, thus disambiguating the linguistic structure of speech.

Unlike the coupled oscillators model, the locus and domain approach does not assume a direct timing effect from higher-level prosodic units on their subconstitients. Thus, polysyllabic shortening, i.e., the inverse relationship between constituent size and the duration of its subconstituents, is refuted as a timing mechanism to signal the linguistic

structure. White (2002) suggested that reported effects of polysyllabic shortening can be re-analysed in terms of well-attested localised timing effects such as accentual lengthening and final lengthening.

One of the main studies that argue for the constituent size effect is that of Port (1981). He used English nonsense words such as “dib”, “dibber”, “dibberly”, in a carrier sentence “I say [target word] again every Monday”, to test the shortening effects on the stressed syllable nuclei. Port (1981) showed that the duration of the stressed vowel was longer in the monosyllabic word “dib” than in the disyllabic “dibber” and longer in “dibber” than in the trisyllabic “dibberly”, providing evidence for word size effect on the duration of the stressed vowel. Port also reported shortening of the stressed syllable’s onset and coda consonants, although less significant than of the stressed vowel, indicating that the whole syllable shortens with the word size.

White (2002) pointed out that a problem in Port’s (1981) study is that the test words were only left-headed, i.e., the stressed syllable was word-initial, and thus it would be difficult to disentangle word-final lengthening effects from polysyllabic shortening. That is, since proximity to the word boundary causes lengthening of the stressed vowel, the stressed vowel should be longer in monosyllabic words than in disyllables and in disyllables than in trisyllables, thus confounding the effect of polysyllabic shortening.

Another important confound in Port’s study is that target words are likely to carry phrasal accent in contexts of the form: “I say [target words] again”. Thus, it is possible that polysyllabic shortening is dependent on phrasal accent and may not be viewed as a general timing mechanism. Indeed, findings from Turk and White (1999) on Scottish English showed that accentual lengthening was greater in stressed syllables in monosyllables (23 %) than in bisyllables (16 %), suggesting a link between polysyllabic shortening and accentual lengthening. Turk and Shattuck-Hufnagel (2000), who found some evidence for polysyllabic shortening in accented and unaccented words, showed that polysyllabic shortening was greater in accented words than in unaccented words, thus also suggesting a dependency of polysyllabic shortening on the presence of accentual lengthening. Note, however, that polysyllabic shortening in Turk and Shattuck-Hufnagel (2000) was reported based on a comparison between monosyllabic and bisyllabic words, thus edge-related lengthening

effects, specifically, initial lengthening and final lengthening, could be a potential confounding factor.

Also, White (2002) pointed out that it is not clear in Port's (1981) study, and other studies that reported polysyllabic shortening (e.g., Nakatani et al., 1981) whether the shortening effect is due to the word size, or phrase size, since the number of syllables in the test sentences was not controlled for as syllables added to target bisyllabic and trisyllabic words.

Turk and Shattuck-Hufnagel (2000) attempted to control for some confounding factors in examining polysyllabic shortening at the word level. First, they control for the effect of final lengthening by examining left-headed (e.g., tune vs. tuna) and right-headed (choir vs. acquire) words, as in the latter final lengthening should be constant in all target words. Second, accent effects are accounted for by examining syllables' duration in accented and unaccented contexts. Third, target words appeared in phrases, within carrier sentences, that contained the same number of syllables to control for the effect of phrase size (e.g., tune acquire vs. tuna choir). Turk and Shattuck-Hufnagel (2000) found polysyllabic shortening in left-headed and right-headed words, both when accented and unaccented, as stressed syllables were shorter in bisyllabic than in monosyllabic words. Importantly, polysyllabic shortening was greater in magnitude in accented than in unaccented words, thus suggesting a dependency of polysyllabic shortening on the presence of accentual lengthening.

White (2002) tested polysyllabic shortening at the word level and designed his test materials by controlling for several confounding factors. The effect of polysyllabic shortening on stressed syllables' duration was tested in left-headed (mace vs. mason) and right-headed (mend vs. commend) words, both when accented and unaccented. In this way, final lengthening is controlled for as it is expected to be constant in right-headed words. The effect of phrasal accent is controlled for as target words either bore an accent or were unaccented. The target words were monosyllabic, bisyllabic, and trisyllabic, to disentangle edge-related lengthening effects from potential polysyllabic shortening in monosyllabic and bisyllabic words. For example, in right-headed mend vs. commend, the onset [m] in /mend/ may be a target for word-initial lengthening, thus confounding the potential polysyllabic shortening process in commend. However, in commend vs. recommend, initial [m] in /commend/ is not word-initial, which would allow for controlling for the effect of word-initial lengthening. The number of syllables in the test sentences was kept constant by adding syllables in sentences

that contained monosyllabic and bisyllabic test words. A Sample of test materials from White (2002)'s study is shown below.

Left-headed unaccented	Right-headed unaccented
I SAW the mace unreclaimed AGAIN	JOHN saw Jessica mend it AGAIN
I SAW the mason reclaimed it ALL	JOHN saw Jessie commend it AGAIN
I SAW the masonry cleaned AGAIN	JOHN saw Jess recommend it AGAIN

Left-headed accented	Right-headed accented
I saw the MACE unreclaimed again	John saw Jessica MEND it again
I saw the MASON reclaimed it all	John saw Jessie COMMEND it again
I saw the MASONRY cleaned AGAIN	John saw Jess RECOMMEND it again

White (2002) found that polysyllabic shortening effects were consistent in accented words. In unaccented words, however, polysyllabic shortening effects were only found for left-headed words. Given that polysyllabic shortening was not consistent in all prosodic prominence levels (only in accented words and only in left-headed unaccented words), it cannot be considered a general timing mechanism in speech production. White (2002) suggested interpreting polysyllabic shortening as timing effects distribution at domain heads and domain edges.

White (2002) interpreted polysyllabic shortening in accented words as the attenuation of accentual lengthening in longer words. In particular, in monosyllabic words, stressed syllables receive all prosodic lengthening due to accent, but in bisyllabic and trisyllabic words, since accentual lengthening targets stressed and unstressed syllables (Turk & White, 1999), lengthening effects on the stressed syllables attenuate.

As for unaccented left-headed words, White (2002) interpreted polysyllabic shortening as progressive final lengthening. Specifically, as closer proximity to the word edge leads to greater lengthening effects, stressed syllables in monosyllables would be lengthened more than in bisyllables, and in bisyllables stressed syllables would be lengthened more than in trisyllables.

Studies that reported constituent size effect on subconstituents' duration (natural rhythm unit: Bouzen and Hirst, 2004; inter-stress intervals: Kim and Cole, 2005; Krivokapić, 2013) did not control for positional effects, which would have strong effects on subconstituents' duration. Hirst (2009) took into account the position of phonemes in the natural rhythm unit (Jassem et al., 1984) in a corpus of British English. Hirst found that the position of phonemes, initial, medial and final, within the natural rhythm unit had a significant effect on phonemes' duration; however, the size of the natural rhythm unit did not, once the position of the phoneme was taken into consideration.

The higher-level effect of an oscillatory process that slows down the time flow of an utterance at its boundaries ($\pi - gesture$), thus leading to lengthening of segments adjacent to the boundary (Byrd & Saltzman, 2003), is not supported by empirical data, at least in American English. Turk and Shattuck-Hufnagel (2007) found that final lengthening is a rather complex process targeting two different sites in final words. It starts with the final stressed syllable and affects the absolute final syllable, skipping segments between those targets. Such a complex process is not accounted for by the ($\pi - gesture$) model.

1.5.3 Interim summary

In the previous sections, we reviewed two different theoretical models that attempt to capture the distribution of speech timing to signal the prosodic structure of speech. The coupled oscillators model (O'Dell & Nieminen, 1999) asserts that the hierarchical relation between prosodic units, e.g., stress feet and syllables, implies an interaction between higher-level and lower-level constituents in surface timing. Empirical evidence for such interaction comes from the findings that the foot's duration increases as the number of syllables within the foot increases, and the size of the foot, in terms of the number of syllables, causes compression to syllables' duration (Eriksson, 1991; Kim and Cole, 2005). This timing interaction is modelled with a system of coupled oscillators, in which each level of the prosodic structure is associated with its own natural frequency. The relative coupling strength between the oscillators is what differentiates between languages. For example, languages with strong stress-foot effects on syllable's duration, i.e., strong compression effects, have greater dominance of the stress foot's oscillator relative to the syllable's oscillator. On the other hand, languages with less syllabic compressibility effects, e.g., Spanish, have greater dominance of the syllable's oscillator. Thus, in the coupled oscillators model, the prosodic

hierarchy is signalled through compressibility effects, e.g., the grouping of syllables within stress feet, with gradient differences between languages in the strength of compressibility effects.

In the locus and domain view (White, 2002, 2014), the prosodic structure influences speech timing at domain heads and domain edges through phonologically specified segments (cf. White, 2002, 2014; van Santen, 1992; van Santen & Shih, 2000; Pointon, 1980). For instance, in English, lexical stress targets the stressed syllable, with greater lengthening effects localised to the syllable's nucleus (Klatt, 1976). The domain of phrase-final lengthening is the word's rhyme, targeting the stressed syllable and the absolute final syllable. Syllables within these targets may be skipped from phrase-final lengthening (Turk & Shattuck-Hufnagel, 2007). Stress-related and edge-related lengthening are widely attested among languages, with gradient differences in the magnitude of lengthening, especially in stress-related lengthening.

As prosodic structure influences timing through the lengthening of phonologically-specified segments, compressibility effects due to constituent size are refuted in the locus and domain approach. Reported constituent size effects (e.g., Port, 1981; Nakatani et al., 1981) may be re-analysed with local timing influences. For example, Port (1981) reported stressed syllable's compression due to constituent size effect in what appear to be accented words. This may be re-analysed as the attenuation of accentual lengthening in longer words. As accentual lengthening targets the whole word, the stressed syllable in monosyllables receives all lengthening effects, while in bisyllables and trisyllables, lengthening effects attenuate at the stressed syllable, since unstressed syllables receive some degree of the accentual lengthening (White, 2002). In unaccented words, compressibility is only observed in left-headed words, which may be interpreted as progressive word-final lengthening (White, 2002). Because different interpretations can be offered for constituent size effect, and due to the inconsistency of constituent size effect (in accented words and in left-headed unaccented words only), it cannot be considered as a general timing mechanism in speech production.

Although compressibility effects, which are implied by the coupled oscillators, are not supported by empirical data, there are some aspects in the coupled oscillators model that may be useful in accounting for timing variation in certain speech situations, such as synchronized speech and timing interaction in natural dialogue. In the following section, we will show how

phase relations between different rhythmic time scales may be useful in capturing temporal coordination patterns in speech.

1.5.4 Entrainment to structure in speech communication

Our aim in this section is to show that phase relation between different rhythmic time scales, which is an integral function in the coupled oscillators model (O'Dell & Niemenin, 1999), can be useful in accounting for temporal coordination between interlocutors, such as in synchronized speech and in dialogue interaction.

We discussed Cummins's (2003, 2009) work earlier, which showed that when two speakers read a text together, they could synchronize their speech, with small differences ranging between 40 ms to 60 ms. Such ability to synchronize speech may reflect the interaction between two oscillatory periods that are coupled and settled at a stable frequency. It may be argued that the ability to synchronize speech is the result of one speaker leading the joint speech and the other lagging behind following the speech rate of the leader speaker. However, Cummins (2003, 2009) showed that there was no consistent leader in the joint speech. Thus, synchronous speech is achieved through temporal coordination, or entrainment, between two interlocutors, and such entrainment may adequately be represented with a model of coupled oscillators.

Importantly, entrainment is not only a process of temporal alignment; rather, it is mediated by structural aspects of speech, such as prominence. For example, Wagner (2019) examined gesture-speech temporal coordination in a drumming task. German participants listened to audio recordings of read sentences and were instructed to drum to the perceived syllables after they had listened to the audio recordings. The intensity of the participants' drums in dB was recorded. It was found that the number of participants' drums was similar to the number of syllables in the test sentences, thus showing temporal alignment. Importantly, the intensity of participants' drums matched the prominence level of syllables. Similarly, when participants were asked to drum to perceived words, drums intensity matched the prominence level of words. Thus, the entrainment process not only involves temporal alignment but also alignment to the abstract prominence structure of speech. As such, entrainment is an important process in the perception and interpretation of speech prosodic structure.

Modelling entrainment between two interlocutors as a system of interacting oscillators involves positing that these oscillators are coupled, or more specifically, that they have *in-phase* relation, such that oscillators representing the utterances of two interlocutors are timed together. The example of synchronous speech represented by Cummins (2003, 2009) is certainly a good example of in-phase relation between two oscillators reflected in the minimum time difference in the speech of two interlocutors. While temporal coordination found in synchronous speech is only produced under an experimentally controlled situation or in other conventionalised speech setting such as prayer recitation and chanting, other evidence from spontaneous speech exists. For example, Włodarczak et al. (2012a,b) studied overlapped speech from corpora of spontaneous speech of American English, German and French. They found that overlapped speech tends to start at the syllable boundary of the overlappee's speech. Specifically, speakers timed their overlapped speech at the vowel onset of the overlappee syllable. This represents evidence for in-phase relation in surface timing, resulting from coupled oscillators of the speech of two interlocutors. Figure 1.5 represents an oscillatory period that entrains with vowel onsets period through in-phase relation.

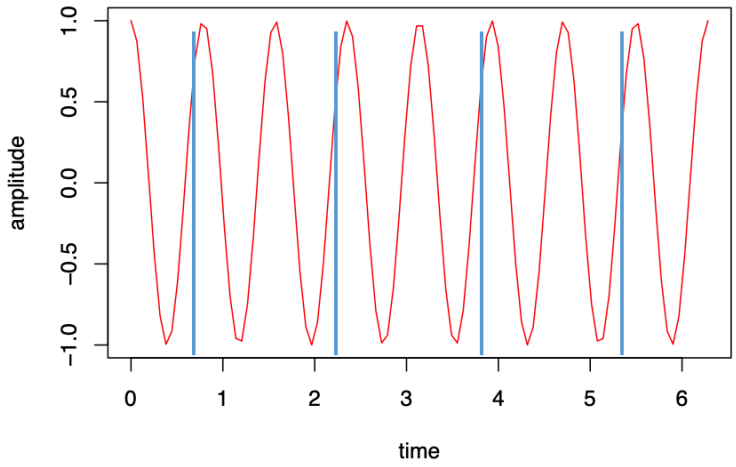


Figure 1.5: An oscillatory period (red) entraining with the period of stressed vowel onsets (blue spikes) through in-phase relation. Adapted from: Wagner et al. (2013, p. 124).

Moreover, they showed that language-specific patterns arise in overlapped speech. English and German speakers timed their overlapped speech just before the vowel onset, at the onset consonant(s) preceding the vowel, while French speakers timed their overlapped speech closer to the vowel onset. It is possible that this pattern is related to p-centre phenomena, the perceptual moment of syllable occurrence (Marcus, 1981; Scott, 1993). It is shown in the p-centre literature that p-centres occur earlier than the vowel onset when there are more onset

consonants in the syllable. Perhaps, since English and German have more complex onset consonant clusters than French, speakers of the former languages tended to initiate overlaps just before the vowel onsets, at onset consonant(s).

From Włodarczak et al.'s (2012a,b) studies, entrainment in speech involves alignment to structure, in this case, syllables, and within syllables, it is vowel onsets (or p-centres) that represent the anchor point for in-phase timing relation in the speech of two interlocutors. Language-specific phonological patterns also affect entrainment, as the difference between English and German and French speakers suggests.

Since phase relation between oscillatory periods has an integral part in the model of coupled oscillators and in explaining entrainment in speech communication, we will describe in the following sections two important, relevant concepts. The first is the phase relations in the inter-limb coordination, and the second is the p-centre, which represents the anchor point for temporal coordination in speech.

1.5.4.1 Temporal coordination in inter-limb movements

Temporal coordination among limb and finger movements has been investigated in the experimental work of Kelso et al. (1979). In this work, participants were asked to move (wage) their fingers in a cyclic manner. Particularly, they were asked to move their fingers to the right and the left direction at the same time. The speed of the fingers' movements was controlled by a pacing metronome, which increased gradually across different trials. The dependent variable in this constrained task was the relative phase of the two fingers, i.e., the difference in the phase of the fingers' movements. The main finding was that participants favoured a certain pattern of the two fingers' movement in which the phase lag between the two fingers was zero. Compared to other phase relations, there was a small amount of variance in the production of a zero lag pattern, which means that this pattern was the most stable. This phase lag of zero, reflects a synchronous pattern, in which the fingers move in simultaneous fashion toward and away from the midpoint of the body. The next most stable is when both fingers move right and left simultaneously. This pattern was termed in Kelso et al. "anti-synchronous" with 1/2 phase lag between the fingers. Moreover, while both the synchronous and anti-synchronous patterns were stable at slow rates of the pacing

metronomes, only the synchronous phase pattern was stable at faster rates. Figure 1.6 represents the in-phase (synchronous) and anti-phase (anti-synchronous) finger movements.

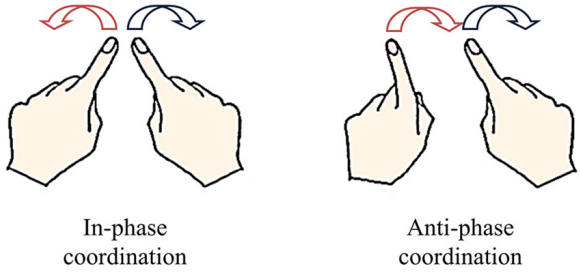


Figure 1.6: In-phase and anti-phase finger movements. Source: Malkoun et al. (2014, p. 4)

The stability of certain phases in the fingers' movements was accounted for by Haken et al. (1985) in a model of an underlying dynamical system of two competing attractors. At the slow rate, two attractors, one representing phase 0 and the other representing the 1/2 phase, are present, while at the faster, only the phase 0 attractor is present. When both attractors are present at the slow rate, it is the 0 phase attractor that is most stable. The existence of multiple attractors in the system is captured by Haken et al. (1985) through a potential function, which is defined as the superposition of two cosine waves, which have a frequency ratio of 1:2:

$$V(\phi) = -a\,cos\phi - b\,cos2\phi$$

The ratio of amplitudes of the two cosine waves b/a corresponds to the control parameter (rate). At a given value of the ratio of the amplitudes (b/a), say, for example, 1, the system is multi-stable, exhibiting stable phases at -0.5, 0, and +0.5. However, as the ratio lowers, corresponding to an increase in rate speed, the phases at -0.5 and +0.5 become less stable, and eventually, only the synchronous phase of 0 becomes stable, attracting the fingers' movements to it. Figure 1.7 shows a schematic illustration of the potential function $V(\phi)$ at different b/a values. As b/a value decreases, the attractors at –π (-0.5) and +π (+0.5) become shallower, until there is only a single attractor at phase 0. Furthermore, once the system is moved from the anti-synchronous state, i.e., the -0.5 and +0.5 phases, to the synchronous state, i.e., the 0 phase, at the lowest ratio value, it will remain at this state even when the ratio is increased further.

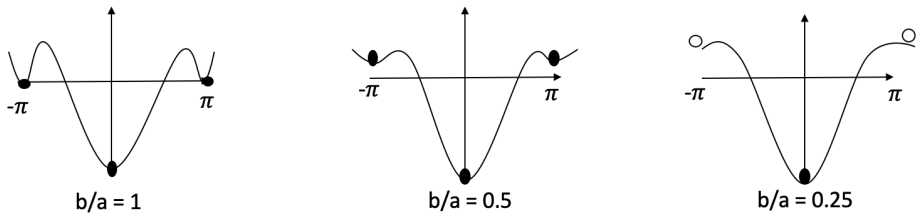


Figure 1.7: An illustration of the potential function $V(\phi)$ at different b/a values. As b/a value decreases (corresponding to an increase in the rate of finger movements), the attractors at –π and +π become less stable until there is only a single stable attractor at phase 0. The filled dots indicate stable attractors, while the unfilled dots indicate unstable attractors. Adapted from Haken et al. (1985, p. 350) and Kelso (1995, p. 11)

Another version of the inter-limb/fingers coordination task is presented by Yamanishi et al. (1980) and Tuller and Kelso (1989). There, subjects were asked to match their fingers' movements with an externally presented signal. In Yamanishi et al. (1980), two pacing metronomes repeating every 1000 ms with a phase difference between the two metronomes at ten values, ranging from 0 to 0.9, were displayed to the participants. One of the metronomes was presented to tap out by the left hand, and the other metronome was presented to tap out by the right hand. The participants were instructed to tap their fingers in synchrony with the two metronome phases. Participants underwent training for the task, and when the difference between the right hand tap and the left hand tap deviated from the target phases by more than +-0.5, a warning signal ("Short" or "Long") was displayed for them. After the training session, they started the actual trial and went through ten cycles of the pacing metronomes with ten values of phase difference. The subjects were asked to continue tapping after the tenth cycle of the pacing metronome and to try to maintain the ten-phase difference of the metronomes. Their main findings were that phases 0 and 0.5 were produced more accurately than other phases and with less variance. These two phase patterns were considered attractors for the fingers' movements. Yamanishi et al. concluded that the coordination relation in finger tapping is controlled by a neural mechanism that is composed of neural coupled oscillators, controlling the movement of the right and left fingers.

In Tuller and Kelso's (1989) study, findings were similar to that in Yamanishi et al. (1980) in that there exist attractors at phases 0 and 0.5; however, there were larger deviations from stable phases than in Yamanishi et al. (1980) but with less standard deviation. Tuller and

Kelso (1989) asserted that the performance in the finger tapping task provides evidence for an intrinsic rhythm that is resultant from the coupling of two oscillators that attract finger movements.

1.5.4.2 Syllable beats or p-centres

We have referred earlier to the concept of anchor points within syllables which listeners may entrain to their period (Cummins, 2003, 2009; Włodarczak et al., 2012a,b). This concept, i.e., anchor points in syllables, is referred to as syllables' perceptual centre of occurrence (p-centre). It was first introduced when Morton et al. (1976) found that a series of syllables, such as ba, pa, ba, pa, with isochronous acoustic onsets, were not judged by participants as being isochronous. This led them to the conclusion that there is some point within syllables, not the physical onset, that accounts for their perceptual moment of occurrence, and perceptual isochrony may be achieved relative to the timing of these points. Later, Marcus (1981) developed the "rhythm adjustment task" in order to define the exact point of the p-centre. In this task, participants heard a sequence of syllables alternating with a sequence of click bursts, e.g., S-C-S-C-S, and their task was to adjust the timing of syllables until subjective isochrony is achieved, and the time point bisecting the sequence of clicks relative to syllables' onsets was considered the point of p-centre in the test syllables. This is shown in the schematic representation in Figure 1.8.

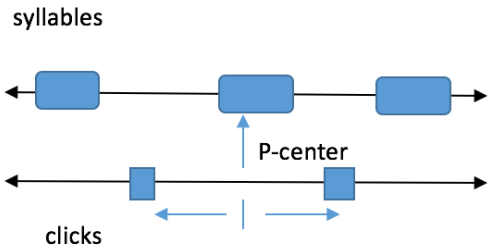


Figure 1.8: Schematic representation of the rhythm adjustment task. Clicks had to be adjusted relative to syllables until perceived isochrony is achieved. The time point bisecting clicks intervals relative to syllables onsets was considered as syllables' p-centres. Adapted from Pompino-Marschall (1989, p. 166).

The main finding in Marcus's (1981) study was that the p-centre occurred close to the vowel onset. In another experiment, the duration of onset consonants, the vocalic nucleus, and the

coda consonants were manipulated. It was found that the p-centre occurred earlier than the vowel onset as the onset consonants increased in duration and later as the vocalic nucleus and coda consonant durations increased. The effect of onset consonant durations, however, was of greater magnitude than the effect of vocalic nucleus and coda consonant durations.

The research after Marcus's (1981) work continued to uncover the acoustic effects on p-centre location and provided several models to predict the point of p-centre. For example, Pompino-Marschall (1989) attempted to predict the p-centre position based on the centre of gravity of the energy envelope calculated over the entire range of the syllable. An effect of the energy envelope rise time was also reported by Howell (1984, 1988a,b). Howell found that an increase in the envelope rise time shifted the P-centre later. Scott (1993) provided a model to predict p-centre position based on local rise time of the energy envelope. Her model is called the Frequency Dependent Amplitude Increase Model (FAIM). In the FAIM model, the p-centre is localised at 50% of the energy envelope's peak value. Scott's (1993) model obtains the energy envelope of a sub-band of the speech signal ranging from (500-1500 Hz). The choice of this sub-band was made after estimating the p-centre for "one" and "two" English words in a rhythm adjustment task. The energy envelope rise time of this sub-band, which corresponds to the F1 formant frequency, was found to be a better predictor of p-centre than higher frequencies. A schematic representation of Scott's model is provided in Figure 1.9.

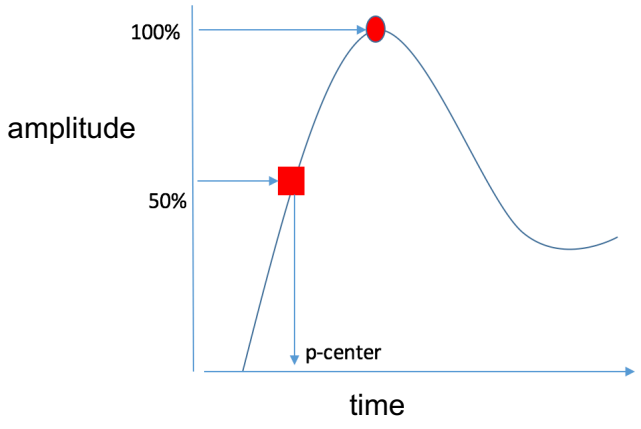


Figure 1.9: Scott's (1993) model for predicting the point of p-centre relative to a local rise in the energy envelope at 50 %.

de Jong (1994) investigated articulatory effects on p-centre. Several articulatory events from the tongue, the lip, and the jaw from a male speaker reading the tokens “toast” and “totes” were recorded using an x-ray microbeam system. Then, a rhythm adjustment task was conducted to define the p-centre relative to “toast” and “totes” test syllables. The p-centre point was in the vicinity of the vowel onset. Afterwards, the effect of the articulatory events, other acoustic correlates on p-centre location was examined. de Jong found that no single articulatory or acoustic event correlated with p-centre position; rather, it was a complex effect of multiple articulatory and acoustic events.

Allen (1972a,b) focused on the relevance of p-centre to the metrical structure of speech. Allen asked colleagues to transcribe utterances obtained from conversational speech, by defining the stress degree of syllables, on a scale that ranges from 0 (lowest stress) to 10 (highest stress). Then, in three experiments, Allen tested the ability of English participants to locate syllables’ beats. In the first experiment, participants were asked to tap their fingers to the beat of certain syllables in test utterances. In the second experiment, participants were asked to adjust clicks with a knob so that they sound on the beat of the target syllables within the test utterances. In the third experiment, participants heard a click, which was placed within the range of 600 ms, centred around the mean of participants’ click adjustments in the second experiment, and were asked to judge whether the clicks were *on the beat* of the target syllable or not. The general finding was that the beat of the target syllables was near the onset of the syllable vocalic nucleus. The exact location of the participants’ taps relative to vowel onsets was negatively correlated with the number of onset consonants, i.e., occurred earlier than the vowel onset when more consonants preceded the vowel, and were positively correlated with the length of the rhyme, i.e., occurred later than the vowel onset for longer rhymes. Also, participants showed higher consistency in aligning taps/clicks with the vowel onsets as the test syllable carried a higher degree of stress (see also Rathcke et al., 2021 for a recent tapping study with similar findings).

P-centre phenomenon was also observed cross-linguistically. Hoequist (1983) compared the performance of Japanese, Spanish and English participants in a production study. He asked the participants to utter a series of CV syllables in synchrony to a sequence of metronomes. He found that the vowel onset was the closest point in all participants’ production to the metronome pulse. Barbosa et al. (2005) had a Brazilian speaker repeat CV syllables to alternating metronome clicks. They found that the speaker aligned the vowel onset to the

metronome click and that the exact location within the vowel onset was affected by metronome rate and the spectral balance of onset consonants. Šturm and Volin (2016) had Czech speakers repeat Czech words with varying degrees of syllabic complexity to metronome pulses. The closest point to metronomes within the words was the vowel onset of the initial stressed syllable, and the location within the vowel onset varied according to the onset consonant durations and, to a lesser degree, the durations of vocalic nuclei and coda consonants.

1.6 Speech cycling

Having established the concept of entrainment to speech structure and the notions that define entrainment, i.e., *in-phase* relation and anchor points within syllables (vowel onsets or p-centres), we now introduce speech cycling, an experimental paradigm that captures entrainment between vowel onsets and higher-level prosodic units in speech production. We utilise this paradigm to explore dialect-specific patterns in Hadari and Bedouin Kuwaiti dialects in entraining vowel onsets of stressed syllables with higher-level prosodic units.

Speech cycling is an experimental paradigm developed by Cummins and Port (1998) and Tajima (1999) to uncover the temporal patterning of hierarchically-nested speech units. In speech cycling, speakers repeat a phrase together with metronome beeps. The interval between repetitions is called the phrase repetition cycle (PRC). It is shown that acoustically salient points, namely vowel onsets of strong syllables (Allen, 1972a,b), tend to lie at certain privileged phases within the PRC. These phases divide the PRC into simple integer ratios, such as 1/3, 1/2 and 2/3, that reflect a certain metrical structure within the PRC. Importantly, the simple phases reflect a hierarchical structure: vowel onsets of strong syllables are constrained to lie at these simple phases within the period of the PRC.

There are two different experimental versions of speech cycling. In Cummins and Port (1998), speakers hear two metronome beeps: a high-tone (H) beep and a low-tone (L) beep, and they repeat simple phrases such as: “Beg for a Dime” with the metronomes multiple times. The requirement was to align the first stressed syllable with the H metronome beep and the second stressed syllable with the L metronome beep. The time interval between H-L metronomes was kept constant at 700ms, while the interval between L beep and H of the following repetition was manipulated. Specifically, using phase conventions, the relative

phase of the L beep between the H-H beeps was manipulated to lie between 0.3 to 0.7 phases within the H-H cycle. This is shown schematically in Figure 1.10.

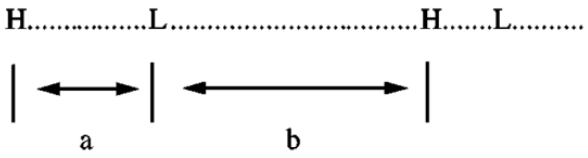


Figure 1.10: A schematic representation of the stimulus used in Cummins and Port (1998). The interval from H tone to L tone, a, was set constant at 700 ms. The relative phase, or timing, of the target L tone, was manipulated by varying the interval from the L tone to the H tone, b, between 0.3 and 0.7 phases. Source: Cummins and Port (1998, p. 152).

The dependent variable was the phase of the second (final) stressed syllable within the Phrase Repetition Cycle. The phase variable is computed by dividing the interval from the vowel onset of the first stressed syllable to the vowel onset of the second (final) stressed syllable by the interval from the vowel onset of the first stressed syllable in the phrase to the vowel onset of the first stressed syllable of the following repetition. This is shown in Figure 1.11.

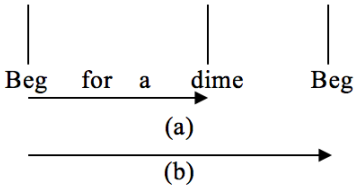


Figure 1.11: Computation of the phase variable of the second (final) stressed syllable in the phrase: "Beg for a dime". Interval a is divided by interval b to yield the phase of the second (final) stressed syllable.

Despite having multiple phase targets of the L beep, participants in Cummins and Port (1998) placed vowel onsets of the final stressed syllable at three distinct phases: 1/3, 1/2 and 2/3. These simple phases are said to be attractors to prominence in the PRC that emerge from repeating sentences at a constant period, and establish a certain metrical organization. The 1/2 phase reflects the structure of two metrical feet in the PRC: [Beg for a] [dime]. The 1/3 phase reflects a structure of two metrical feet, with a silent foot at the end of the phrase: [Beg for a] [dime] []. The 2/3 phase reflects the structure of three metrical feet, with an artificial

prominence assigned to the function word “for”: [Beg] [for a] [dime]. These findings support the notion that there is a metrical hierarchical relation between vowel onsets and the PRC, as vowel onsets are constrained to lie at simple phases within the PRC that establish a certain metrical structure.

Tajima (1999) used a non-targeted version of speech cycling in which speakers repeat phrases with only one metronome beep at the beginning of the cycle. The rationale for not using a targeted version of speech cycling is the lack of lexical stress in Japanese. The alignment of vowel onsets at simple phases of target L-tone metronomes may be facilitated with strong contrast between stressed and unstressed syllables as in English. On the other hand, the difference in prominence between moras in Japanese is not as strong as the stress contrast in English. Therefore, Japanese speakers may have difficulties in the alignment of certain moras with target L metronomes. Tajima pointed out, however, that accented moras in Japanese may show tendencies to align with simple phases in the PRC if the task is simplified by having speakers repeat sentences with a single metronome in the PRC. We will discuss in more detail in section 1.6.1 the effects of targeted speech cycling on the performances of speakers between languages with different prominence systems.

The general findings from Tajima (1999), which compared the performances of English speakers with Japanese speakers, are similar to that of Cummins and Port (1998): vowel onsets of prominent (accented in Japanese) syllables showed tendencies to align with simple phases in the PRC, such as 1/3, 1/2 and 2/3.

Tajima (1999) differentiated between two phase measurements: the external phase and the internal phase. The external phase measures the phase of the final prominent syllable in the phrase relative to the PRC, while the internal phase measures the phase of the medial prominent syllable relative to the initial and final prominent syllables within the phrase. The measures of the external and internal phases are shown in Figure 1.12, with the English phrase: “`Buy the `girl a `doughnut” as an example.

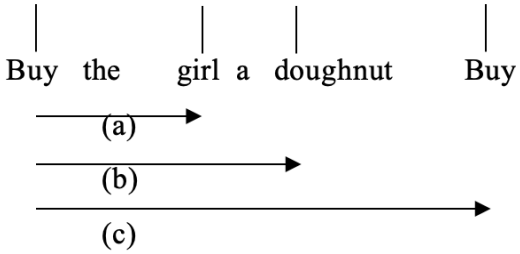


Figure 1.12: A schematic representation of the computation of the external and internal phases. The external phase is computed by dividing interval b by c (b/c), and the internal phase is computed by dividing interval a by b (a/b).

1.6.1 Cross-linguistic differences in speech cycling

Since the main objective of the dissertation is to explore rhythmic differences between dialects, it is imperative to demonstrate how language-specific differences may arise in speech cycling and to form predictions for dialect-specific differences based on the reported cross-linguistic differences in speech cycling.

Languages can vary with regard to their ability to align prominent syllables at beats, or metrically important points at given points in time based on the degree of temporal contrast between strong and weak syllables. For example, Cummins (2002) asked English, Spanish and Italian speakers to read sentences with two stressed syllables, each followed by an unstressed syllable, and to align the first stressed syllable to a high-tone beep and the second stressed syllable to a low-tone beep. English speakers found the task easy to perform. On the other hand, Italian and Spanish speakers found it more difficult, as they needed more than 30 minutes of practice, and even after practice, they found the task extremely uncomfortable. A possible explanation for these differences can be made by referring to the degree of temporal contrast between strong and weak syllables. English is known to have a high degree of temporal contrast between stressed and unstressed syllables, marked by lengthening of stressed syllables and vowel reduction of unstressed syllables, which in turn adds to the contrast between stressed and unstressed syllables. On the other hand, the degree of temporal contrast between strong and weak syllables in Italian and Spanish is lower than in English, with less stress lengthening, especially for Spanish, and limited exploitation of vowel reduction in unstressed syllables (Grabe & Low, 2002; White & Mattys, 2007a). Since simple phases in the PRC are attractors to prominence, the higher degree of stress contrast in English

affords emphasising the stressed syllables' beats, through closer alignment to simple phases within the PRC. On the other hand, the lower contrast between stressed and unstressed syllables in Italian and Spanish makes the task unnatural, as high degrees of emphasis on stressed syllables relative to unstressed syllables are not available in their language. In other words, stress beats are more salient in English than in Italian and Spanish; thus, English speakers would find the closer alignment of stressed syllables to simple phases, compared with unstressed syllables, more natural and easier than Spanish and Italian speakers. Another example is from Zawaydeh et al. (2002), who compared English to Jordanian Arabic in the non-targeted version of speech cycling. In this study, English speakers were able to align stressed syllables closer to a simple phase of 1/2 than Arabic speakers, who aligned stressed syllables further away from the 1/2 phase. Jordanian Arabic employs vowel reduction of unstressed syllables less than English (Vogel et al., 2017), which makes the alignment of stressed syllables to simple phases less affordable in the former. Thus, it is the degree of temporal contrast between strong and weak syllables that affords the alignment to simple phases.

Another factor that would afford the alignment to simple phases is syllabic compressibility, which is a consequence of the higher degree of contrast between strong and weak syllables. Tajima (1999) compared English and Japanese speakers' performance in the non-targeted speech cycling. The external phase of the final stressed/prominent syllable was examined. Speakers read sentences at different metronome rates, varying from slow to fast. The rationale for speaking at different rates is to see whether speakers of English and Japanese would be persistent in keeping the final stressed syllable at a simple phase angle across different rates, or that rate increase will lead final stressed/prominent syllables to drift away from a simple phase angle. English speakers were able to keep final stressed syllables at a simple phase angle throughout the different metronome rates, while Japanese speakers showed an incremental change in the timing of final prominent syllables, placing them at later points in the cycle relative to a simple phase angle. Figure 1.13 clearly shows the difference between English and Japanese.

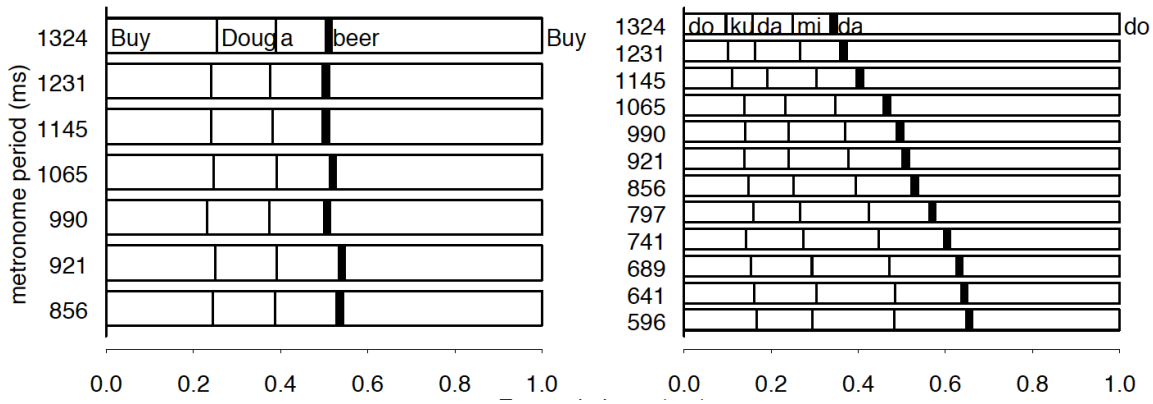


Figure 1.13: The change in the phase of final prominent syllables in English and Japanese as a function of rate increase. Only seven rates are provided for English, compared to twelve rates in Japanese. Source: (Tajima, 1999, p. 35)

From the example provided in Figure 1.12 of the English phrase, it seems that the long duration of stressed syllables afforded more compressibility of stressed syllables with increasing rate, which would afford keeping stressed syllables vowel onsets at a consistent phase throughout different metronome rates. Also, since English exploits unstressed vowel reduction, we would expect English to tolerate the compressibility of unstressed syllables, which would afford consistent phase alignment in English. On the other hand, shorter durations of prominent syllables and the absence of vowel reduction in non-prominent syllables in Japanese make compressibility intolerable, and consequently, the final prominent syllable would fall out of the stable, simple phase at faster metronome rates.

The effect of compressibility and its role in affording keeping stressed syllables at a simple phase angle persisted in Tajima's experiments when timing was manipulated by changing the number of syllables in English and Japanese rhythmic feet. This is shown below:

Pattern A: [`σσ] [`σσ] [`σ]

Pattern B: [`σσσ] [`σσ] [`σ]

The internal phase of the medial stressed syllable was examined. Tajima found that the difference in the alignment of the medial stressed syllable to a simple phase angle between pattern A and pattern B was smaller in English than in Japanese. This also reflects that compression of unstressed syllables affords keeping stressed syllables (here medial stressed

syllables) at a harmonic phase angle, while in Japanese, adding more syllables, combined with the absence of compressibility, would lead to a shift from a harmonic phase angle.

Thus, we may summarise the factors that would afford the alignment of prominent syllables to metrically important beats or points in time as follows:

a- A top-down effect, which refers to the relative perceptual contrast between strong and weak syllables. As simple phases in the PRC are attractors to prominent events in speech, greater perceptual stress contrast, e.g., through greater stress lengthening and unstressed vowel reduction, may afford closer alignment to metrically important points in time.
b- A bottom-up effect, which refers to the tolerance of phonetic compressibility of syllables. Longer durations of stressed syllables and vowel reduction in unstressed syllables would make the compressibility of stressed and unstressed syllables more tolerable. Hence, when timing is manipulated, such as with increasing rate, compressibility would afford keeping stressed syllables at simple phases throughout different rates.

Based on these proposed effects in speech cycling, we predict differences in phase alignment between Hadari and Bedouin since the former has greater stress contrast. We will detail the predictions regarding each dialect's performance in speech cycling in the following chapter.

Chapter 2. Corpus recording and Experiment 1 (a): phase measurements

2.1 Introduction

Three sets of analyses in a speech cycling corpus will be used to elucidate patterns of temporal coordination between stressed vowel onsets and the Phrase Repetition Cycle in Bedouin and Hadari Kuwaiti dialects. The first is phase measurements to explore different dialectal patterns in aligning vowel onsets at harmonic phase angles. As we mentioned in section (1.6.1), there are two effects that may lead to variable phase alignments between languages: a top-down effect and a bottom-up effect. The top-down effect refers to the relative degree of contrast between strong and weak syllables. For instance, stressed long syllables would afford stronger temporal coordination, i.e., closer alignment, with harmonic phase angles, than stressed short syllables. The bottom-up effect refers to the tolerance of syllabic compressibility, such as unstressed syllables reduction. As timing is manipulated in speech cycling tasks, e.g., through manipulating metronome rates, syllabic compressibility may afford consistent harmonic phase alignment at different metronome rates. The variable structural aspects of Bedouin and Hadari dialects are predicted to lead to variable phase alignment between the two dialects. Hadari has greater relative contrast between stressed and unstressed syllables than Bedouin, as the former tends to exhibit a greater degree of unstressed syllables reduction than the latter dialect. Therefore, we predict that Hadari would afford aligning stressed syllables closer to a harmonic phase angle than Bedouin. Also, greater tolerance to syllabic compressibility through unstressed syllables reduction in Hadari is predicted to afford consistent alignment to a harmonic phase at various timing manipulation conditions. We will demonstrate in section (2.8) more details on how dialects may differ in phase alignment with regard to the various experimental manipulations. We will also use the speech cycling corpus in other experimental chapters. In Experiment 1 (b), we will examine the syllabic durational profile in order to explain how dialect-specific patterns of phase alignment may arise from different syllabic durational patterns, such as unstressed syllables reduction. In Experiment 1 (c), we will investigate the metrical structure of stress beats in terms of relative prominence through an analysis of spectral balance. In Experiment 2, we will investigate the potential mutual timing influences between stress and syllables in speech cycling through an analysis of amplitude envelope statistics.

2.2 Methods

We used the non-targeted speech cycling (Tajima, 1999), where speakers align the beginning of sentences with a metronome beep, to investigate possible hierarchical timing differences between Hadari and Bedouin Kuwaiti dialects. We did not use the targeted version of speech cycling (Cummins & Port, 1998), where the speakers' task is to align stresses with high and low tone metronome beeps. We avoided the use of targeted speech cycling because it was shown in earlier experiments that speakers found this task to be difficult and needed long training sessions before they could do the task. In Cummins's (2002) study, Spanish and Italian speakers found targeted speech cycling difficult, which resulted in misalignment with target phases of low tone metronomes and extreme variability in their production. As discussed in section 1.6.1, this difficulty may be attributed to the low stress contrast in Spanish and Italian, which makes the task unnatural for Spanish and Italian speakers. On the contrary, English speakers, which have strong stress contrast, completed targeted speech cycling in Cummins's (2002) study more easily, aligning stressed syllables with different phase targets of low tone metronomes. As for Arabic dialects, despite gradient differences amongst them in stress contrast, they have lower degrees of stress contrast compared with English, especially with stressed light, CVC, vs. unstressed syllables, due to less unstressed vowel reduction. Therefore, to make the comparison in speech cycling between Hadari and Bedouin Kuwaiti speakers more plausible by avoiding potential extreme variation between the dialects, we used the non-targeted speech cycling (Tajima, 1999).

We first conducted a pilot study to see whether speakers could comfortably adapt to the task demands. The participants in the pilot study were three Hadari and three Bedouin speakers; all were male speakers. They were university students in Newcastle(UK), and recordings took place in a soundproofed room at Newcastle University. The sentences used were composed of six syllables. Each trial consisted of 12 metronome beeps at a constant rate. Participants' task was to listen to the first four metronome beeps at the beginning of the trial and then start repeating the sentence at the fifth beep and stop repeating with the last beep. There were ten trials, i.e., ten metronome periods, with the longest being 1800 ms and the shortest 963 ms. The metronome period of the following trial was 93 % of the previous one. Speakers were instructed to repeat sentences in each trial with a single breath and not to breathe between repetitions, as this was found to bias timing towards certain phases (Tajima, 1999; Cummins & Port, 1998), thus may not reflect prosodic constraints on phase production.

Speakers were told if they needed to take a breath during a trial, to skip a repetition cycle, and to continue with the next metronome cycle. Although speakers could repeat sentences in a single trial without inserting breaths between repetitions, there were significant disfluencies in their production, especially towards the shortest metronome periods. Therefore, we decided to use three comfortable metronome periods, long at 1800 ms, normal at 1512 ms, and short at 1270 ms. Speakers produced no disfluencies when repeating sentences at these metronome periods. It is noteworthy that using fewer metronome periods, three, compared with the number of metronome periods in Tajima (1999), fourteen, may contribute to a more plausible comparison between Hadari and Bedouin dialects. This is because a greater number of metronome periods may lead to more variability in rhythmic modes (e.g., 1/3, 1/2, and 2/3). However, a fewer number of metronome periods may lead to a distinct rhythmic mode (e.g., 1/3 only), to which we can compare the degree of close alignment that Hadari and Bedouin speakers may achieve.

2.3 Participants and recordings

After conducting the pilot experiment, the actual data were collected from 23 Bedouin speakers and 22 Hadari speakers; all male speakers. Recordings took place in a soundproofed room at the media lab in Kuwait University. Speakers were students in Kuwait university and participated voluntarily. The age range was from 21 to 28. Participants filled a questionnaire about their dialectal background. Bedouin speakers belonged to two Bedouin tribes: "Al-Shammari" and "Al-Enzi". Most of them lived in "Al-Jahra" city, which is located to the north of Kuwait City and most of its population consists of Bedouins. Some of Bedouin participants reported that although their dialect may have been subjected to some changes due to contact with Hadari speakers, they still speak in a way that sounds distinct from Hadari since their dialect is part of their social identity. Hadari participants have origins from Iraq, Iran, and Hasa (eastern part of Saudi Arabia). They lived in different areas in Kuwait, some of which have a majority of Bedouin speakers. However, they reported that this did not lead to changes in their dialect since Hadari is more prestigious than Bedouin (Rosenhouse, 2006).

Participants' speech was recorded using a Zoom H4 recorder (sampling rate 44100 Hz, quantization 16 bit). They spoke through an AKG D5 dynamic microphone, held to a stand, while reading the sentences presented on a computer screen. They heard the metronome beeps through Sennheiser headphones. Metronome beeps were played to the participants

through a MacBook Pro machine. Metronomes were also recorded on a second channel. Our objective to record metronomes was to compare the alignment of certain points in the speech to metronome beeps; however, these data were not used in the current analyses.

2.4 Materials

Participants read 12 sentences, six of which contained words with a trochaic stress pattern, and six contained words with an iambic stress pattern. Each sentence was composed of six syllables. Table 2.1 shows the sentences used in the study. Note that in Table 2.1 we presented the Bedouin pronunciation of words since it is more conservative than Hadari's pronunciation, with no reduction of unstressed vowels. All sentences were of a Verb-Subject-Object (VSO) syntactic structure. Consistency in syntactic structure was to avoid any potential effect of syntactic boundaries on timing. Stressed syllables were of CVC and CVV(C) structures. CVC stressed syllables will be termed light, while CVV(C) heavy. The reason to label stressed CVC syllables as light is because in Kuwaiti dialects and other Arabic dialects, when CVC and CVV syllables are in the same word, CVV receives stress (Watson, 2011a,b). In the trochaic pattern, there was an equal number of light CVC and heavy CVV(C) stressed syllables. In the iambic pattern, however, all stressed syllables in the first word were CVC syllables. There were two CVC stressed syllables and four CVVC syllables in the second word (there were no CVV stressed syllables in the iambic case because CVV syllables do not exist in the Hadari and Bedouin dialects in the iambic structure). There were four CVC stressed syllables and two CVVC in the third word. The uneven number of CVV and CVVC syllables was due to the limited number of words that have an iambic structure in a VSO type of sentences in Kuwaiti dialects.

2.5 Procedure

The procedures were based on the findings from the pilot experiment. Participants read each sentence in Table 2.1 at three different metronome periods: slow at 1800 ms, medium at 1512 ms, and fast at 1270 ms, where each metronome period corresponds to a single trial. The participants repeated all the sentences at the slow rate first, then at the medium rate, and ended with the fast rate. The amount of metronome period reduction from slow to medium and from medium to fast was constant at 16%. In each trial, participants heard 12 metronome beeps over headphones.

Table 2.1: Text material. Stressed syllables in bold. Bedouin pronunciation is shown since it is more conservative than Hadari's pronunciation, with no reduction of unstressed vowels. Syllable structure of each word is shown, with G indicating a geminate.

Trochaic	Iambic
mar:at **ba:s**ma **fa:**tˤ ma **CVG**.GVC **CVVC**.CV **CVVC**.CV "Baasma provided a ride to Fatma"	so**bag** na**be:l** ʕa**bi:r** CV.CVC CV.**CVVC** CV.**CVVC** "nabeel outran ʕabiir"
taakil **bas**ma **birħ**i **CVV**.CVC **CVC**.CV **CVC**.CV "Basma eats dates"	to**bax** ħa**mad** fa**geʕ** CV.**CVC** CV.**CVC** CV.**CVC** "Hamad cooked mushroom"
baaʕat **sal**ma **bad**la **CVV**.CVC **CVC**.CV **CVC**.CV "Salma sold a dress"	to**baʕ** ba**dir** ma**qa:l** CV.**CVC** CV.**CVC** CV.**CVVC** "Badir printed a paper"
ta:kil **mar**yam **ge:**mar **CVV**.CVC **CVC**.CVC **CVV**.CVC "Maryam eats sweat cream"	ka**lat** be**door** dʒo**bin** CV.**CVC** CV.**CVVC** CV.**CVC** "bodoor ate cheese"
bad:al **sa:**lim **da:**rah **CVG**.GVC **CVV**.CVC **CVV**.CVC "Saalim changed his room"	ni**ʃar** ra**bi**:**ʕ** xa**bar** CV.**CVC** CV.**CVVC** CV.**CVC** "Rabee distributed an information"
dazlik **ba:**sil **daʕ**wa(h) **CVC**.CVC **CVV**.CVC **CVC**.CV(C) "Baasil sent you an invitation"	ʃa**rab** ħa**bi:b** la**ban** CV.**CVC** CV.**CVVC** CV.**CVC** "Habeeb drank yougurt"

They were asked to listen to the first four metronome beeps to familiarise themselves with the speed of the beeps. Then, on the fifth metronome beep, they started to repeat a single sentence until the metronomes stopped at the 12th beep. They were asked to align the beginning of the sentence with the metronome beep. The total number of repetitions from each sentence was 8 in every metronome period trial. The total number of recorded tokens (repetitions) for Bedouin was *6624*, and for Hadari *6336* (see below for details on analysed tokens). Half of the participants started with the trochaic sentences, while the other half started with the iambic sentences. Each participant completed the task in approximately 20 minutes.

Before the actual experiment, participants had a training session. They repeated two sentences, not from Table 2.1, at a comfortable metronome rate. When they felt confident about doing the task, they started the actual experiment. This training session took around two minutes.

Participants were asked to repeat a sentence in a single breath and not take breaths between repetitions. They were told if they needed to breathe, to skip one cycle of repetition, and start repeating on the next one. In cases where speakers inserted a breath between two repetitions, the cycle was excluded from the analysis. When they finished repeating a sentence, they were given some time to breathe and drink water before repeating the next sentence. Also, the first, and the last (8^{th}) repetitions were excluded from the analysis to avoid transient effects.

The metronome beeps were generated using Praat (Boersma & Weenink, 2018) through a script made by Hugo Quenè (https://www.hugoquene.nl/tools/index.html) that the author adapted to match the design of the task. Each metronome beep was a 400 Hz pure tone.

The total number of tokens, i.e., cycles, that were analysed for the external phase measure was *5948*, and for the internal phase measure was *5665*. Note that the minor disparity between the number of analysed tokens between the external and internal phase measures is probably due to measurement errors in boundary marking. Tables 2.2, and 2.3 illustrate detailed tokens analysed in each condition of our experiment (see section 2.8) in Bedouin, and Tables 2.4 and 2.5 show Hadari's data.

Table 2. 2: Analysed tokens in the external phase measure for Bedouin. The total number is 2788.

Total tokens per trail	Slow		Normal		Fast	
	962		940		886	
Stress pattern	Trochaic	iambic	Trochaic	Iambic	Trochaic	Iambic
Heavy	246	149	237	146	219	143
Light	245	322	241	316	224	300
Total	491	471	478	462	443	443

Table 2. 3: Analysed tokens in the internal phase measure in Bedouin. The total is 2734.

Total tokens per rate trial	Slow		Normal		Fast	
	929		919		886	
Stress patterns	Trochaic	iambic	Trochaic	Iambic	Trochaic	Iambic
Heavy	252	288	249	286	239	272
Light	250	139	248	136	239	136
Total	502	427	497	422	478	408

Table 2. 4: Analysed tokens for the external phase measure in Hadari. The total is 3160.

Total tokens per trial	Slow		Normal		Fast	
	1049		1057		1036	
Stress pattern	Trochaic	iambic	Trochaic	Iambic	Trochaic	Iambic
Heavy	258	173	269	182	268	170
Light	263	355	263	361	259	339
Total	521	528	532	543	527	509

Table 2. 5: Analysed tokens in the internal phase measure in Hadari. The total is 2931.

Total tokens per rate trial	Slow		Normal		Fast	
	969		1007		955	
Stress patterns	Trochaic	iambic	Trochaic	Iambic	Trochaic	Iambic
Heavy	259	295	270	309	263	286
Light	267	148	274	154	268	138
Total	526	443	544	463	531	424

2.6 Data preparation

The recording was made for all trials, from the longest to the shortest metronome period, for each speaker without stopping the recorder. Repetitions of each sentence from each metronome period trial were marked with Praat TextGrid boundaries. We then used a Praat script made by Pablo Arantes (https://github.com/parantes/slicer) to extract each sentence's repetitions at each metronome trial into a separate audio file.

The data were annotated with the aid of a Praat script called *BeatExtractor* made by Barbosa (2003), which implements Cummins and Port's (1998) Beat Extractor script, with some modifications. The *BeatExtractor* detects vowel onsets by using a low-pass filtered energy envelope. It detects a certain energy threshold in the normalized envelope, 0.12 of amplitude peak. Boundaries are inserted on points that meet this threshold in the form of TextGrid boundaries. The script provided a good approximation in boundary placement at vowel onsets when the preceding consonant was an oral stop or a nasal; however, boundaries were after the vowel onset most of the time. The script added spurious boundaries for fricatives and approximants. Thus, boundaries were checked manually, according to segmentation procedures reported in the literature (Turk et al., 2006), to ensure they were at vowel onsets. Boundaries at the vowel-consonant transition and between consonants were also added for subsequent syllable duration analysis and spectral balance analysis. Segmentation criteria are provided below:

1- The start and end of the vowels were at the start and end of a pitch period at zero-crossing points in the waveform.
2- In vowel-oral stops sequences, the main determiner for vowel boundary was a drop in waveform amplitude associated with a change in formant structure. In oral stops-vowel sequences, the boundary was placed after the stop's release burst, or, if present, the boundary was placed after a period of frication following the release burst, before the appearance of vowel formant structure.
3- In fricative-vowel transition, the end of the fricative was marked by a drop in frication energy or by the end of silence phase after frication, before vowel formant structure appears. In vowel-fricative transition, the main determiner of the boundary was the onset of frication energy.

4- In nasals-vowel sequences, the end of nasals was decided based on an increase in amplitude waveform. In vowel-nasal sequences the end of vowels was decided based on the appearance of nasal formant structure and waveform amplitude minimum.
5- The start and end of trills after and before vowels were associated with a drop and an increase in waveform amplitude, respectively. Thus, boundaries were determined based on waveform amplitude change.
6- The main cue to the boundaries of laterals was the drop in amplitude in the higher frequency bands relative to the preceding and the following vowel.
7- The boundaries from approximants to vowels and vice versa were accompanied by a smooth transition in the formant structure. Changes in intensity at these transient points, coupled with a change in the waveform amplitude, were considered as the boundaries for approximants.
8- The pharyngeal (ʕ) is a special case of an approximant, realized with high tension in the glottis, which results in a drop in fundamental frequency after vowels. Thus, the boundaries were added at a drop in the fundamental frequency value from the previous vowel and a rise in fundamental frequency at the beginning of the following vowel (Haselwood, 2004).
9- Criteria mentioned above for laterals, such as lower energy at higher bands relative to vowels, and approximants, such as smooth formant transitions and changes in fundamental frequency at transients from and to vowels, were sometimes difficult to detect due to similar formant structure. In these cases, we based boundary placement on auditory impressions.
10- There were variable cases of sequences of heterosyllabic consonants. For oral stop-oral stop sequences, boundaries were placed after the disappearance of burst energy or after the disappearance of low energy frication following bursts, at the beginning of closure. For oral stops-fricatives, there was low energy frication following bursts most of the time, and boundaries were placed at points where high energy frication of fricatives was visible. When only bursts were visible, the different energy distribution of frication at different frequency bands than burst energy was considered the cue to the beginning of fricatives. When nasals were followed by fricatives and stops, or vice versa, nasal formant structure was considered for boundary placement. In approximant-stop and approximant-fricative sequences, boundaries were placed when the approximant formant structure disappeared. From trills to approximants, boundaries were placed at an increase in waveform amplitude. When a pharyngeal (ʕ)

was followed by another approximant, an increase in fundamental frequency value was considered as a point for boundary placement. When no clear increase in fundamental frequency was visible, boundaries were added based on auditory impression. In fricative-fricative sequences, different distribution of frication energy at different frequency bands was considered in boundary placement. For example, in x#ħ sequences, energy was at higher bands for x than ħ.

2.7 Phase measurements

Another text grid tier was added for vowel onset boundaries, and two phase measurements, shown in Figure 2.1, were computed:

1- The External Phase: The duration from the beat, i.e., the vowel onset of the first stressed syllable within a phrase to the beat of the last stressed syllable divided by the duration of the whole cycle. This measure defines the phase of the last stressed syllable.

2- The Internal Phase: The duration from the beat of the first stressed syllable to the beat of the second stressed syllable and divided by the duration from the beat of the first stressed syllable to the beat of the final stressed syllable. This measure defines the phase of the medial stressed syllable.

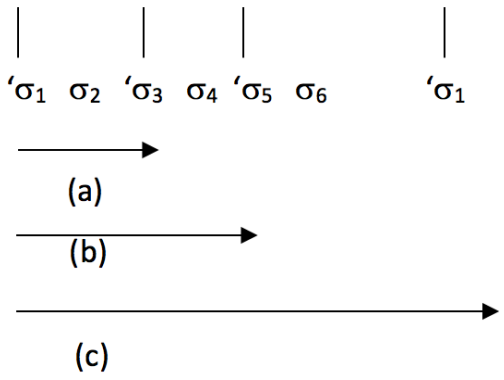


Figure 2.1: A schematic representation of the computation of the external and internal phases. The external phase is computed by dividing interval b by c (b/c) and the internal phase is computed by dividing interval a by b (a/b).

2.8 Predictions

We have mentioned two factors that would promote the alignment of stressed syllables to simple harmonic phases: top-down effect, such as phonological vowel length, and bottom-up effect, which refers to tolerance of syllabic compressibility. There are three factors in our materials that are expected to have an effect on Hadari and Bedouin speakers' phase alignment in speech cycling: syllable weight, stress pattern, and metronome period. We will detail below predictions on how each factor may influence phase alignment in Hadari and Bedouin dialects.

2.8.1 Syllable weight

Medial and final stressed syllables are either heavy, CVV(C), or light, CVC. The predicted effects associated with syllable weight are top-down in nature. Both dialects are expected to align heavy stressed syllables close to a simple phase angle. The phonological length of stressed, heavy syllables would provide greater affordance to the alignment to simple phases. The situation is different for stressed, light syllables. In Hadari, they contrast with unstressed syllables through vowel reduction in the latter, i.e., unstressed syllables. In Bedouin, the contrast between stressed, light syllables and unstressed syllables is of lesser magnitude than Hadari due to the limited exploitation of vowel reduction. Thus, the greater contrast between stressed, light syllables and unstressed syllables in Hadari is predicted to provide greater affordance to the alignment to simple phase angle. In contrast, the lower contrast between stressed light and unstressed syllables in Bedouin would afford the departure from simple phase angle.

2.8.2 Stress pattern

The trochaic pattern is expected to facilitate alignment to a simple phase angle more than the iambic pattern. That is, in the trochaic pattern, phrase-initial stressed syllables, more specifically, vowel onsets of initial stressed syllables can easily be aligned with the metronome beep, which signals an important metrical position, i.e., the beginning of the cycle. In the iambic pattern, however, it is more difficult to align vowel onsets of phrase initial stressed syllables to the metronome beep because phrase initial unstressed syllables naturally intervene between the metronome beep and the initial stressed syllables. Thus, if the

target phase of the initial stressed syllable, i.e., Phase 0, starts with the metronome beep, it would be easier to align the first stressed syllable close to Phase 0 in the trochaic pattern than in the iambic pattern. Thus, in the trochaic pattern, no shift from Phase 0 is predicted, while a shift is predicted in the iambic pattern. This shift in the iambic pattern would consequently, other things being equal, cause a shift from the simple phase angle, i.e., later alignment, of the medial and final stresses in the cycle.

The predicted effects of the trochaic and iambic patterns are bottom-up in nature, which refers to compressibility. Given that compressibility of unstressed syllables is more tolerable in Hadari than Bedouin, Hadari is predicted to compress phrase initial unstressed syllables in the iambic pattern to afford the alignment of phrase initial stressed syllables to the metronome beep. Consequently, Hadari would also align vowel onsets of medial and final stressed syllables close to simple phase angles. However, in Bedouin, less compressibility of initial unstressed syllables would promote greater lag of phrase initial stressed syllables relative to the metronome beat, and thus of the following medial and final stressed syllables from simple phase angles.

2.8.3 Metronome period

The predicted effect associated with metronome period is a bottom-up effect. As metronome period decreases, speaking rate may increase, and the temporal alignment of syllables may be disrupted due to increased speaking rate. We assess the persistence of internal and external phase angles by examining the effect of metronome period on phase angle change. As before, the exploitation of vowel reduction in Hadari is predicted to make compressibility more tolerable and would afford keeping a stable phase angle across different metronome periods. The limited vowel reduction in Bedouin, however, will make compressibility less tolerable, thus, in shorter metronome periods, the alignment of medial and final stressed syllables in Bedouin is predicted to occur later relative to a harmonic phase angle.

2.8.4 Higher-level interactions

Higher-level interactions are also predicted to influence phase alignment. Specifically, based on our primary experimental factors, we predict the following higher-order interaction:

1- Dialect, weight and metronome period might interact. Due to the relative salience of a simple phase angle in the case of heavy syllables, both dialects are predicted to resist temporal perturbation caused by the change in metronome period when uttering heavy syllables. However, in the case of light syllables, only Hadari is expected to show a persistent phase angle across the different metronome periods. This is because, as mentioned earlier, light syllables in Hadari are contrasted from unstressed syllables through vowel reduction in unstressed syllables. Vowel reduction is predicted to afford compressibility at shorter metronome periods, and would make the alignment of stressed, light syllables in Hadari to a simple phase angle more persistent across the different metronome periods. In Bedouin, however, the limited vowel reduction would make compressibility intolerable, and thus, the change in metronome period is predicted to perturb the alignment of light syllables relative to a stable phase angle.

2- A three-way interaction between dialect, stress pattern, and syllable weight is also predicted. Since stressed, heavy syllables are predicted to prompt closer alignment to a simple phase, stressed, heavy syllables might resist the temporal perturbation in the iambic pattern in both dialects. However, stressed, light syllables are predicted to resist temporal perturbation in the iambic pattern only in Hadari but not in Bedouin. As mentioned earlier, light syllables in Hadari are contrasted from unstressed syllables through vowel reduction, and vowel reduction would make compressibility of unstressed syllables more tolerable, hence resisting the maximal distortion to metrical alignment caused by the iambic pattern. On the other hand, less exploitation of vowel reduction in Bedouin, hence less tolerance to compressibility, would cause light syllables to be thrown off the simple phase alignment in the iambic pattern.

3- A four-way interaction between all four factors (dialect, weight, stress pattern, and metronome period) is also predicted. Both the shortest metronome period and the iambic pattern are expected to cause the maximal temporal distortions to the alignment to simple phase angles, and only heavy syllables, in both dialects, are expected to resist this temporal distortion. Light syllables, however, are expected to resist distortions to metrical alignment only in Hadari, while in Bedouin, resistance to temporal distortions is not predicted. Thus, the four-way interaction is predicted to arise because the shortest metronome period and the iambic pattern would cause maximal temporal distortions for light syllables in the Bedouin dialect only.

2.9 Statistical analysis

We ran linear mixed-effects models for both phase measures (external and internal). There were four predictors: dialect (Bedouin vs. Hadari), stress pattern (iambic vs. trochaic), syllable weight (heavy vs. light), and metronome period, which was treated as a continuous variable. Also, to test the effect of stress pattern, syllable weight, and metronome period on dialectal differences, two-way interactions between dialect and stress pattern, dialect and syllable weight, and dialect and metronome period were included in the model. In addition, Higher-order interactions that were expected to arise between the predictors were also included. These are three-way interactions between dialect, stress pattern, and syllable weight, and dialect, syllable weight, and metronome period, and a four-way interaction between dialect, stress pattern, syllable weight, and metronome period.

Other interactions were also included in the model to control for other sources of variance in the data. These are two-way interactions between stress pattern and syllable weight; stress pattern and metronome period; syllable weight and metronome period. Three-way interactions were also included between syllable weight, stress pattern and metronome period; dialect, stress pattern and metronome period.

In defining the random structure of the model, we initially included speaker and sentence as random intercepts. Also, by speaker random slopes for stress pattern, syllable weight and metronome period were included in the model, since these factors vary within speakers. By sentence random slopes for dialect and metronome period were included in the model as dialect and metronome period vary within sentences. However, in the external phase analysis, the model did not converge with by sentence random slope for dialect thus we removed this random slope from the random structure. For the internal phase measure, the model converged with the maximal random structure.

All categorical variables (dialects, weight, and stress pattern) were sum-coded and the continuous variable (metronome period) was centred. The intercept will have a meaningful interpretation by sum-coding the categorical variables and centring the continuous variable: the intercept will represent the mean in phase ratio of all predictors. The slopes will represent the amount of change from the grand mean, i.e., the mean of all predictors, which is represented in the intercept.

We conducted likelihood ratio tests to test the significance of each of our predictors as well as the interaction terms, comparing the full model to the nested models. The nested models contain all factors except for the one tested. If the difference between the full model and the nested model is significant, this means that dropping that specific variable would decrease the likelihood of the model, which means that it has a significant effect on the dependent variable.

The linear mixed-effects model and the likelihood ratio tests were conducted using the *afex* package (Singmann et al., 2016) in R software (R Core Team, 2020). Pairwise comparisons, through by-subject two-tailed t-test, with Bonferroni correction for multiple comparisons, were conducted using R package *phia* (Rosario-Martinez et al., 2015)

2.10 Results

2.10.1 External phase

The estimated value of the intercept, which represents the mean value of all predictors, is 0.443. This is shown in the density plot of the external phase ratios in Figure 2.2 below. This value is close to 0.5 external phase ratio, reflecting a 1/2 rhythmic mode in the phrase repetition cycle. The 1/2 rhythmic mode reflects a structure of four beats in phrases made of three stresses, with the fourth beat being a silent one.

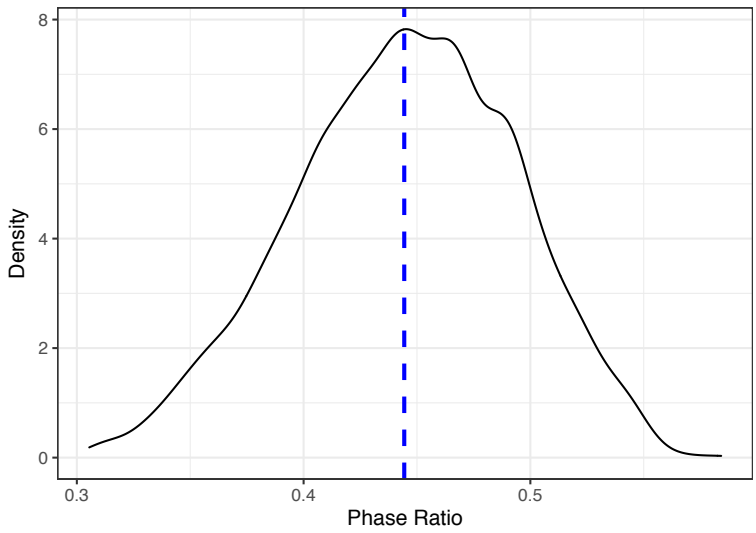


Figure 2.2: Kernel density estimation, y-axis, of phase ratios, x-axis. The dashed line represents the value of the intercept, which is 0.443. This value is close to 0.5 phase ratio, reflecting a 1/2 rhythmic mode in the phrase repetition cycle.

There was no effect of dialect, $X^2(1) = 1.15$, $p = .2$. Figure 2.3 shows the phase ratio of Hadari and Bedouin dialects.

There was no effect of syllable weight on external phase ratios, $X^2(1) = 1.56$, $p = .2$. Figure 2.4 plots the phase ratios of heavy and light syllables.

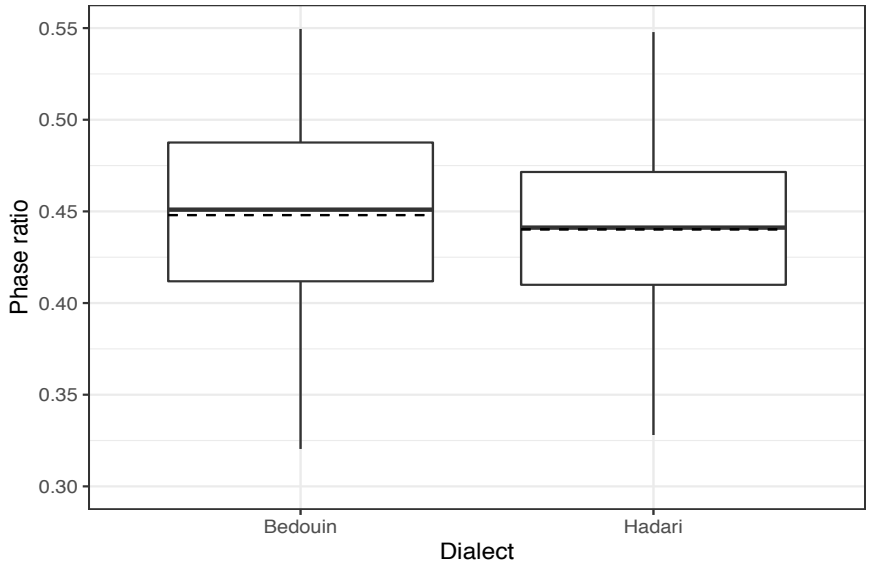


Figure 2.3: External phase ratio of Bedouin and Hadari dialects. The model's means are in dashed lines, and the thick lines represent the medians. The model mean for Bedouin is 0.449, and for Hadari is 0.440.

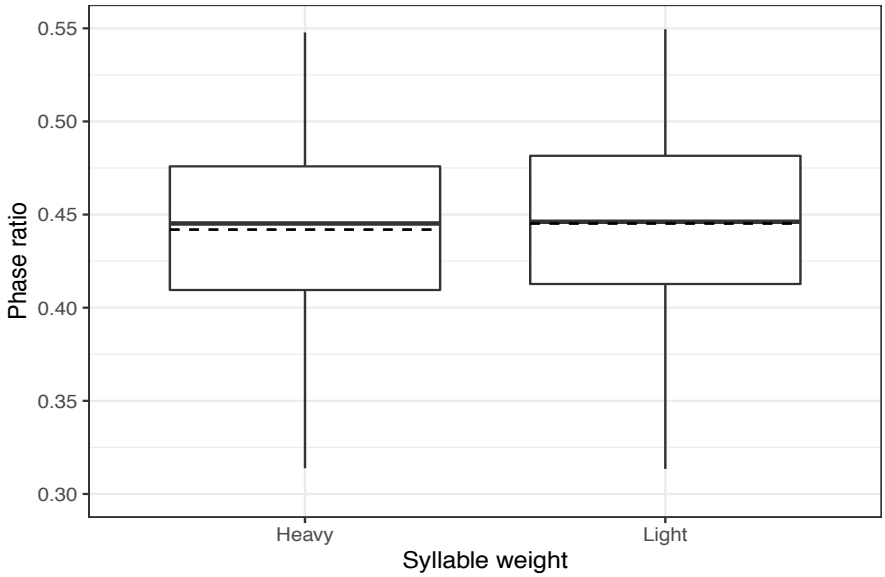


Figure 2.4: Phase ratio of heavy and light syllables. The predicted mean of heavy syllables is 0.442, and for light syllables is 0.446.

The effect of stress pattern on external phase ratio was significant, $X^2(1) = 5.19$, $p = .02$, with $\beta = -0.009$ and SE = 0.004.

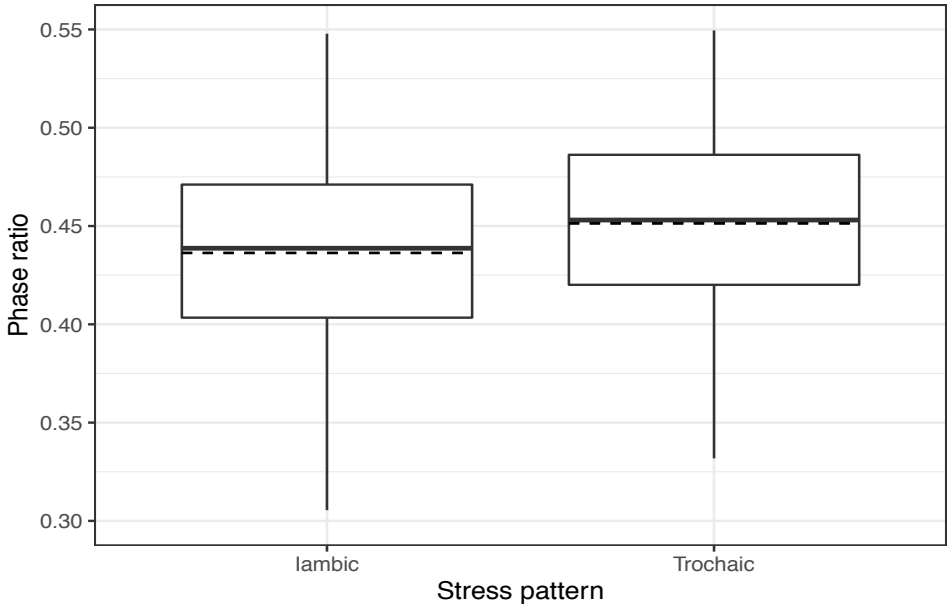


Figure 2.5: Phase ratio of the iambic and trochaic stress patterns.

The slope represents the change in phase ratio from the trochaic pattern, reference level, to the iambic pattern around the intercept, 0.443. Thus, the model's prediction of the external phase ratio in the iambic pattern is obtained by adding β to the intercept, 0.443 + (-0.009) = **0.434**. To obtain the model's prediction of external phase ratio of the trochaic pattern we reverse the sign of β, 0.443 + (+0.009) = **0.452**. This effect of stress pattern is demonstrated in Figure 2.5, where the phase ratio in the trochaic pattern is closer to the harmonic phase 0.5 than in the iambic pattern.

The effect of metronome period on external phase ratios was significant, $X^2(1) = 69.81$, $p < .001$, with $\beta = 0.038$ and SE = 0.002. The slope, β, represents the change in external phase ratio around the intercept, 0.443, at the shortest metronome period, 1270 ms, which was assigned a value of 1.01 after treating metronome periods as continuous variables. Thus, prediction for the shortest metronome period is 0.443 + 0.038 = **0.481**. For the longest and medium metronome periods, β needs to be multiplied by the continuous value of each period. Thus, prediction for the longest metronome period is 0.443 + 0.038 * -0.98 = **0.405**, and for the medium period, 0.443 + 0.0038 * 0.015 = **0.443**. Figure 2.6 demonstrates the effect of

metronome period, where the shortest metronome period is closer to the harmonic 0.5 phase than the longer periods.

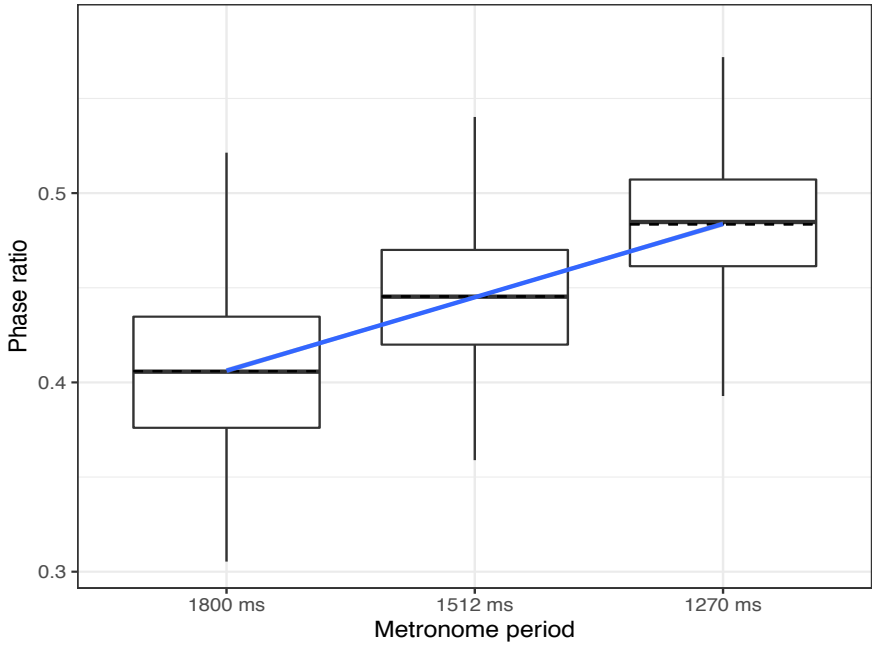


Figure 2.6: Metronome period effect on external phase ratio. The longest metronome period was 1800 ms, the medium period was 1512 ms, and the shortest period was 1270 ms. The dashed lines are the model's predicted means, the thick lines are medians, and the blue line is the model's fitted regression line.

Thus far, only two factors have had significant effects on the external phase ratio, stress pattern and metronome period. Trochaic stress pattern and the shortest metronome period encouraged closer alignment to harmonic external phase ratio of 0.5 than the iambic pattern and the longest and the medium metronome periods. We shall now explore whether interactions between the predictors affect the production of a harmonic phase relation between vowel onsets of stressed syllables and the phrase repetition cycle.

There was a significant two-way interaction between dialect and syllable weight, $X^2(1) = 5.21$, $p < .02$. We used the function *predict* from R package *stats* (R Core Team, 2019) to obtain predicted values of the interaction levels. Figure 2.7 plots the interaction with the predicted external phase ratio values. There are larger differences between heavy and light syllables in Bedouin (0.442 vs. 0.453) than in Hadari (0.437 vs. 0.442). The smaller differences in phase ratio between heavy and light syllables in Hadari may reflect a greater

degree of unstressed syllables compression, which was hypothesized to lead to similar phase alignment between heavy and light syllables.

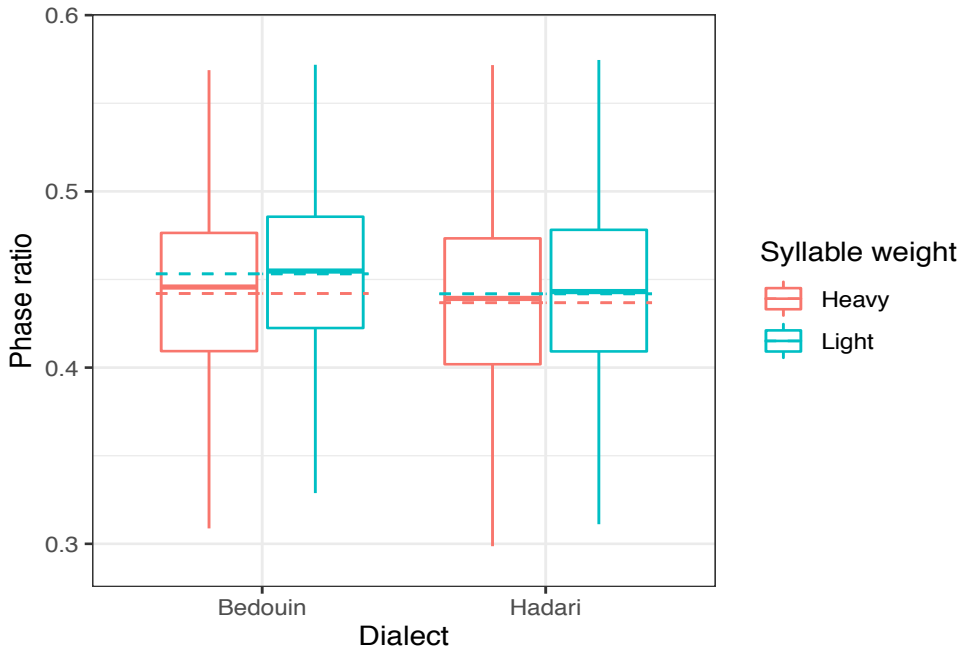


Figure 2.7: The effect of two-way interaction between dialect and syllable weight on external phase.

In pairwise comparison, we compared the difference in phase ratio between heavy and light syllables in each dialect. There were no differences between heavy and light syllables in Bedouin $p > .05$, and Hadari $p > .05$. In another test, we tested whether the difference in external phase between heavy and light syllables *differs between* the pairs of dialect groups.[2] The statistical test showed that the difference in phase ratio between heavy and light was significantly larger in Bedouin than in Hadari, $p < .0001$. Thus, Bedouin tends to show greater contrast in external phase between heavy and light syllables than Hadari.

There were no two-way interactions between dialect and stress pattern, $X^2(1) = 0.43$, $p = .4$, or between dialect and metronome period, $X^2(1) = 1.36$, $p = .2$. Also, there were no three-way

[2] The first pairwise comparison tested for *simple effects* of the interaction, in which the contrast between the levels of one factor is evaluated, when the other factor is held constant at a certain level. In our case, we compared the difference in external phase between heavy and light when dialect factor is held constant at Hadari or Bedouin. The second pairwise comparison tested the *differences of differences*, in which the contrast across the levels of one factor is evaluated between the levels of the other factor. In our case, the contrast between heavy and light is evaluated between Hadari and Bedouin. These two tests can capture different effects of the interaction, as we found in examining the interaction between dialect and weight.

interactions between dialect, weight and stress pattern, $X^2(2) = 3.37$, $p = .1$, or between dialect, weight and metronome period, $X^2(2) = 0.09$, $p = .9$. There was, however, a significant four-way interaction between all predictors, dialect, weight, stress pattern, and metronome period, $X^2(1) = 12.9$, $p < .001$. The four-way interaction is illustrated in Figure 2.8 with the model's predicted phase ratio values.

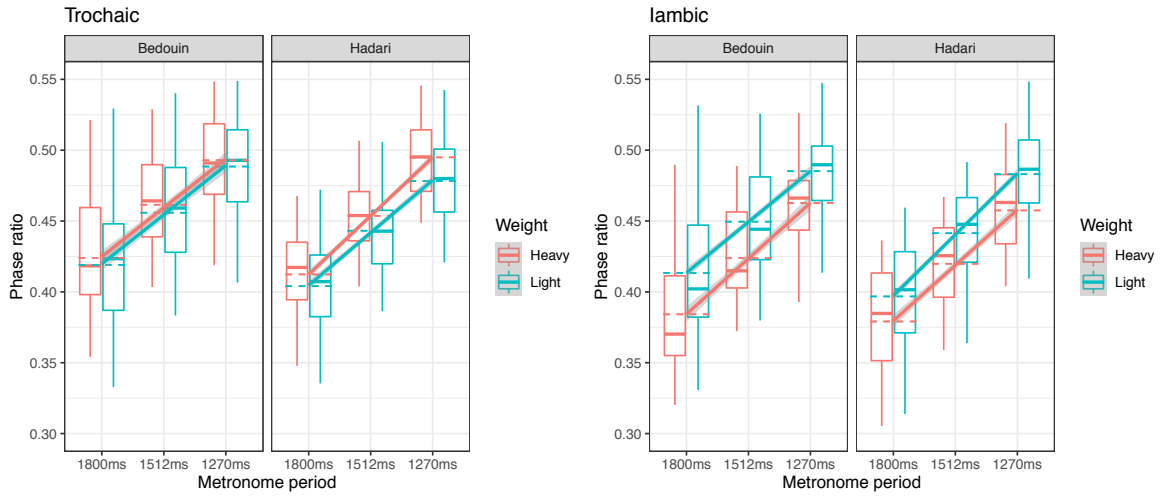


Figure 2.8: The effect of four-way interaction between dialect, weight, stress pattern, and metronome rate on external phase.

A noticeable between-dialects difference arises in the trochaic pattern, in the shortest metronome period, where syllables' alignment is closest to harmonic external phase ratio of 0.5. Bedouin has a similar external phase ratio for heavy and light syllables, 0.494 vs. 0.493, whereas light syllables have an earlier external phase ratio than heavy syllables in Hadari, at 0.481 and 0.497, respectively.

Post-hoc multiple comparisons revealed a significant difference in phase ratio between heavy and light syllables in Hadari, $p < .001$, but not in Bedouin, $p > .05$.

Variable phase alignment may reflect different syllabic durational patterns between the two dialects, which themselves emerge from differences in temporal stress contrast between Hadari and Bedouin. We will discuss in more detail below how variable syllable duration may afford different phase alignment between dialects.

As for other interactions, there were no two-way interactions between syllable weight and stress pattern, $X^2(1) = 3.44$, $p = .06$ or between stress pattern and metronome period, $X^2(1) =$

0.63, p = .4, and there were no three-way interactions between weight, stress pattern and metronome period, $X^2(1) = 0.46$, p = .7, or between dialect, stress pattern and rate, $X^2(1) = 0.06$, p = .8.

2.10.1.1 Summary and discussion

There are three key points arising from these analyses of the external phase ratio results. First is the lower external phase ratio, i.e., earlier phase alignment, in the iambic pattern than in the trochaic pattern. Initially, we predicted that because the first stressed syllables in the iambic pattern are separated from the metronome beep by an unstressed syllable, there would be a shift towards later phases, resulting in higher phase ratios in the iambic pattern than in the trochaic pattern. However, our results show the opposite. Possibly, this has to do with our text materials rather than prosodic timing constraints. The unstressed syllables in the iambic pattern are all of a CV structure, while all unstressed syllables in the trochaic pattern are of a CVC structure. This would lead to shorter durations of the iambic phrases than the trochaic phrases. As a consequence, final stressed syllables in the iambic pattern occur earlier in the cycle, with lower phase ratios than in the trochaic pattern. This is exemplified in Figure 2.9.

Figure 2.9: Illustration of phase alignment in iambic, left, and trochaic sentences, where simpler unstressed syllables in the former lead to earlier phase alignment in the cycle, as indicated by the blue arrows.

The second point relates to the closer phase alignment to a harmonic 0.5 external phase ratio at the shortest metronome period. As the shortest metronome period encourages speakers to speak more rapidly, it may be the case that there is a preferred speaking rate for temporal coordination with a harmonic 0.5 phase ratio.

Findings from Tajima's (1999) speech cycling study also suggest that speaking rate mediates temporal coordination, as stable rhythmic modes varied at various metronome rates. Figure 2.10 demonstrates the change in stable rhythmic modes in Japanese, from 1/2, in the slower rates, to 2/3, in the faster rates. In our case, however, there are no distinct rhythmic modes;

the fast rate simply leads to closer alignment with the harmonic 0.5 phase. The lack of distinct rhythmic modes at different rates in our data is probably due to the fewer metronome periods in our experiment, compared to Tajima's (1999) experiment, which involved more than ten rates.

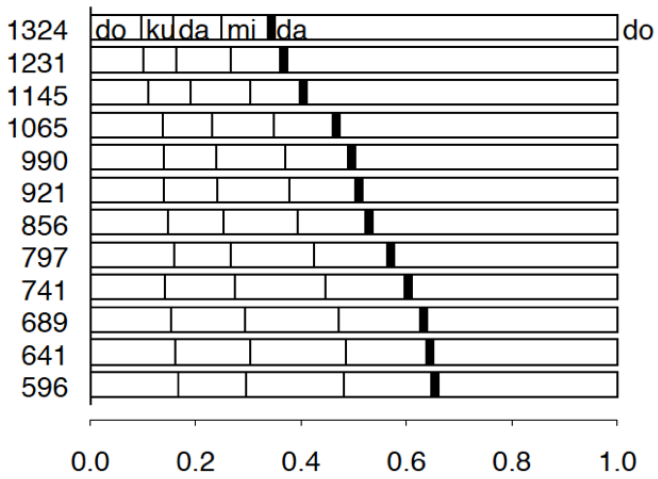


Figure 2.10: External phase ratios in Japanese as a function of metronome period. At longer metronome periods, 1065, the rhythmic mode is 1/2, 0.5 phase ratio, while at shorter periods, 596 ms, the rhythmic mode changes to 2/3, 0.666 phase ratio. Source: Tajima (1999, 49).

The third point relates to interactions. There was a two-way interaction between dialect and weight, with Hadari exhibiting smaller differences in phase ratio between heavy and light syllables than Bedouin. However, in the higher-order interaction between dialect, weight, stress pattern, and metronome period there was a different picture. In the shortest metronome period at the trochaic pattern, Hadari showed greater differences between heavy and light syllables than Bedouin, as Hadari aligned light syllables earlier than heavy syllables. Two processes may explain the contrasting alignment patterns of heavy and light syllables in Hadari and Bedouin dialects in the trochaic pattern at the shortest metronome period. The first is that heavy syllables, by virtue of their phonological length, have greater contrast with weak syllables, thus prompt stronger coordination, i.e., closer alignment, with the harmonic external phase, 0.5. This potential top-down effect seems to be present in both dialects, as both exhibit close alignment of heavy syllables to the harmonic external phase, 0.5. The second factor is syllabic compressibility, which may explain differences between dialects in light syllables' external phase. In particular, earlier external phase ratio of final stressed syllables in Hadari may be due to greater compressibility in non-final syllables, especially unstressed syllables. Figure 2.11 provides an example of the earlier external phase of light syllables in Hadari.

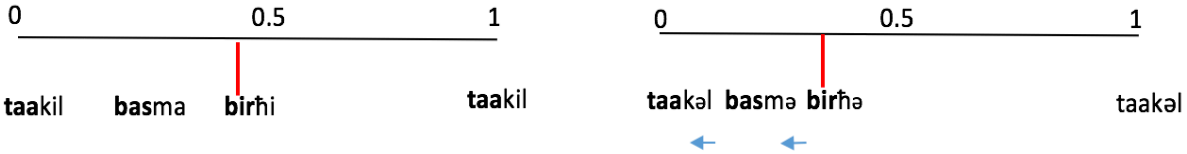


Figure 2.11: Schematic of syllable compression effect on external phase alignment. In the right panel, vowel reduction leads to earlier external phase alignment in the cycle, indicated by the blue arrows.

Thus, syllabic compressibility does not seem to afford closer alignment to the harmonic external phase, 0.5, in Hadari in the shortest metronome period in the trochaic pattern. However, because speaking rate mediates temporal coordination, we may speculate that affordance of syllabic compressibility to stronger temporal coordination may emerge at metronome rates faster than those used in our experiment.

2.10.2 Internal phase results

The Internal phase relates to the temporal ratio between vowel onsets of medial stressed syllables and the interval between initial and final stresses within the phrase. The estimate of the intercept value is 0.503. This value indicates that vowel onsets of second stressed syllables tend to be in the middle of the phrase and that there is a harmonic phase angle of 1/2 within the phrase. The intercept value is represented in the density plot of the phase ratios in Figure 2.12.

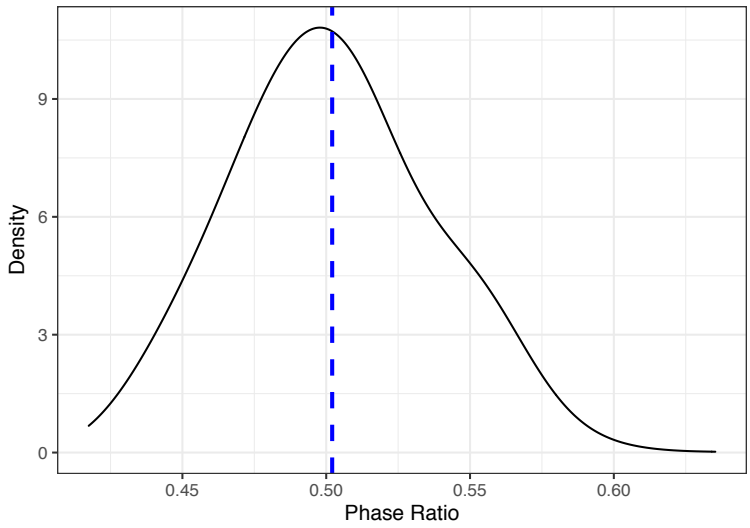


Figure 2.12: Kernel density estimation, y-axis, of internal phase ratios, x-axis. The dashed line represents the value of the intercept, which is 0.503.

There was no significant effect for dialect, $X^2(1) = 0.02$, $p = .8$. The internal phase ratio of both dialects is shown in Figure 2.13.

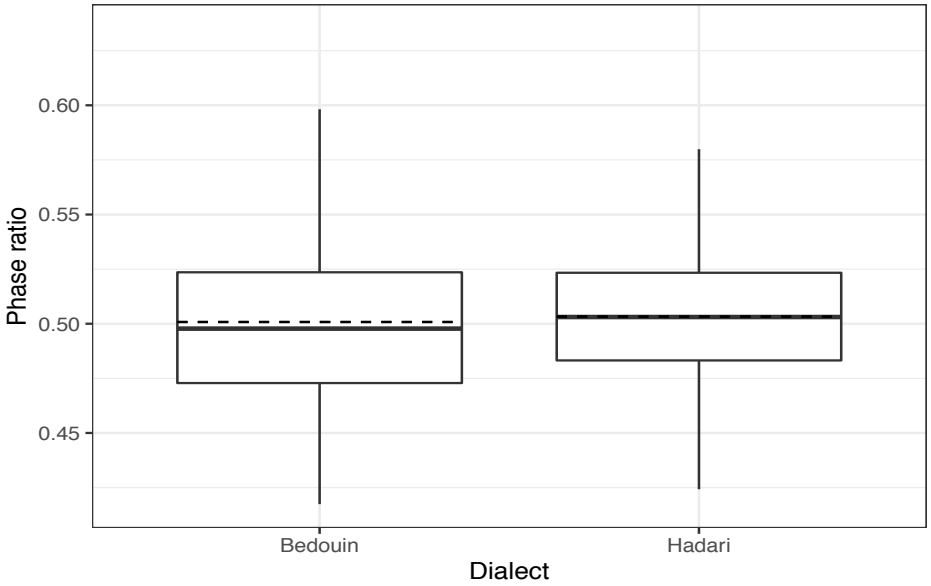


Figure 2.13: Internal phase ratio of Bedouin and Hadari.

There was no effect of syllable weight on internal phase ratio, $X^2(1) = 2.13$, $p = .1$. Figure 2.14 shows the internal phase ratio for heavy and light syllables.

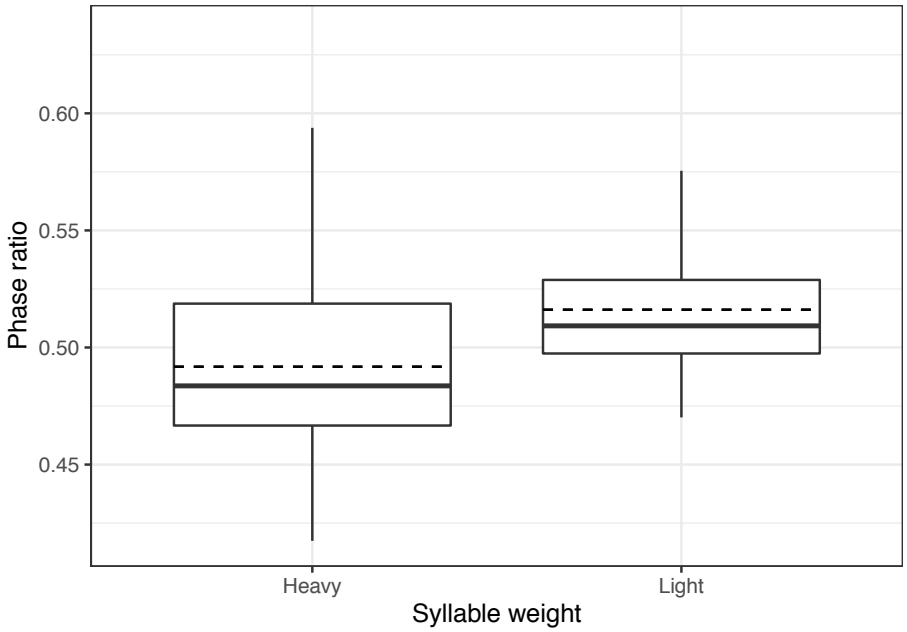


Figure 2.14: Internal phase ratio of heavy and light syllables.

There was a significant effect for stress pattern on the internal phase ratio, $X^2(1) = 5.09$, $p = .02$. with β = -0.02, and SE = 0.006. To obtain predicted phase ratio of the iambic pattern we add β to the intercept, 0.503 + (-0.016) = **0.487**, and we reverse β's sign for the trochaic pattern, 0.503 + 0.016 = **0.519**. Figure 2.15 shows phase alignment in both stress patterns.

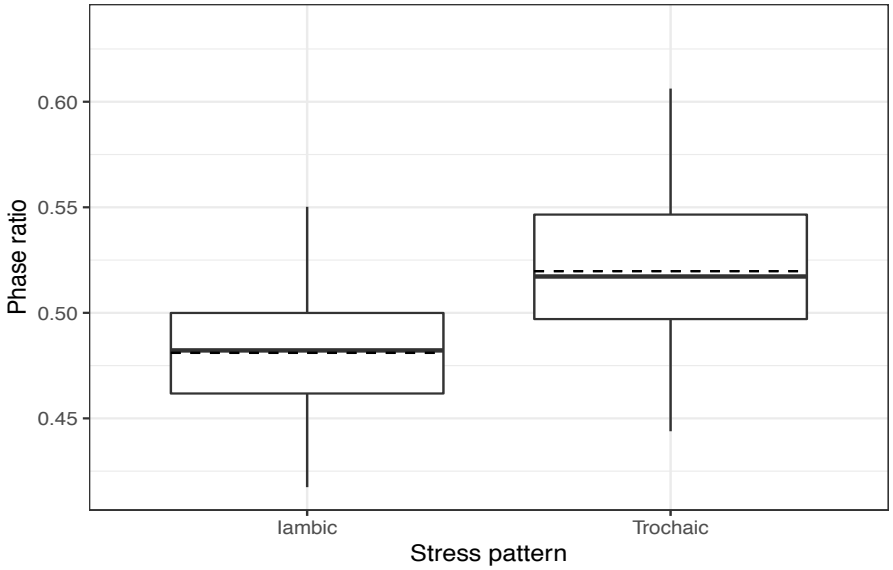


Figure 2.15: Internal phase ratio of iambic and trochaic stress patterns.

Figure 2.16 shows the effect of the metronome period on the internal phase ratio.

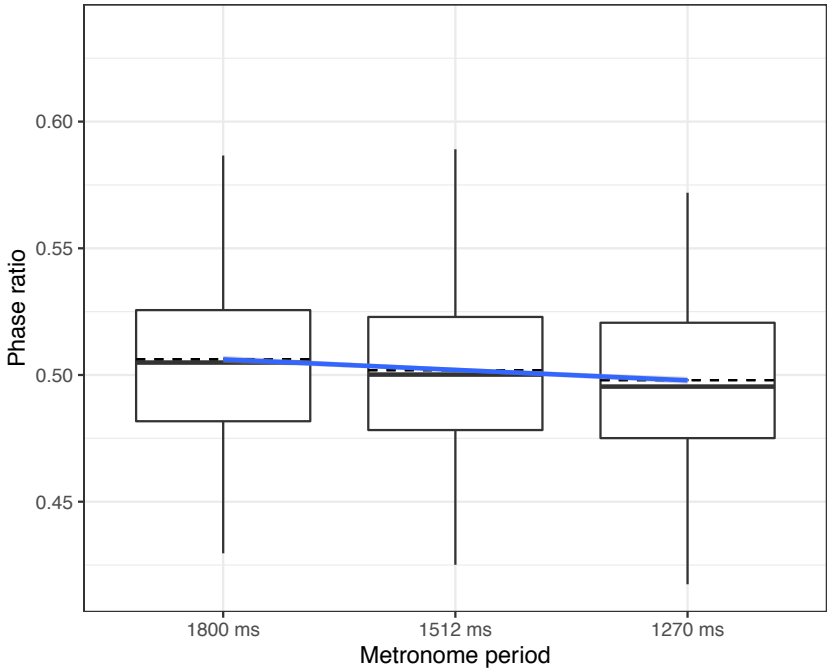


Figure 2.16: Metronome period effect on internal phase ratio.

There was a significant effect of metronome period on internal phase ratio, $X^2(1) = 17.79$, $p = 0.001$, with $\beta = -0.0047$, and SE = 0.0009. The slope, β, represents the change in phase ratio at the shortest metronome period, 1270 ms, around the intercept, 0.503 + (-0.0047) = **0.498**. For the longest period the predicted phase ratio is: 0.503 + (-0.0047) *-0.98 = **0.507**, and for the medium period: 0.503 + (-0.0047) *0.015 = **0.503**. The difference in phase alignment between the medium and the shortest metronome periods is very small, and there is only a tendency for phase alignment to be later in the medium period than in the shorter period. The earlier phase alignment in the shortest period could be due to the shorter sentence duration in the shortest metronome period.

Only two factors influenced internal phase ratio: stress pattern and metronome period. Regarding the effect of stress pattern, there was a tendency for phase ratio in the iambic pattern to be earlier than in the trochaic pattern. It is possible that this pattern is due to the text materials, rather than prosodic timing constraints, as unstressed syllables in the iambic pattern are of simpler structure, CV, than in the trochaic pattern, CVC, which may lead to earlier phase alignment in the former. As for metronome period effects, the medium and the shortest periods were closer to a harmonic internal phase ratio of 0.5, than the longest period.

For the internal phase ratio, there was a significant two-way interaction between dialect and syllable weight, $X^2(1) = 5.64$, $p = .01$. Figure 2.17 plots the predicted means of the interaction levels. Both dialects tended to align heavy syllables closer to the harmonic internal phase of 0.5 than light syllables, reflecting that heavy syllables prompt stronger coordination with the harmonic 0.5 phase. Hadari, however, tended to align light syllables more similarly to heavy syllables and closer to 0.5 phase than Bedouin. Pairwise comparison showed no difference in the internal phase between heavy and light syllables, 0.498 vs. 0.510, in Hadari, $p > .05$, while there were significant differences in the internal phase between heavy and light syllables, 0.489 vs. 0.517, in Bedouin, $p < .05$. We will discuss in more details why Hadari affords closer alignment of light syllables to the harmonic 0.5 phase than Bedouin.

There were no two-way interactions between dialect and stress pattern, $X^2(1) = 0.67$, $p = .4$, or between dialect and metronome period, $X^2(1) = 0.87$, $p = .3$.

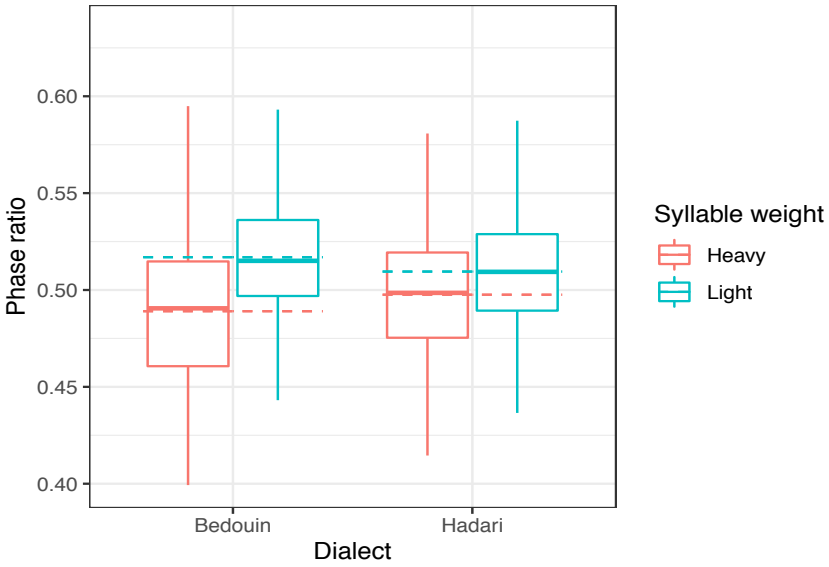


Figure 2.17: The effect of two-way interaction between dialect and weight on the internal phase.

There were no three-way interactions between dialect, weight and metronome period, $X^2(2) = 3.36$, $p = .06$, or between dialect, weight and stress pattern, $X^2(2) = 2.55$, $p = .1$. There was no four-way interaction between all predictors, $X^2(1) = 2.3$, $p = .6$. There was a significant two-way interaction between syllable weight and metronome period, $X^2(1) = 5.23$, $p = .02$. Figure 2.18 plots the interaction.

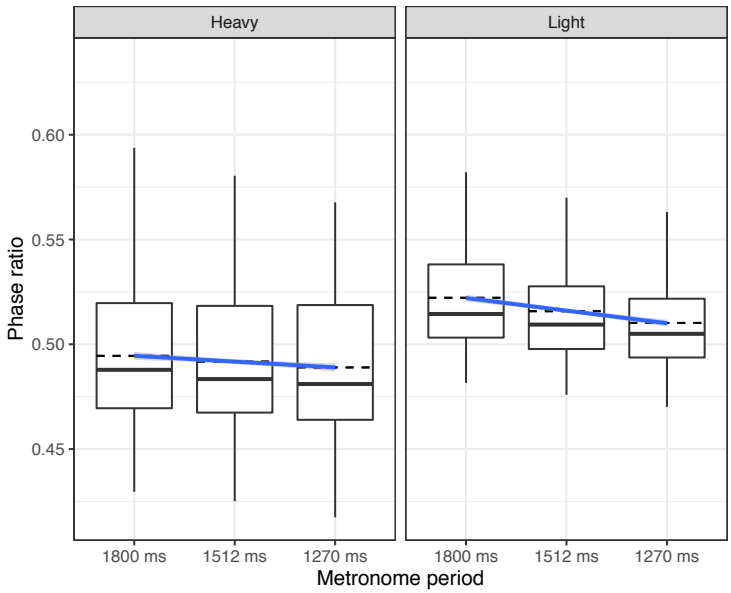


Figure 2.18: The effect of two-way interaction between weight and metronome period on the internal phase.

Figure 2.18 shows more consistent internal phase alignment, relative to the harmonic 0.5, in heavy syllables than in light syllables across the longest, medium, and shortest metronome periods. Post-hoc comparisons showed no difference in the internal phase ratio between heavy syllables in the longest period and in the medium period, 0.497 vs 0.493, $p > .05$, and there is no difference in internal phase ratio between heavy syllables in the medium period and the shortest period, 0.493 vs. 0.491, $p > .05$. As for light syllables, there was a significant difference in internal phase ratio between light syllables in the longest metronome period and the medium period, 0.520 vs. 0.513, $p < .05$, and a significant difference in the internal phase ratio between light syllables in the medium period and in the shortest period, 0.513 vs 0.508, $p < .05$.

More consistent internal phase alignment relative to the harmonic 0.5 phase in heavy syllables across different metronome periods indicates that heavy syllables prompt stronger coordination with the harmonic internal phase of 0.5, while coordination with harmonic phase in light syllables seems to be facilitated in shorter metronome periods.

There was no two-way interaction between weight and stress pattern, $X^2(1) = 0.81$, $p = .3$, and there was no three-way interaction between stress pattern, syllable weight and metronome period, $X^2(1) = 0.96$, $p = .4$.

2.10.2.1 Summary and discussion

There are four primary findings in these analyses of internal phase ratios in speech cycling tasks with speakers of Hadari and Bedouin dialects of Kuwaiti Arabic. First, the internal phase ratio in the iambic pattern was earlier than in the trochaic pattern. This is consistent with the findings of earlier external phase in the iambic pattern than in the trochaic patterns. Similar to our explanation of the earlier external phase in the iambic pattern, the earlier internal phase in the iambic pattern is plausibly due to the simpler structure of unstressed syllables in the iambic pattern, CV, than unstressed syllables in the trochaic pattern, CVC. This is illustrated in the schematic demonstration in Figure 2.19. Thus, the effect of stress pattern on internal phase is due to the text materials rather than prosodic timing constraints.

Figure 2.19: Schematic illustration of the possible effect of simpler unstressed syllables, CV, in the iambic pattern, left panel, on earlier internal phase than in the trochaic pattern, as indicated by the blue arrows.

Second, medial stressed syllables in the medium and in the shortest metronome periods were closer to a harmonic phase ratio of 0.5 than in the longest period. This is in line with the finding in the external phase measure, that shorter metronome periods encourage a stronger rhythmic pattern than the longer periods. As shorter metronome periods encourage speakers to speak more rapidly, this indicates that speaking rate mediates temporal coordination with the harmonic internal phase of 0.5.

Third, we found that both dialects tended to align heavy syllables closer to the harmonic phase 0.5, whereas, in light syllables, Bedouin showed later alignment from 0.5 than Hadari. The affordance of heavy syllables to closer alignment to the harmonic 0.5 phase in both dialects is possibly because of heavy syllables phonological length, which leads to stronger contrast with unstressed syllables. The closer alignment of light syllables to the harmonic 0.5 phase in Hadari than in Bedouin is plausibly due to greater unstressed syllable reduction in Hadari. The potential effect of unstressed syllable reduction on earlier phase alignment, and closer to the harmonic 0.5 phase, in Hadari, is exemplified in Figure 2.20.

Figure 2.20: Schematic illustration of the potential effect of unstressed syllables reduction in Hadari, right, on earlier phase alignment in Hadari, as indicated by the blue arrows.

Fourth, the two-way interaction between weight and metronome period supports the idea that heavy syllables afford closer alignment with the harmonic phase 0.5. Heavy syllables showed consistent close alignment to the harmonic 0.5 phase across the different metronome periods.

On the other hand, light syllables showed greater differences in the internal phase across the different metronome periods, as closer alignment to 0.5 phase was more affordable at shorter metronome periods.

2.11 General discussion

In the external phase analysis, the intercept value, which represents the mean phase ratio of all predictors was close to 0.5, reflecting a 1/2 rhythmic mode in the phrase repetition cycle. The 1/2 rhythmic mode reflects a structure of four beats in phrases made of three stresses, with the fourth beat being a silent one. In the internal phase, the intercept value was close to 0.5, which indicates that medial stresses tend to be in the middle of the phrase.

As for dialectal differences, our general hypothesis was that as Hadari tends to have greater contrast between stressed and unstressed syllables than Bedouin, there would be different patterns of syllables' alignment with harmonic phase angles between the two dialects. There were dialectal differences in the external phase and the internal phase ratios. In the external phase, there was a four-way interaction between dialect, syllable weight, stress pattern, and metronome period. In the trochaic pattern in the shortest metronome period, Bedouin tended to align heavy and light syllables similarly, close to the harmonic external phase 0.5, while in Hadari, heavy syllables were aligned close to 0.5 while light syllables were aligned earlier. We interpreted the close alignment of heavy syllables to a harmonic external phase of 0.5 in both dialects as reflecting a top-down effect; the greater contrast of heavy syllables with unstressed syllables, due to their phonological length, attracts the alignment with a harmonic phase angle. The contrasting patterns of light syllables' alignment, closer to a harmonic phase of 0.5 in Bedouin and earlier in Hadari, may reflect different degrees of unstressed syllables compressibility. Hadari tends to compress unstressed syllables to a greater degree than Bedouin (see section 1.6.1 for a review), which may lead to earlier alignment of light syllables in Hadari in the shortest metronome period in the trochaic pattern. For its relevance, we reproduce Figure 2.11, which showed a schematic of the potential effect of unstressed syllables compressibility on the alignment of light syllables in the two dialects. Thus, unstressed syllables compressibility in Hadari does not afford closer alignment of light to a harmonic phase in the trochaic pattern in the shortest metronome period.

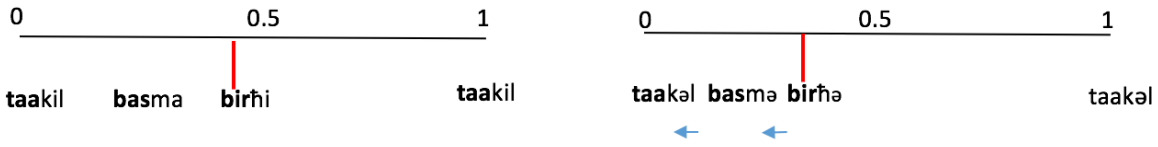


Figure 2.21: Schematic of unstressed syllables' compressibility effect on external phase alignment. In the right panel, vowel reduction in Hadari leads to earlier external phase alignment in the cycle, indicated by the blue arrows.

In the internal phase measure, there was a two-way interaction between dialect and syllable weight. Both dialects tended to align heavy syllables closer to a harmonic internal phase of 0.5 than light syllables, reflecting that heavy syllables attract harmonic phase alignment due to their greater contrast with unstressed syllables. Hadari, however, tended to align light earlier in the phrase, and closer to the harmonic 0.5 phase than Bedouin. It is plausible that closer alignment of light syllables to 0.5 in Hadari is due to greater unstressed syllables compressibility. We reproduce Figure 2.20 which illustrated a schematic of the potential effect of unstressed syllables compressibility on light syllables' internal phase in the two dialects.

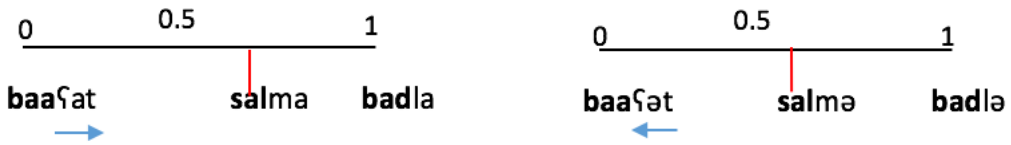


Figure 2.22: Schematic illustration of the potential effect of unstressed syllables reduction in Hadari, right, on earlier phase alignment in Hadari, as indicated by the blue arrows.

The effect of unstressed syllable compressibility in Bedouin and Hadari may seem different between the external phase and the internal phase measures, however, this is due to the higher-order, four-way interaction which influenced the external phase.

As our explanation of the between-dialects differences refers largely to unstressed syllables compressibility, we will investigate in Experiment 1 (b) the potential varying degrees of unstressed syllables compression between dialects, which might influence phase alignment.

One of the findings that may require further discussion is the effect of metronome period on phase alignment. We found that external phase and internal phase ratios were closer to a

harmonic phase angle of 0.5 at shorter metronome periods. As shorter metronome periods encourage speakers to speak more rapidly, this indicates that speaking rate mediates temporal coordination with harmonic phase angle. Cummins and Port (1998) addressed the potential effect of speaking rate on temporal coordination in targeted speech cycling, in which speakers were required to align the first stressed syllable to a high tone (H) metronome and the second stressed syllable, i.e., final stressed, with a low tone (L) metronome. In the slow rate condition, the H-L interval was fixed at 700 ms, and H-H cycle ranged from 1 s to 2.3 s, while in the fast rate condition, the H-L interval was fixed at 450 ms, and the H-H cycle ranged from 638 ms to 1500 ms. The phase targets for the second stressed syllable evenly varied between 0.3 and 0.7. Three speakers showed no effect of rate, as they were able to produce 1/3, 1/2, and 2/3 phase angles at the slow and the fast rate. One speaker, however, produced 1/3, 1/2, and 2/3 at the slow rate only, but could not produce 2/3 at the fast rate. Cummins and Port concluded that for this speaker, phase alignment may not reflect hierarchical temporal coordination between vowel onsets and the phrase repetition cycle. The alignment of the second stress in the fast rate at 1/3 and 1/2 targets only, but not 2/3, may reflect a tendency to align vowel onsets' intervals at a certain preferred durational window, such as 450 ms or less, which may be more available at earlier phase targets, 1/3 and 1/2, than in later phase targets, 2/3.

According to the latter assertion it may be argued that, in our speech cycling experiment, as phase ratio increased from slow to fast metronome rates, phase alignment may only reflect the distribution of vowel onsets at preferred durational windows rather than hierarchical organization. However, we believe that preference of phase alignment at a certain speaking rate and hierarchical organization need not be mutually exclusive. In the external phase, dialectal difference in phase alignment of heavy and light syllables emerged in the trochaic pattern in the fast metronome rate, where syllables' alignment was closer to a harmonic phase of 0.5. This indicates that the fast rate attracts hierarchical temporal organization, as reflected in the variable temporal organization of heavy and light syllables between the two dialects. If rate effects were confined to constraining the distribution of stresses at a certain durational window, dialectal differences would not have been observed at the fast rate only.

Chapter 3. Experiment 1 (b): syllabic durational profile

3.1 Introduction

There are several timing patterns in the external and internal phase measures that require explanation in terms of the syllabic durational patterns that would result in the alignment of stressed syllables at certain phases. First, we saw that the iambic pattern was associated with lower phase ratios than the trochaic pattern in the external and internal phase measures. It was suggested that this might be due to the shorter unstressed syllables in the iambic pattern which were of a CV structure, than unstressed syllables in the trochaic pattern which were of a CVC structure. Therefore, syllabic durations in the iambic and trochaic patterns will be investigated.

Second, the shortest metronome period was associated with closer alignment to the harmonic phase, 0.5, than the longest and the medial periods. As discussed earlier, the shortest period is associated with faster speaking rate, which may provide a preferable rate for temporal coordination with 0.5 phase. As faster rates are associated with shorter syllabic durations, we predict syllables' duration to be shorter at faster rates which are associated with shorter metronome periods.

Third, in relation to dialectal differences, we found in the external phase measure that Bedouin aligned heavy and light syllables similarly, close to 0.5 phase, in the shortest metronome period in the trochaic pattern, while Hadari aligned light syllables earlier than heavy syllables, in a four-way interaction. In the internal phase measure, Hadari showed smaller differences in the alignment of heavy and light syllables than Bedouin in a two-way interaction. We speculated that these dialectal differences might arise due to greater unstressed syllables reduction in Hadari than in Bedouin. Also, unstressed syllables reduction between dialects might differ based on the position in the phrase. Since the initial and final positions mark phrase boundaries, they may be more prominent than the medial position. Thus, syllabic duration, including the degree of unstressed syllables reduction, may vary based on the prominence level of phrasal position.

3.2 Methods

From the sound files that included multiple repetitions of a single phrase, each phrase repetition was extracted into a separate sound file for syllable duration analysis. Boundaries to the speech segments were added, in Praat, based on the segmentation criteria described in section (2.6). Then, in another tier, boundaries were added to the phonological syllables. Syllable duration was extracted using a Praat script. The total number of tokens analysed was *37183*. Tables 3.1 and 3.2 show detailed analysed tokens in Bedouin and Hadari, respectively[3]. Cells arrangement in the tables is based on the highest level of interaction in the linear mixed-effects model (dialect*rate*position*syllable stress); see section 3.3 below. The statistical model also included a two-way interaction between stress pattern and syllable stress, but was not nested in a higher level of interaction (explanation in the text below); thus we provided details for the number of tokens of syllables (heavy, light, unstressed) in both stress patterns, iambic (im) and trochaic (tr).

*Table 3. 1: Analysed tokens of syllable duration analysis in Bedouin. The total is 20238. Cells arrangement is based on the highest interaction level in the linear mixed-effects model (dialects*rate*position*syllable stress). We also provided detailed number of tokens of syllables (heavy, light, unstressed) in the iambic (im) and trochaic (tr) stress patterns.*

Tokens per rate trial	Slow			Medium			Fast		
	6853			6810			6575		
Phrase position	Initial	Medial	Final	Initial	Medial	Final	Initial	Medial	Final
Heavy	**284**	**643**	**440**	**289**	**627**	**440**	**277**	**596**	**422**
im	*0*	*354*	*150*	*0*	*343*	*151*	*0*	*325*	*148*
tr	*284*	*289*	*290*	*289*	*284*	*289*	*277*	*271*	*274*
Light	**861**	**488**	**707**	**849**	**494**	**700**	**824**	**473**	**679**
im	*573*	*206*	*423*	*565*	*204*	*414*	*525*	*198*	*403*
tr	*288*	*282*	*284*	*284*	*290*	*286*	*272*	*275*	*276*
Unstressed	**1140**	**1143**	**1147**	**1135**	**1139**	**1137**	**1102**	**1109**	**1103**
im	*572*	*571*	*573*	*564*	*564*	*562*	*551*	*549*	*554*
tr	*568*	*572*	*574*	*571*	*575*	*572*	*551*	*550*	*549*
Total	**2285**	**2274**	**2294**	**2273**	**2260**	**2277**	**2203**	**2178**	**2204**

[3] Note that there are disparities between the total analysed tokens in Bedouin (20238) and in Hadari (16945), which are due to an oversight in not segmenting the phrases of one Hadari speaker into syllables.

Table 3. 2: Analysed tokens of syllable duration analysis in Hadari. The total is 16945.

Tokens per rate trial	Slow			Medium			Fast		
	5493			5934			5518		
Phrase position	Initial	Medial	Final	Initial	Medial	Final	Initial	Medial	Final
Heavy	**219**	**524**	**343**	**251**	**558**	**379**	**236**	**519**	**356**
im	*0*	*294*	*119*	*0*	*310*	*132*	*0*	*283*	*113*
tr	*219*	*230*	*224*	*251*	*248*	*247*	*236*	*236*	*243*
Light	**696**	**389**	**574**	**741**	**431**	**609**	**679**	**402**	**566**
im	*466*	*172*	*349*	*495*	*182*	*362*	*447*	*165*	*336*
tr	*230*	*217*	*225*	*246*	*249*	*247*	*232*	*237*	*230*
Unstressed	**915**	**917**	**916**	**984**	**988**	**993**	**920**	**920**	**920**
im	*467*	*466*	*467*	*492*	*492*	*497*	*448*	*448*	*447*
tr	*448*	*451*	*449*	*492*	*496*	*496*	*472*	*472*	*472*
Total	**1830**	**1830**	**1833**	**1976**	**1977**	**1981**	**1835**	**1841**	**1842**

3.3 Analysis

In order to quantify the patterns of syllabic duration, a linear mixed-effects model was fitted to the data. Syllable duration in milliseconds was the dependent variable. There were five predictors: dialect, stress pattern, metronome period, syllable stress, and phrasal position. There were two levels for dialect (Hadari and Bedouin) and for stress pattern (iambic and trochaic). Syllable stress factor contained three levels (unstressed, stressed light and stressed heavy), so did position factor (initial, medial and final). Metronome period was treated as a continuous variable.

Explaining dialectal differences in phase alignment in terms of syllabic duration, stressed and unstressed syllables, and taking into consideration the potential effect of metronome period and position requires multiple interaction levels. We included two-way interactions between dialect and syllable stress, dialect and metronome period, and dialect and phrasal position in the model. We did not think that an interaction between dialect and stress pattern (iambic and trochaic) is required since the effect of stress pattern did not reflect prosodic constraints on phase alignment, rather it seemed confined to differences in text materials and the simpler phonological structure (hence shorter duration) of unstressed syllables in the iambic sentences. Therefore, the interaction that is necessary for uncovering the effect of stress

pattern is a two-way interaction between stress pattern (iambic and trochaic) and syllable stress (stressed heavy, stressed light, and unstressed).

Higher order three-way and four-way interactions were also motivated. Specifically, dialects might differ in the degree of unstressed syllables reduction relative to stressed syllables at faster speaking rates, thus a three-way interaction between dialect, syllable stress, and metronome period was included in the model. We also mentioned that phrasal position is an important factor that might affect syllabic duration, and dialects might show different syllable duration and different degrees of unstressed syllables reduction based on position. Therefore, three-way interaction between dialect, position, and syllable stress was included in the model. The interaction between dialect, position, and syllable stress could be modulated by rate as unstressed syllables reduction between dialects at faster rates might be different based on the metrical strength of phrasal position. Thus, four-way interaction between dialect, position, syllable stress, and metronome period was included in the model.

We also included in the model all possible lower-level interactions that are nested in the three-way and four-way interactions as controls for potential variation. These included two-way interactions between syllable stress and position, syllable stress and metronome period, position and metronome period; three-way interactions of syllable stress, position and metronome period and between dialect, position, and metronome period.

Regarding the random structure, speaker was included as a random intercept, and metronome period was included as a random slope by speaker. The model did not converge with a more complex structure.

Likelihood ratio tests, based on model comparisons, were conducted for significance testing using package *afex* (Singmann et al., 2016) in R software. Pairwise comparison to compare the means of interactions levels, through by-subjects two-tailed t-test, was conducted using R package *phia* (Rosario-Martinez et al., 2015).

3.4 Results

The model's intercept, which represents the mean of the mean syllable duration in all predictors is 182 ms.

Figure 3.1 shows syllable duration in Bedouin and Hadari. There was no effect of dialect, $X^2(1) = 0.13$, $p = .7$.

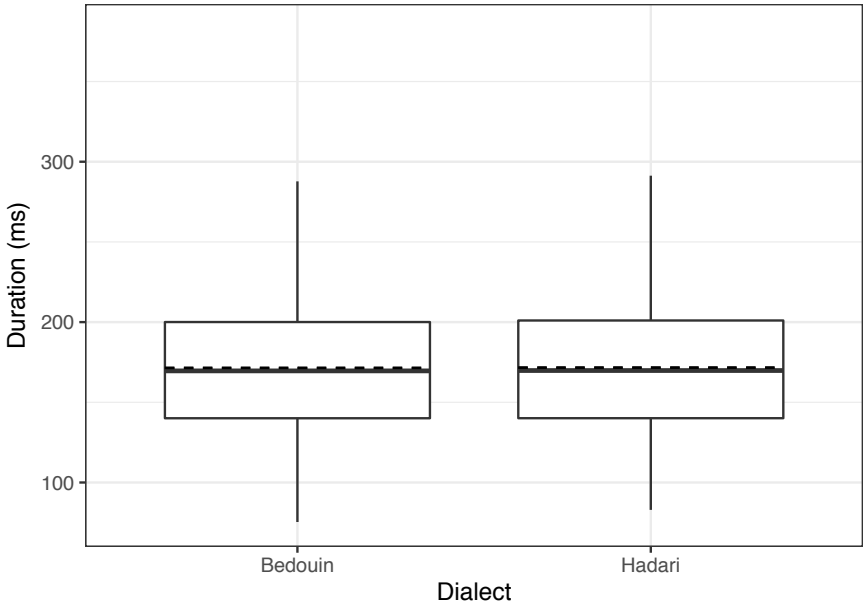


Figure 3.1: Syllable duration in Bedouin and Hadari.

Figure 3.2 shows the model's predicted means for syllable duration by syllable stress, where heavy syllables are the longest followed by light and unstressed syllables. There was a significant effect for syllable stress, $X^2(2) = 2147.62$, $p = .001$. For heavy syllables, β = 24.21 ms with SE = 1.03 ms, with the reference level being unstressed syllables, and for light syllables, β = 12.04 ms with SE = 0.87 ms relative to unstressed syllables. As β represents the change around the intercept, 182, the model's prediction for heavy syllables is: 182 + 24.21 = **206 ms**, and for light syllables: 182 + 12.04 = **194 ms**. As for unstressed syllables, the reference level, we reverse βs' signs and add them to the intercept: 182 – 24.21 – 12.04 = **145 ms**.

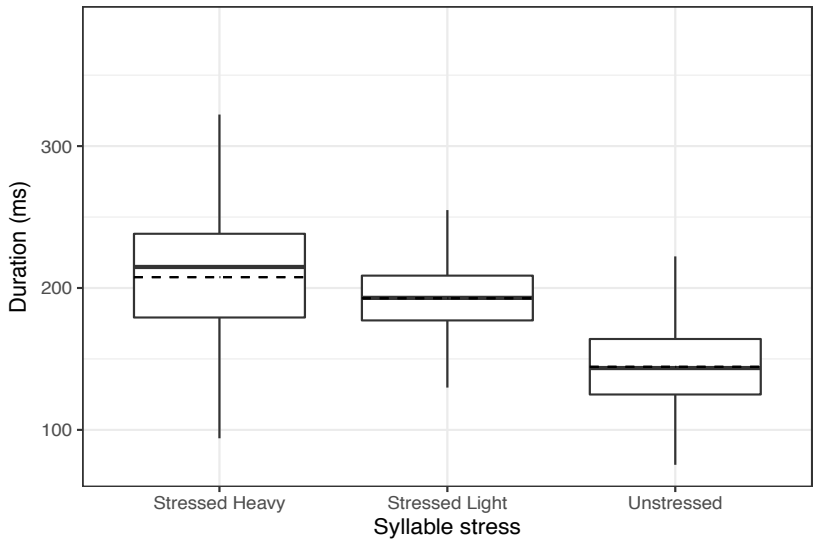


Figure 3.2: The effect of syllable stress on syllable duration.

Figure 3.3 shows the model's predicted means for syllable duration in iambic and trochaic patterns, where mean syllable duration is longer in the iambic pattern than in the trochaic. There was a significant effect for stress pattern, $X^2(2) = 4.3$, $p < .03$, with $\beta = 1.24$ ms, and SE = 0.63 ms. Prediction for mean syllable duration in the iambic pattern is 182+1.24 = **183.2 ms**, and for the trochaic stress pattern 182+(-1.24) = **180.7 ms**.

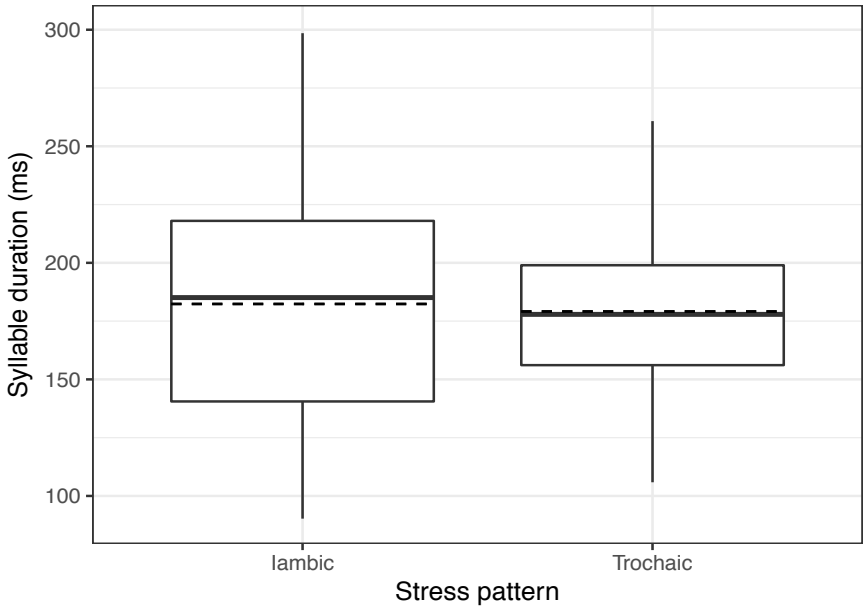


Figure 3.3: The effect of stress pattern on syllable duration.

Figure 3.4 shows the model's predicted means for syllable duration by phrasal position, where mean syllable duration is longer in final position, followed by the medial and the initial. There was a significant effect of phrasal position on syllable duration, $X^2(2) = 803.11$, $p < .0001$. For Initial position β = -20.54 ms, with SE = 0.93 ms relative to the final position and for medial position β = -1.97 ms, with SE = 0.87 ms relative to the final position. Predicted mean for initial position is 182+(-20.54) = **161.4 ms**, for medial position: **180.03 ms**, and for final position we reverse βs' signs and add them to the intercept: 182+20.54+1.97 = **204 ms**.

Figure 3.5 shows the effect of metronome period on mean syllable duration; mean syllable duration is shorter at shorter metronome periods. There was a significant effect for metronome period on syllable duration, $X^2(1) = 45.06$, $p < .01$, with β = -12.31 ms and SE = 1.31 ms. The slope, β, represents the change in syllable duration around the intercept, 182 ms, at the shortest metronome period, 1270 ms, which was assigned a value of 1.01 after treating metronome periods as continuous. Thus, prediction for the shortest period is, 182+(-12.31) = **169.6 ms**. For the longest period, we multiply β by the continuous value of the longest period: 182+(-12.31) *-0.98 = **194 ms**, and for the medium period: 182+(-12.31) *0.015= **182 ms**.

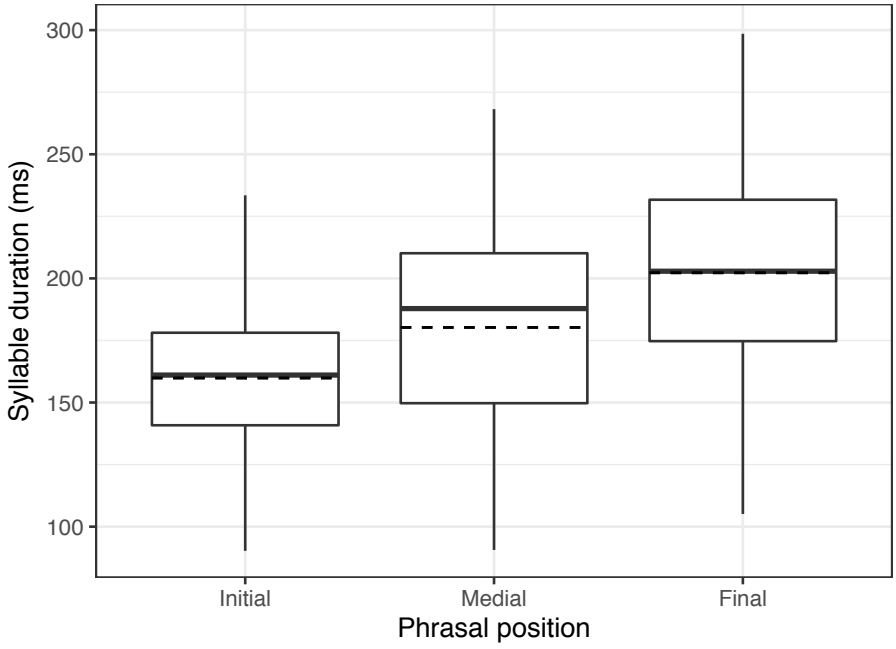


Figure 3.4: The effect of phrasal position on syllable duration.

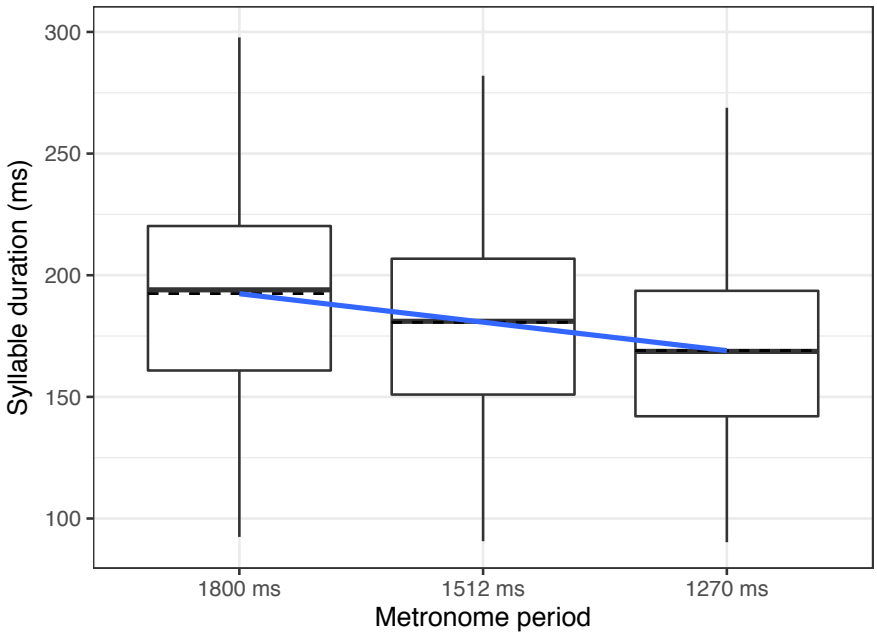


Figure 3.5: Metronome period effect on syllable duration.

Four predictors had significant main effect on syllable duration: stress pattern (iambic, trochaic), syllable stress (heavy, light, unstressed), phrasal position (initial, medial, final) and metronome period. Mean syllable duration in the iambic pattern was higher than in the trochaic pattern – the difference, however, was very small, less than 3 ms.

Not surprisingly, the difference between heavy stressed syllables and unstressed (64 ms) was larger than the difference between light stressed syllables and unstressed syllables (49 ms). Positional effects showed that mean syllable duration in final position was longer than in medial and initial positions. We cannot be sure whether the durational differences between syllables in phrase-final position and in non-final positions are due differences in prominence levels, or due to adjacency to phrase boundary. We will address this matter in Experiment 1 (c) using non-durational cues to prominence. As expected, mean syllables duration showed incremental decrease from the longest metronome period to the shortest metronome period.

We now investigate interaction effects between different predictors on syllables duration. There were no two-way interactions between dialect and syllable stress, $X^2(2) = 2.80$, $p = .1$, or between dialect and position, $X^2(2) = 1.22$, $p = .5$, or between dialect and metronome period, $X^2(1) = 1.47$, $p = .2$. There was a significant two-way interaction between stress pattern (iambic, trochaic) and syllable stress (heavy, light, unstressed), $X^2(2) = 690.42$, $p <$

.001. We used function *predict* from R package *stats* (R Core Team, 2019) to generate predictions of the interaction levels. Figure 3.6 plots predicted means of the interaction between syllable stress and stress pattern. We can see that the contrast in duration between stressed syllables, heavy and light, and unstressed syllables is greater in the iambic pattern than in the trochaic. Pairwise comparison showed that the difference in duration between heavy vs. unstressed is significantly larger in the iambic pattern than in the trochaic pattern, $p < .05$, and the difference in duration between light vs. unstressed is significantly larger in the iambic pattern than in the trochaic pattern, $p < .05$. It is not surprising that unstressed syllables are shorter in the iambic pattern than in the trochaic pattern, as unstressed syllables are of simpler phonological structure in the iambic pattern, CV, than in the trochaic pattern, CVC.

Shorter durations of unstressed syllables, with greater difference with stressed syllables, in the iambic pattern than in the trochaic pattern, may explain the earlier external and internal phase in the iambic pattern than in the trochaic pattern.

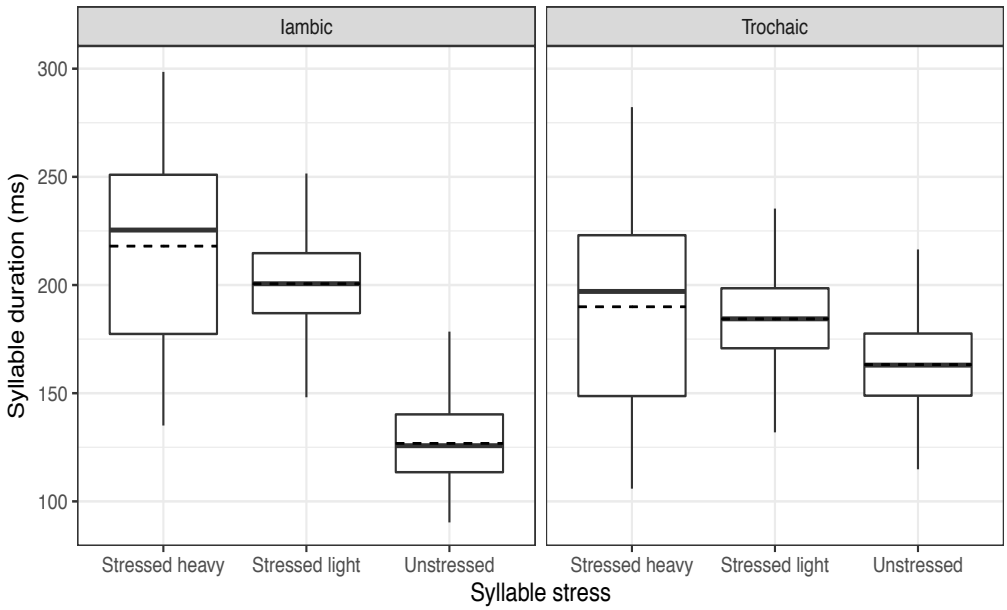


Figure 3.6: The effect of two-way interaction between stress pattern (iambic, trochaic) and syllable stress (heavy, light, unstressed) on syllable duration.

It is also noteworthy that the longer duration of unstressed syllables in trochaic than in the iambic sentences is not only due to unstressed syllables in the former being in absolute phrase-final position, thus receiving phrase-final lengthening. Rather the difference between

unstressed syllables across trochaic and iambic sentences is consistent in all positions, **38 ms**, and the effect of phrase-final lengthening on unstressed syllables durations is observed in both stress patterns, as shown in Figure 3.7.

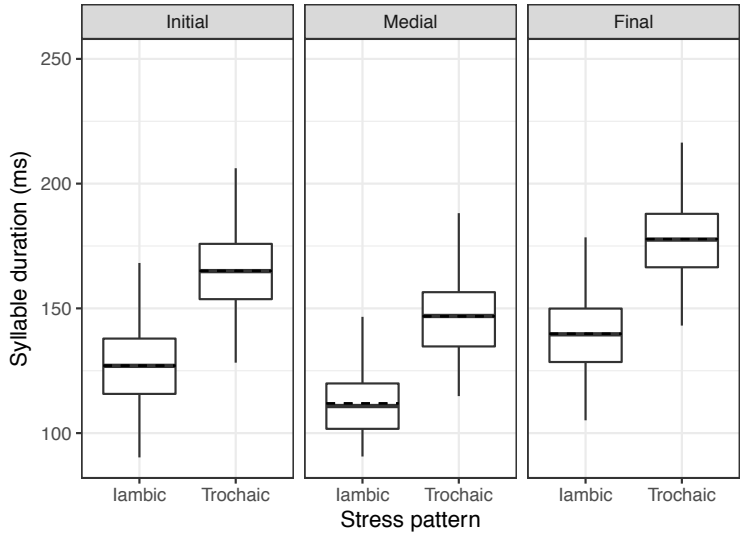


Figure 3.7: Unstressed syllables duration in iambic and trochaic patterns across different phrase positions.

There was a significant two-way interaction between syllable stress and position, $X^2(4) = 520.85$, $p < .001$. Figure 3.8 shows this interaction. From Figure 42, we note that, unexpectedly, in initial position heavy syllables are shorter than light syllables. This is probably due to a confounding factor in our text material; heavy syllables in phrase-initial position occurred only in trochaic sentences, which makes them in post-pausal position.

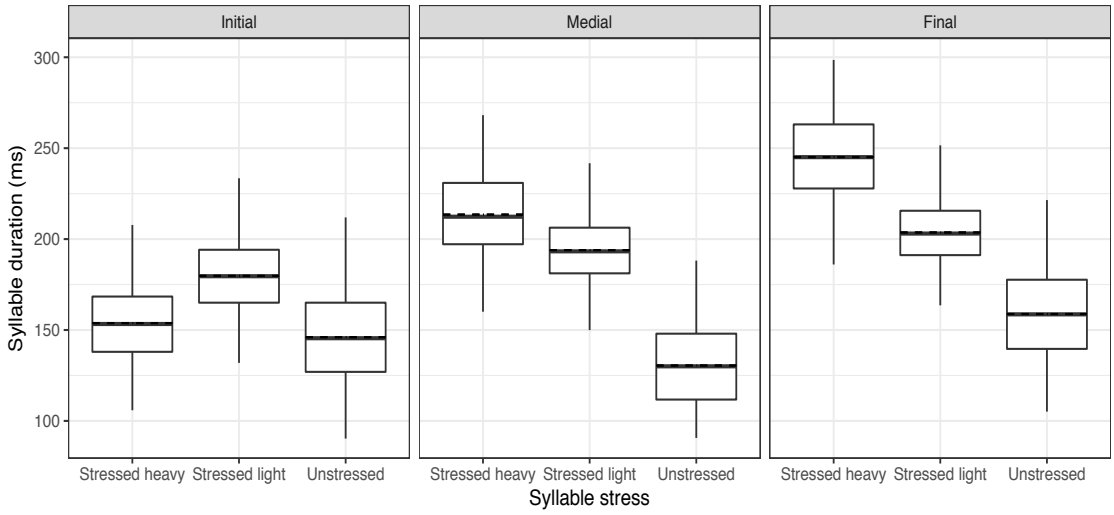


Figure 3.8: The effect of two-way interaction between syllable stress and position on syllable duration.

As such, boundaries of oral stops in the onset of heavy syllables (all onset consonants of heavy syllables in post-pausal position were oral stops) were marked by the burst energy, as closure phase could not be detected in post-pausal position. This might have contributed to the shorter duration of heavy syllables in initial position as shown in Figure 3.8. Also, this confound seems to influence the lower contrast of light vs unstressed syllables in initial position than in other positions, since some light syllables with oral stops occurred in post-pausal position.

Beyond this confound, we can see that syllables, heavy, light, unstressed, in phrase-final position are longer than in other positions, reflecting phrase-final lengthening. Pairwise comparisons showed that syllables (heavy, light, unstressed) in phrase-final position are significantly longer than in initial position; in all comparisons, $p < .05$, and syllables in phrase-final position are longer than in medial position; in all comparisons $p < .05$. Whether these trends reflect difference in prominence level or is only due to phrase-final lengthening is unclear, and we will use non-durational cues in Experiment 1 (c) to address this matter.

There were no two-way interactions between syllable stress and metronome period, $X^2(2) = 4.50$, $p = .1$, or between position and metronome period, $X^2(2) = 1.58$, $p = 0.4$. There were no three-way interactions between dialect, position and syllable stress, $X^2(4) = 7.35$, $p = .1$, or between dialect, syllable stress and metronome period, $X^2(2) = 0.05$, $p = .1$, or between dialect, position and metronome period, $X^2(2) = 3.31$, $p = 0.1$. There was a significant three-way interaction between syllable stress, position, and metronome period, $X^2(4) = 12.07$, $p = 0.01$. Figure 3.9 shows three-way interaction between syllable stress, position and metronome period. We can see in Figure 3.9 the effect of the previously discussed confound of heavy syllables in initial position being in post-pausal position, as heavy syllables are shorter than light syllables and show small contrast with unstressed syllables, in all metronome periods. Pairwise comparison showed there was no difference between heavy syllables and unstressed syllables in initial position in all metronome periods; in all comparisons $p > .05$, while in other cases (light syllables vs unstressed in initial position, heavy vs unstressed and light vs unstressed in medial and final positions in all metronome periods) there were significant differences; in all comparisons, $p < .05$.

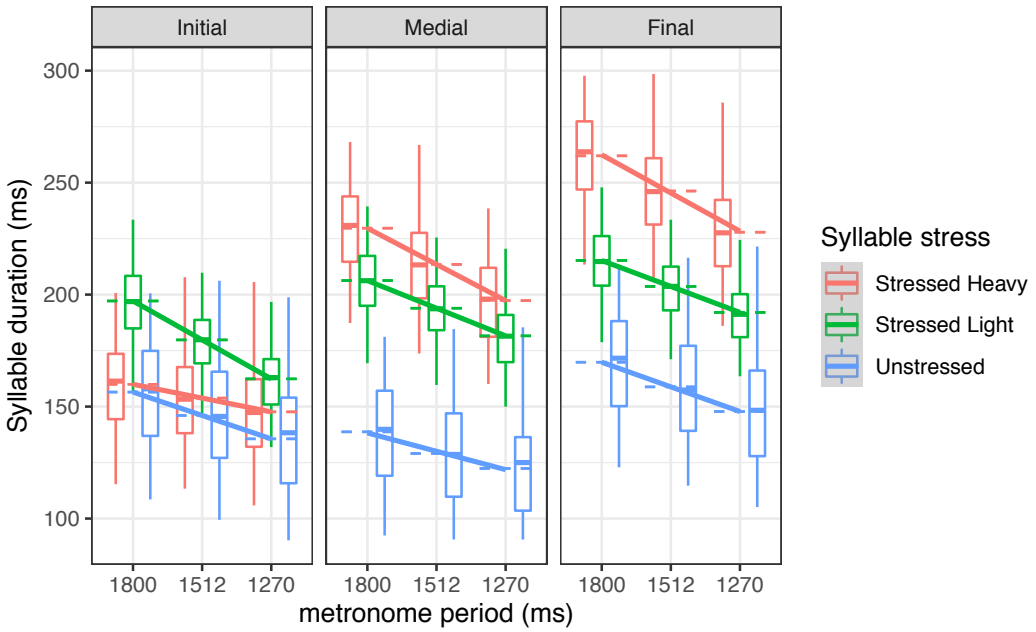


Figure 3.9: The effect of three-way interaction between syllable stress, position and metronome period on syllable duration.

There was a significant four-way interaction between dialect, syllable stress, position and metronome period, $X^2(4) = 12.07$, $p = .01$. Figure 3.10 plots the interaction. We can see that there is a noticeable between dialect difference in the duration of unstressed syllables in phrase-initial position, as there is a steeper slope across metronome periods in Hadari than in Bedouin. To compare the degree of unstressed syllable reduction in Hadari and in Bedouin, we will compare the duration of unstressed syllables at each metronome period to stressed light syllables. It may not be useful to compare unstressed syllables to stressed heavy due to the confounding effect of heavy syllables in phrase-initial position being in post-pausal position. Possibly, due to this confound, we can see in Figure 3.10 that heavy syllables, unexpectedly, are shorter than light syllables in all metronome periods, and in Bedouin they are similar in duration to unstressed syllables in all metronome periods. Therefore, the degree of unstressed syllable reduction between dialects will be assessed based on comparison with stressed light syllables.

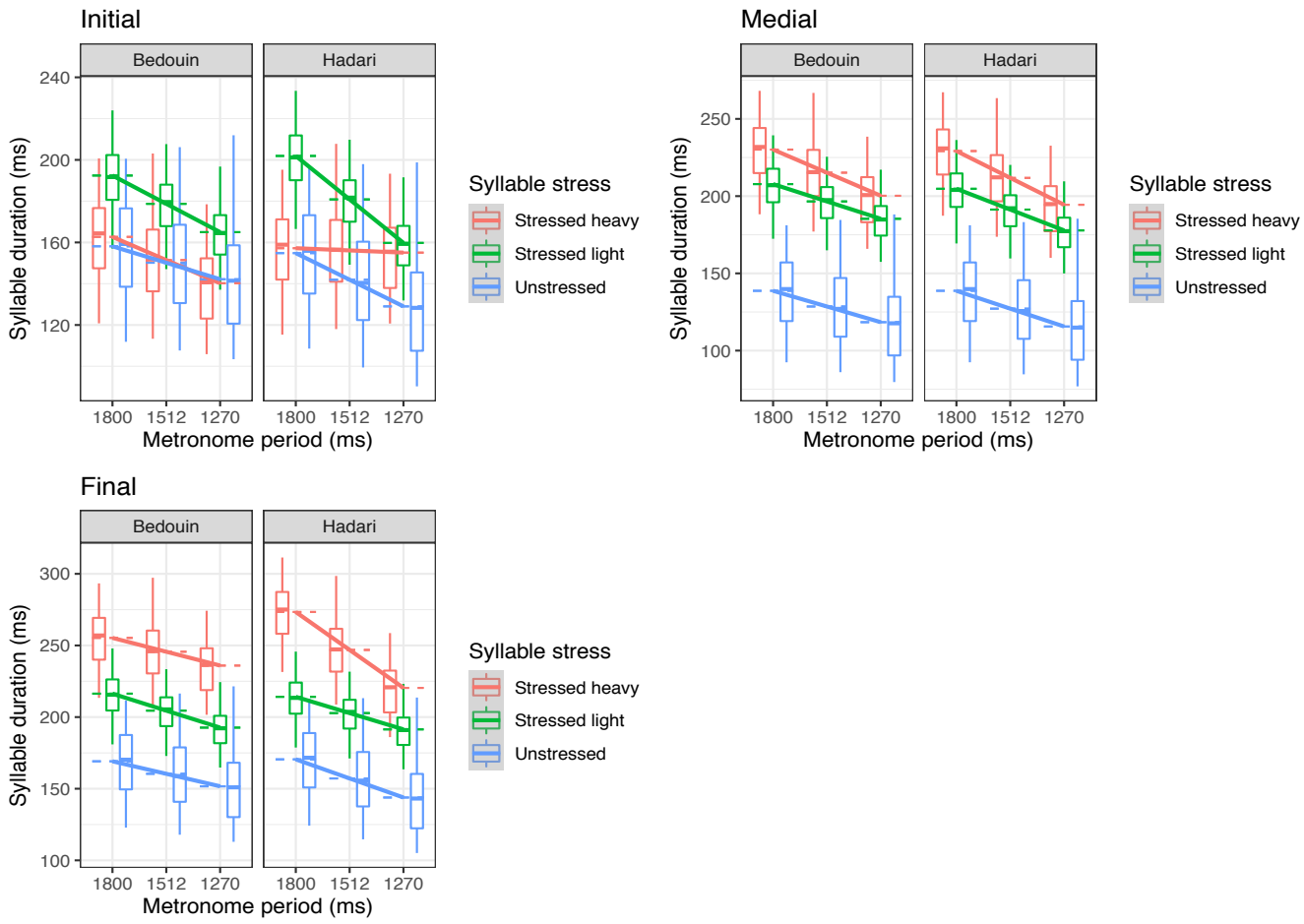


Figure 3.10: The effect of four-way interaction between dialect, syllable stress, position and metronome period on syllable duration.

We conducted two pairwise comparison tests. First, we compared the duration of stressed light to unstressed syllables in phrase-initial position in each metronome period in each dialect. Unstressed syllables were significantly shorter than stressed light in each metronome period in each dialect in phrase-initial position, $p < .0001$. Second, to compare the degree of unstressed syllable reduction between dialects, we examined if the difference in duration of stressed light vs. unstressed, was significantly different between dialects. Table 3.3 summarizes the latter test. From Table 3.3, we can see that the difference in durational contrast between light vs. unstressed syllables is of greater magnitude in Hadari than in Bedouin in phrase-initial position at the longest metronome period, $p < .05$, and at the shortest metronome period, $p < .05$. In the medium period, Bedouin showed greater contrast, and it was significant, $p < .05$; however, the difference between dialects in the durational contrast between light vs. unstressed syllables was small; 2 ms only.

Table 3.3: Pairwise comparison of the contrast between stressed light and unstressed syllables between dialects in initial position across different metronome periods.

Metronome period		Bedouin	Hadari	P-value
1800 ms	Light vs Unstressed	197-161 = **36**	207-152 = **55**	$p < .05$
1512 ms	Light vs Unstressed	182-148 = **34**	183 – 151 = **32**	$p < .05$
1270 ms	Light vs Unstressed	170-146 = **24**	164-127= **37**	$p < .05$

Thus, these is a greater degree of unstressed syllable reduction in Hadari than in Bedouin in phrase-initial position, most notably, at the longest and the shortest metronome periods.

3.5 Summary and discussion

In analysing syllable duration, we were interested in finding syllabic durational patterns that may explain phase alignment. Specifically, we speculated that the earlier internal and external phase alignment in the iambic sentences was due to the simpler structure of unstressed syllables in the iambic sentences, CV, than in the trochaic sentences, CVC (see Figures 2.9 and 2.19 for schematic illustrations). Not surprisingly, we found that unstressed syllables are shorter in the iambic sentences than in the trochaic sentences, thus potentially leading to earlier alignment of stressed syllables. Therefore, the contrasting phase alignment patterns between the iambic and the trochaic sentences are due to the different text materials between the two sets of sentences, rather than prosodic timing constraints.

As for dialectal differences, we found in the external phase measure, in a four-way interaction, that Bedouin tended to align heavy and light syllables similarly, close to 0.5 phase, in the trochaic pattern in the shortest metronome period, while Hadari tended to align light syllables earlier than heavy syllables. In the internal phase measure, in a two-way interaction, both dialects tended to align heavy syllables close to 0.5 phase, whereas the alignment of light syllables differed; Hadari aligned light syllables earlier in the phrase, and closer to 0.5, while Bedouin aligned light syllables later. We speculated that the tendency in Hadari to align light syllables earlier in the phrase is due to greater unstressed syllable reduction in Hadari than in Bedouin (see Figures 2.11 and 2.20 for schematic illustrations). The analysis of syllable duration revealed greater unstressed syllable reduction in Hadari than in Bedouin. In particular, we found in a four-way interaction that Hadari exhibited greater unstressed syllable reduction than Bedouin in phrase-initial position in the longest and the

shortest metronome periods, thus potentially leading to earlier phase alignment in Hadari than in Bedouin.

The effect of unstressed syllable reduction on phase alignment was also found in the cross-linguistic comparison between English and Japanese. As discussed earlier in section (1.6.1), Tajima (1999) examined the internal phase between English and Japanese, when the number of syllables in phrase-initial words was manipulated. This is demonstrated below. In Pattern A, there were two syllables in the phrase-initial word, while in Pattern B, there were three syllables. Japanese aligned medial stressed syllables in Pattern B later, compared to Pattern A, while English aligned the medial stressed syllables in Pattern B earlier in the phrase, and more similar to Pattern A, due to greater unstressed syllable reduction in the phrase-initial word.

Pattern A: [‘σσ] [‘σσ] [‘σ]
Pattern B: [‘σσσ] [‘σσ] [‘σ]

Thus, the degree of unstressed syllable reduction is an important predictor for the temporal organization of stressed syllables in speech cycling, between dialects and between languages.

We pointed at the beginning of this chapter to the potential difference in the metrical strength of different phrasal positions in our speech cycling corpus. We found that mean syllable duration in phrase-final position was longer than in phrase-initial and phrase-medial positions. It is not clear whether lengthening in phrase-final position is due to boundary-adjacency only, or also reflects difference in prominence level. We will use non-durational acoustic cues in the following chapter to investigate the metrical structure, in terms of relative prominence, of phrases in speech cycling.

It is noteworthy, however, that the effect of phrase-final lengthening in Bedouin and Hadari Kuwaiti dialects seems similar to other Arabic dialects. In particular, the effect of phrase-final lengthening was the greatest in phonologically heavy syllables, followed by stressed light and unstressed syllables, as indicated by the two-way interaction between position and syllable stress. This is in line with the findings in other dialects which show that stressed heavy syllables are lengthened in phrase-final position more than stressed light and

unstressed syllables (e.g., de Jong & Zawaydeh, 2002; Kelly, 2021). Probably, less phrase-final lengthening effects on short syllables, light and unstressed, is due to the functional load; as vocalic and consonantal length is phonemic in Arabic dialects, final lengthening may endanger the phonemic contrast between long and short segments, thus the effect of lengthening attenuates in short syllables, which contain short segments. Also, another aspect of similarity between Kuwaiti Arabic dialects and other Arabic dialects is that the temporal stress contrast between stressed heavy and unstressed syllables is substantially higher than the temporal contrast between stressed light and unstressed syllables. As most of the literature on Arabic dialects examined vowel duration only, we demonstrate in Figure 3.11 vowel duration, stressed long, stressed short, and short unstressed in our data. Stressed long vowels are twice as long as unstressed vowels with a difference of 54.4 ms (51 %), while the difference between stressed short and unstressed vowels is 13 ms (19 %). The difference between stressed long and unstressed vowels is well within the range reported for other Arabic dialects which ranges between 50 % and 66 %. The difference between stressed short and unstressed vowels, however, is larger than that reported for other Arabic dialects which is less than 10 %. The substantial difference between stressed long and unstressed vowels may support the idea that Arabic may have adapted stress in its phonological system to enhance the length contrast between short and long vowels (Ahn, 2002; Vogel et al., 2017).

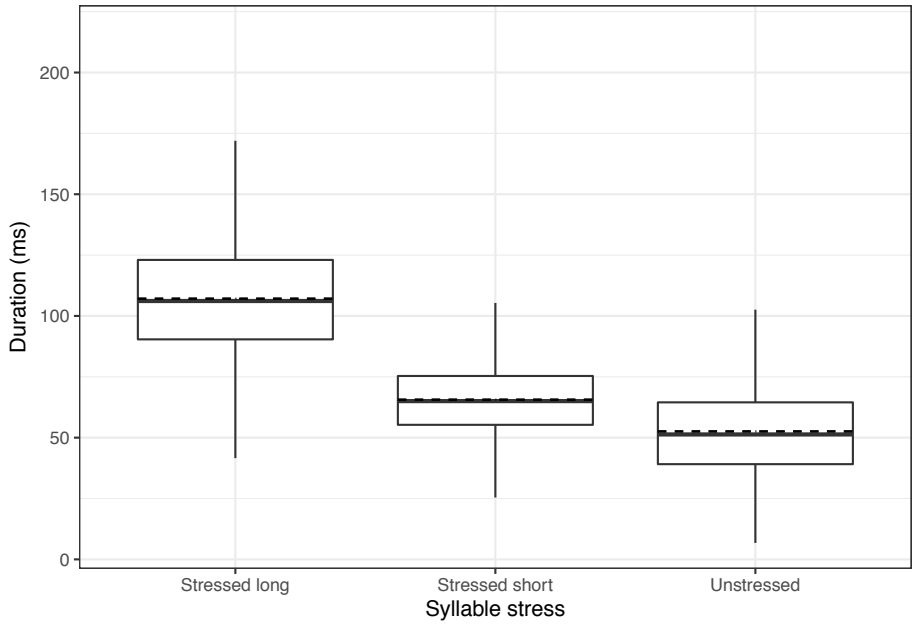


Figure 3.11: Differences in stressed long, stressed short and unstressed vowel duration collapsed over Hadari and Bedouin dialects.

Chapter 4. Experiment 1 (c): the relative strength of stress beats: supporting evidence for a hierarchical metrical structure

4.1 Introduction

The aim of this chapter is to explore the metrical structure of the relative strength of stress beats in the production of Bedouin and Hadari speakers in speech cycling.

When exploring the external phase, we found that beats of stressed syllables lie at certain points that divide the phrase repetition cycle into a simple integer ratio, specifically 1/2. The simple division of the cycle is considered as evidence for the nesting of lower prosodic units, beats of stressed syllables, within a higher prosodic unit, that is, the phrase repetition cycle, thus, reflecting a hierarchical structure. The lower units within the hierarchical structure, i.e., beats of stressed syllables, may have a metrical organization, specifically manifesting as differences in relative prominence. Thus, when we consider the 1/2 rhythmic mode, it reflects a structure where the phrase repetition cycle is divided into four beats in sentences made of three stresses, with a silent beat at the end of the sentences. Within the phrase repetition cycle, these beats may exhibit a metrical organization in terms of relative prominence. We may predict that the beats of stressed syllables at phrase-initial position will be the strongest beats, as speakers may apply greater vocal effort due to the requirement of timing phrase-initial stressed syllables with metronome beats. The relative strength of stress beats can be represented with a metrical grid (Liberman, 1975; Hayes, 1985), as in Figure 4.1.

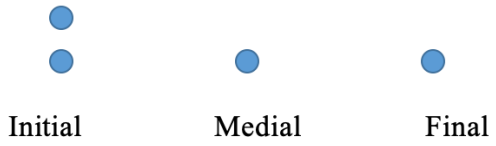


Figure 4.1: Representation of a metrical grid of relative prominence of stress beats, with the phrase initial stress beat representing the strongest beat.

We have analysed in the previous chapter syllable duration at different phrasal positions. However, uncovering prominence through duration could be confounded with other timing effects, such as boundary-related lengthening. Figure 4.2 shows the mean duration of stressed syllables (collapsed over light and heavy) by position.

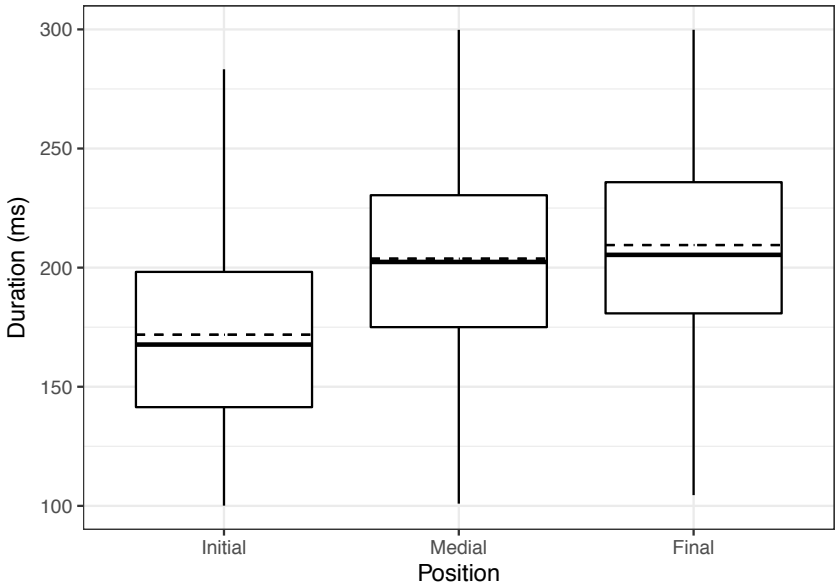


Figure 4.2: Mean duration of stressed syllables (collapsed over heavy and light) by position. Phrase final stressed syllables are the longest at 220 ms, followed by medial syllables at 206 ms and initial syllables at 172 ms.

From Figure 4.2, we can see an incremental increase in duration towards the phrase boundary, which suggests edge-related lengthening effects rather than prominence. Another potentially durational confound is related to the shorter duration of phrase-initial stressed syllables in our speech cycling data. As discussed in Experiment 1 (b), the shorter duration of phrase-initial stressed syllables, especially heavy syllables, could be due to the boundary marking of oral stops in stressed syllables onsets. The boundaries for oral stops were taken from the burst energy, as the closure phase was not detectable in post-pausal position. Marking initial stressed syllables onsets as such might have contributed to their shorter durations.

Therefore, another non-durational acoustic cue should be used to examine the relative prominence of stress beats. We will use spectral balance, which has been shown to be a reliable cue for prominence in several Arabic dialects (see section 1.4) and in other Germanic languages such as English and Dutch (Sluijter & van Heuven, 1996). This measure captures the increased intensity at higher frequency bands, thus reflecting higher vocal effort (glottal effort). It has been shown that increased vocal effort, e.g., when producing stressed syllables, is associated with an increase in intensity in harmonics above 500 Hz. The relative increase in harmonics above 500 Hz was shown to correlate more strongly with stress than overall intensity (Sluijter & van Heuven, 1996). Several researchers computed the intensity increase

in harmonics at higher bands (e.g., Heldner, 2003; Tranmüller & Erickson, 2000) by low-pass filtering the signal with a cut-off value of 1.5 of the fundamental frequency mean. Then, the overall intensity was subtracted from the intensity of the low-pass filtered signal (see van Heuven, 2018, for a review of other spectral balance measurements). The higher the value, the higher the vocal effort.

Relative prominence structure will be explored using spectral balance for both dialects, Bedouin and Hadari, to investigate whether the metrical structure as in Figure 4.1 is supported, or varies between the two dialects.

4.2 Methods

We adapted a Praat script made by Chen Gafni (https://github.com/chengafni/praat) to implement spectral balance as appeared in Heldner (2003) and Tranmüller and Erickson (2000). Spectral balance is measured as the difference between the overall intensity and the intensity in a low-pass filtered signal. The filter cut-off is 1.5 times the mean f0 of the whole utterance. Thus, we first extracted phrases, every single repetition from our speech cycling materials, to compute the mean f0 of each utterance. Then, vowels were low-pass filtered with a filter cut-off 1.5 times mean f0. Afterwards, the overall intensity of vowels was subtracted from the intensity in the low-pass filtered signal to yield a measure of spectral balance in dB. The higher the spectral balance, in dB, the higher the vocal effort in the production of a vowel. Our analysis is confined to vowels only since most prominence effects are manifested in vowels. The total number of tokens analysed was *37183*. Tables 4.1 and 4.2 show detailed analysed tokens in Bedouin and Hadari, respectively. Cells arrangement in the tables is based on the highest level of interaction in the linear mixed-effects model (dialect*position*syllable stress); see section 4.3 below. The number of tokens in metronome rates and stress patterns was not provided because these variables were considered controls in the spectral balance statistical model. However, as the total number of tokens in the spectral balance analysis is similar to that of the syllable duration analysis, the reader may refer to Tables 3.1 and 3.2 for the number of tokens in the metronome rate and stress patterns.

*Table 4. 1: Analysed tokens for spectral balance analysis in Bedouin. The total is 20238. Cells arrangement in the tables is based on the highest level of interaction in the linear mixed-effects model (dialect*position*syllable stress).*

Tokens per position	**Initial**	**Medial**	**Final**
	6761	6702	6775
Heavy	850	1866	1302
Light	2534	1455	2086
Unstressed	3377	3381	3387

Table 4. 2: Analysed tokens for spectral balance analysis in Hadari. The total is 16945.

Tokens per position	**Initial**	**Medial**	**Final**
	5641	5648	5656
Heavy	706	1601	1078
Light	2116	1222	1749
Unstressed	2819	2825	2829

4.3 Analysis

Syllable position in the phrase is an important factor that is predicted to influence metrical strength levels. We predict that phrase-initial position will have higher metrical prominence than other positions, as the alignment of phrase-initial syllables with metronomes may be associated with greater vocal effort. Phonological syllable stress - heavy, light, and unstressed - is predicted to affect prominence; the literature shows that lexically stressed syllables manifest greater phonetic prominence effects, such as higher spectral balance, and phonologically heavy syllables exhibit higher spectral balance than light syllables (Heldner, 2003). Thus, dialect (Bedouin and Hadari), position (initial, medial, final), and syllable stress (heavy, light, unstressed) were included in the linear mixed-effects model as predictors. Metronome period and stress pattern (iambic, trochaic) were included as controls since we do not have specific predictions of their effects on prominence levels.

Two-way interactions between dialect and syllable stress, dialect and position, and syllable stress and position were included in the model, as well as three-way interactions between dialect, syllable stress and position.

All predictors were centred to obtain a meaningful interpretation of the intercept. When centring the predictors, the intercept would represent the mean of all predictors, and the slopes of main effects would represent the amount of change around the intercept.

Speaker and sentence were included as random intercepts and, by speaker random slopes for syllables stress, position, stress pattern and metronome period were included in the model, since these factors vary within speakers. By sentence random slopes for dialect and position were included in the model, as these factors vary within sentences. The model did not converge with a by sentence random slope for metronome period, thus it had to be removed from the model.

Linear mixed-effects model, as well as likelihood ratio tests for significance testing, were conducted using package *afex* (Singmann et al., 2016) in R software. Pairwise comparison to compare the means of interactions levels, through by-subjects two-tailed t-test, was conducted using R package *phia* (Rosario-Martinez et al., 2015).

4.4 Results

The intercept value which represents the mean of all predictors is 3.59 dB. Figure 4.3 shows the effect of dialect on spectral balance, with Hadari having higher vowel spectral balance than Bedouin. There was a significant effect for dialect, $X^2(1) = 9.06$, $p = .003$. For Bedouin, β = -0.66 dB, and SE = 0.20 dB, with Hadari as the reference level. To obtain the model's prediction for Bedouin we add β to the intercept, 3.59+(-0.66) = **2.93 dB**, and for Hadari, the reference level, we reverse β's sign, 3.95+0.66 = **4.25 dB**.

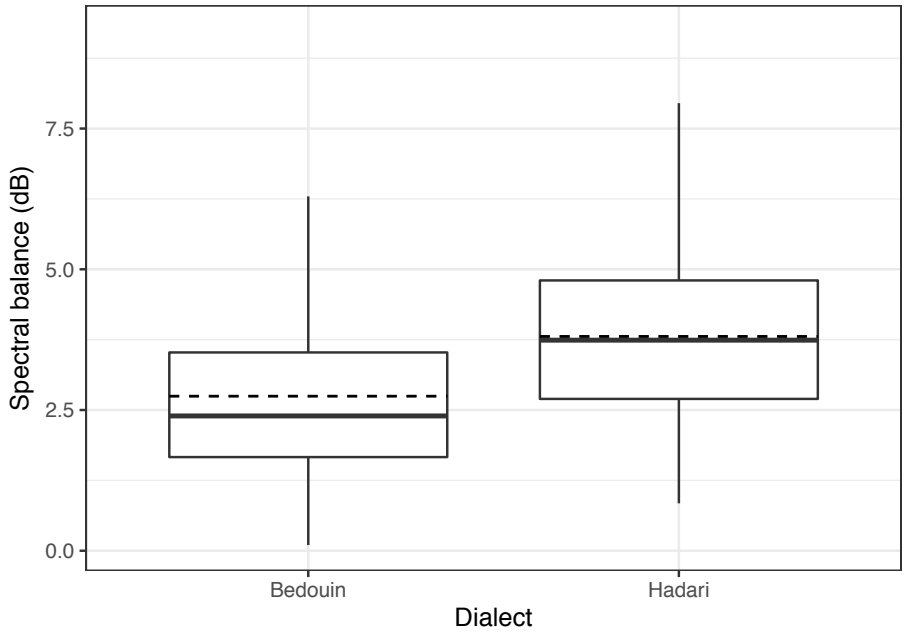


Figure 4.3: The effect of dialect on the mean vowel spectral balance.

Figure 4.4 illustrates the effect of syllable stress on spectral balance, with heavy syllables having the highest spectral balance followed by light and unstressed syllables.

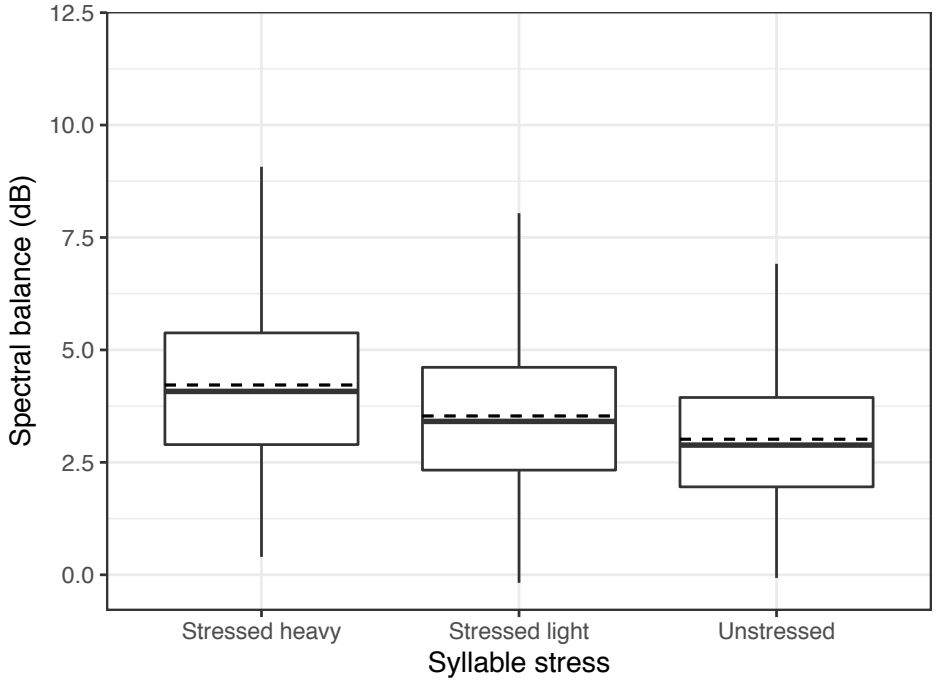


Figure 4.4: Syllable stress effects on spectral balance.

There was a significant effect for syllable stress, $X^2(2) = 73.656$, $p < .001$. For heavy syllables, β = 0.63 dB, and SE = 0.04 dB, with unstressed in the reference level and for light syllables, β = -0.06 dB, with SE = 0.02 dB, with unstressed being in the reference level. Larger β for heavy syllables than light syllables indicate that heavy syllables show greater contrast with unstressed syllables than light syllables. Model's prediction for heavy syllables is 3.59 + 0.63 = **4.22 dB**, for light syllables 3.59-0.06 = **3.53 dB**, and as unstressed is the reference level we reverse βs' signs and add them to the intercept: 3.59+(-0.65)+(0.06) = **3 dB.**

Figure 4.5 shows the effect of position on mean vowel spectral balance, with the initial position showing the highest spectral balance value followed by the final and medial positions.

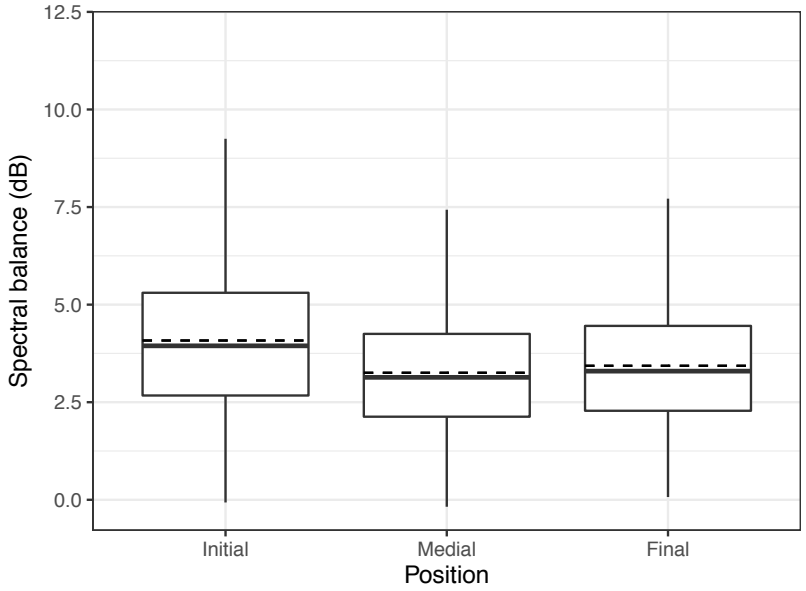


Figure 4.5: The effect of phrasal position on spectral balance.

There was a significant effect for phrasal position, $X^2(2) = 63.63$, $p < .001$. For the initial position, β = 0.49 dB, and SE = 0.05 dB, with the final position in the reference level, and for the medial position β = -0.34 dB, and SE = 0.04 dB, with the final position as the reference level. Predicted means for the initial position is 3.59 + 0.49 = **4.09 dB**, for medial position, 3.59+(-0.34) = **3.25 dB**, and for the final position, the reference level we reverse βs' signs: 3.59+(-0.49)+(0.03) = **3.44 dB.**

Figure 4.6 demonstrates spectral balance values at different metronome periods. There was no significant effect for metronome period, $X^2(1) = 2.99$, $p = 0.08$.

Spectral balance by stress pattern is shown in Figure 4.7. There was no significant effect for stress pattern on spectral balance, $X^2(1) = 0.23$, p = 0.6.

Main effects of dialect, syllable stress and position are in line with our predictions. Hadari exhibited higher spectral balance value than Bedouin. Not surprisingly, the contrast between stressed heavy and unstressed syllables is higher than between stressed light syllables and unstressed syllables. Phrase-initial position associated with spectral balance value higher than other positions indicating that it is the position with highest metrical prominence level.

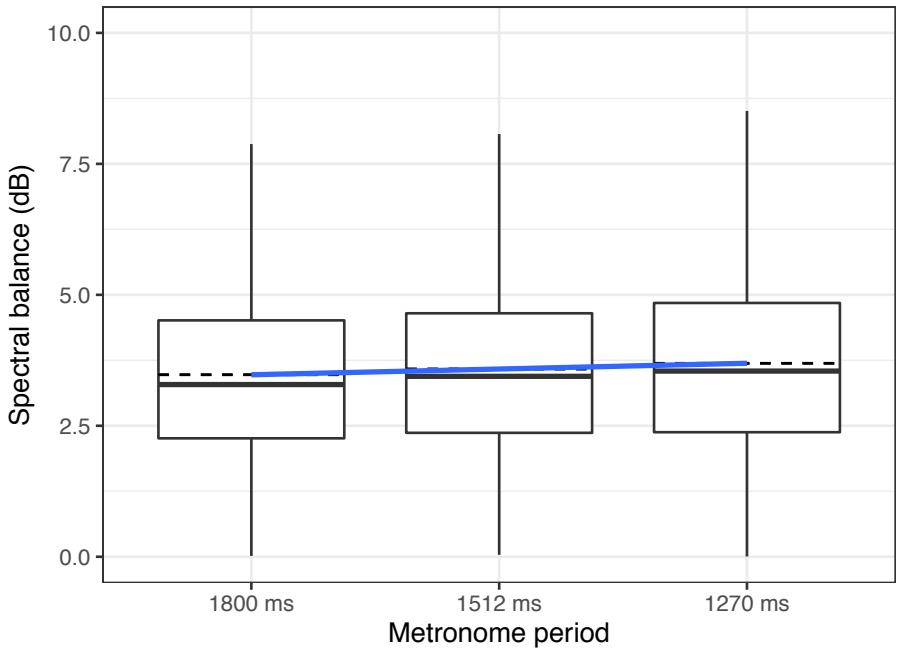


Figure 4.6: The effect of metronome period on spectral balance.

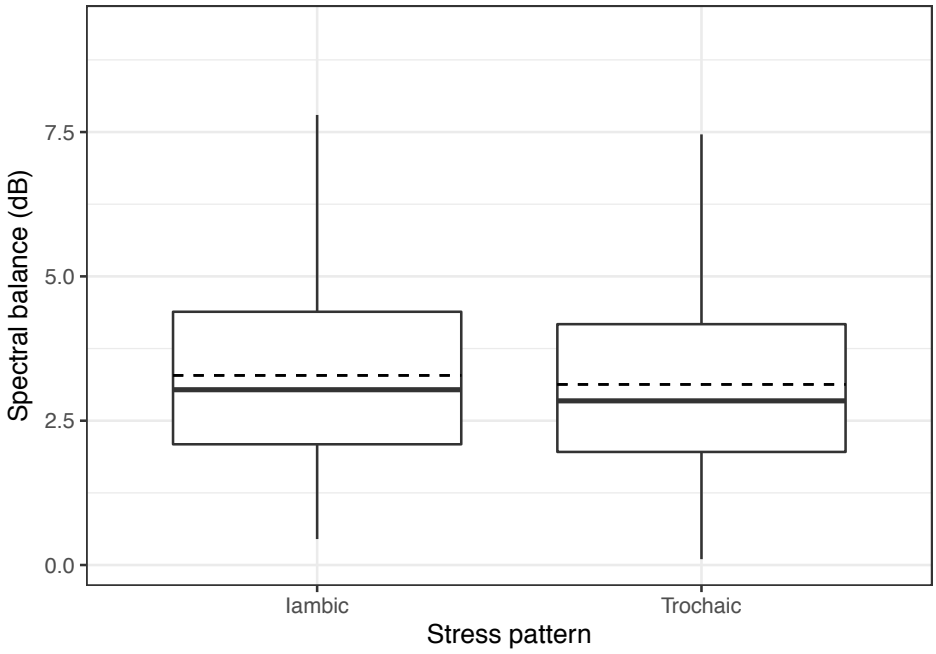


Figure 4.7: The effect of stress pattern on spectral balance.

We now investigate whether the syllable stress contrast and positional effects vary between dialects through two-way interactions. There was a significant two-way interaction between dialect and syllable stress, $X^2(2) = 7.18$, $p = .02$. We used function *predict* from R package *stats* (R Core Team, 2013) to generate predictions of the interaction levels. Figure 4.8 plots

the interaction between dialect and syllable stress. In pairwise comparisons, we compared the difference between stressed syllables, heavy and light, and unstressed syllables in each dialect. There was a significant difference in spectral balance between heavy and unstressed syllables in Hadari, $p < .0001$, and in Bedouin $p < .0001$, and there was a significant difference between light and unstressed syllables in Hadari, $p < .0001$, and in Bedouin, $p < .0001$. To examine the degree of stress contrast between dialects, we examined whether the *difference* in spectral balance between stressed syllables (heavy or light), and unstressed syllables *differs* between pairs of dialect group. Thus, the dependent variable in this comparison is the difference in spectral balance between stressed, heavy and light, and unstressed syllables. Table 4.3 summarises the statistical test.

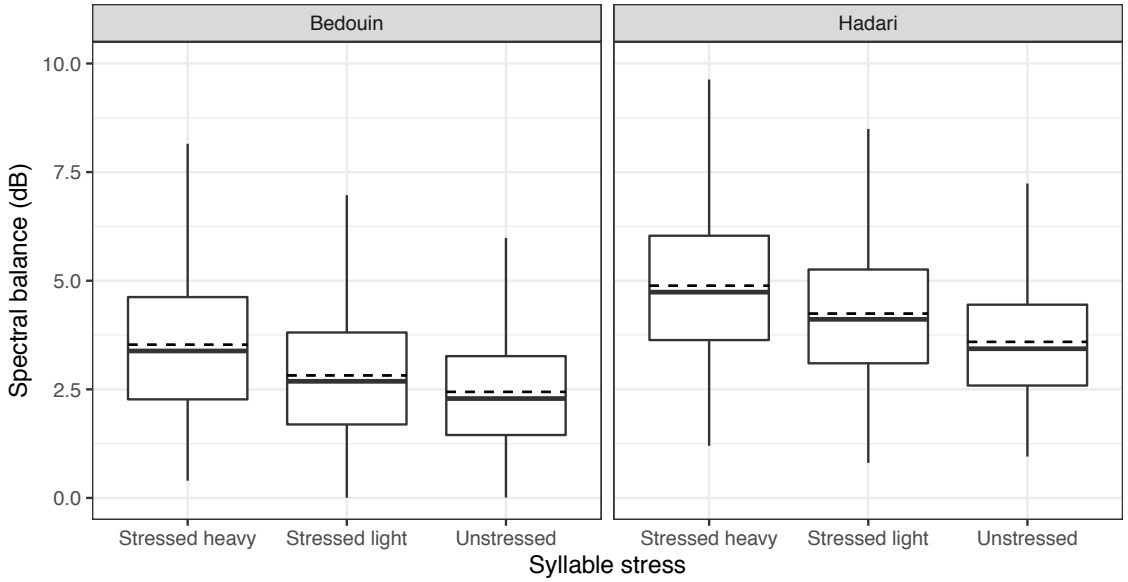


Figure 4.8: The effect of two-way interaction between dialect and syllable stress on spectral balance.

Table 4.3: Pairwise comparison of the difference in spectral balance between stressed and unstressed syllables across dialects.

	Bedouin	Hadari	P-value
Heavy vs Unstressed	3.53-2.44 = **1.09**	4.91-3.59 = **1.32**	$p = .2$
Light vs Unstressed	2.82-2.44 = **0.38**	4.24-3.59 = **0.65**	$p = .01$

From Table 4.3, we can see that the only significant between dialects difference is in the contrast of light vs unstressed syllables, $p = .01$, with Hadari tending to show greater contrast than Bedouin.

There was a significant two-way interaction between dialect and phrasal position, $X^2(2) = 10.01$, $p = .004$. Figure 4.9 plots the interaction.

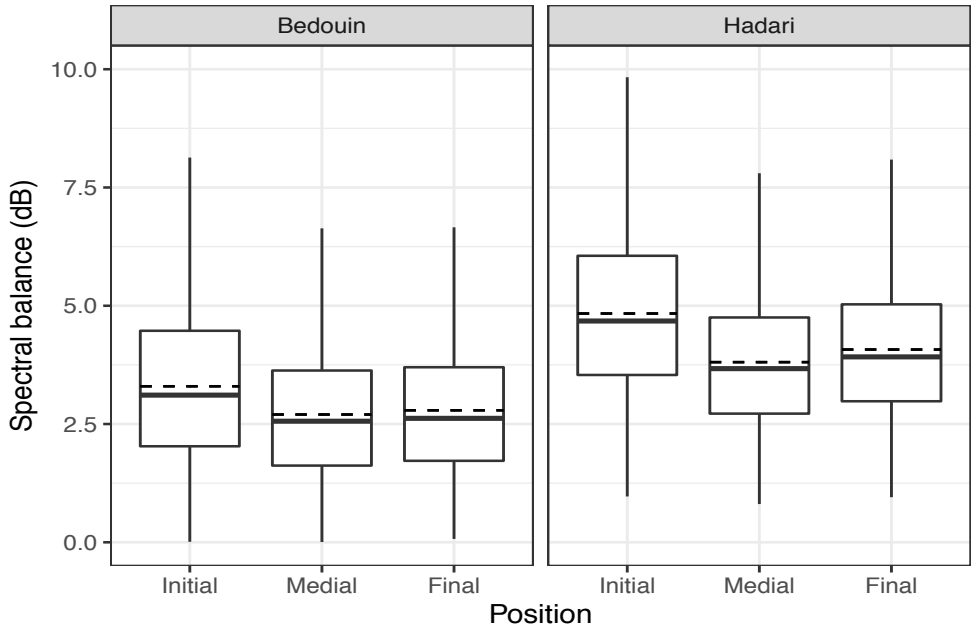


Figure 4. 9: The effect of two-way interaction between dialect and position on spectral balance.

In pairwise comparison, we examined the difference in spectral balance between different positions in each dialect. There was a significant difference in spectral balance between initial and medial positions in Bedouin, $p < .001$, and in Hadari, $p < .001$, and a significant difference between initial and final positions in Bedouin, $p < .001$, and in Hadari, $p < .001$. There was no difference between medial and final positions in Bedouin, $p = .3$, whereas in Hadari the difference was significant, $p = .03$, with the final position showing higher spectral balance value than the medial, **4.07 dB** vs **3.8 dB.** This indicates that Hadari contrasts between three levels of phrasal prominence: initial position vs. medial position, initial position vs. final position, and final position vs. medial position, whereas Bedouin only contrasts between initial and non-initial positions.

There was a significant two-way interaction between syllable stress and phrasal position, $X^2(4) = 504$, $p < .0001$. Figure 4.10 plots the interaction. The interaction seems to be due to the lower contrast between light and unstressed syllables in the medial position. Post-hoc pairwise comparison showed no difference in spectral balance between light and unstressed syllables in medial position, $p = .1$, while the difference between heavy and unstressed

syllables in medial position was significant, $p < .0001$, and in initial and final positions the difference between heavy and unstressed syllables and light and unstressed syllables was significant; in all comparisons, $p < .0001$. The non-significant difference between light and unstressed syllables in medial position, may explain the lower spectral balance of medial position main effect, than initial and final positions, Figure 4.5.

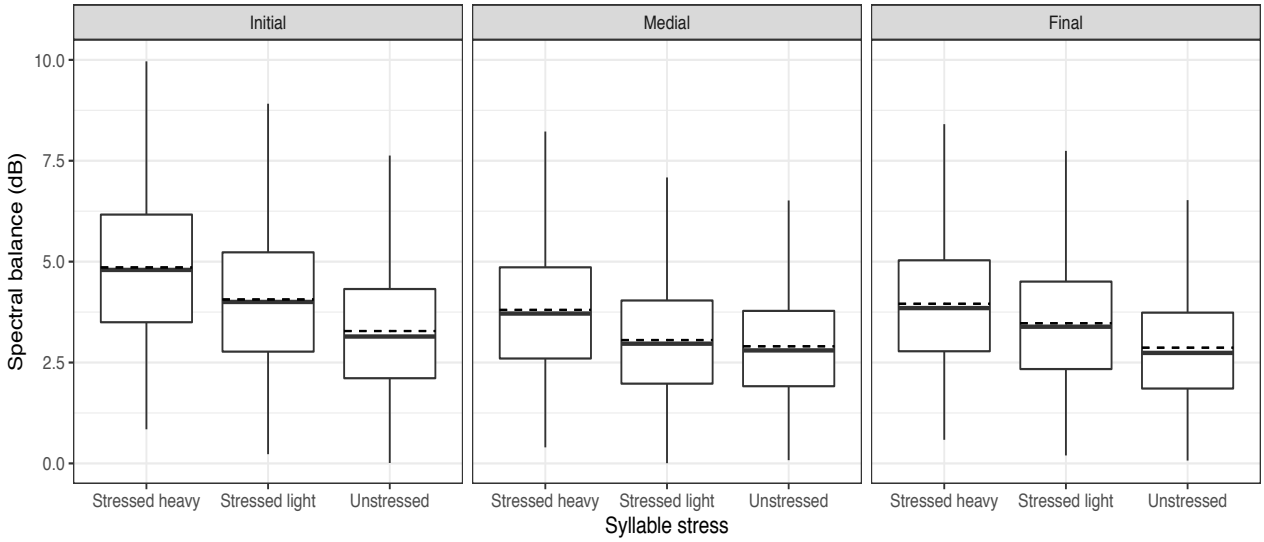


Figure 4.10: The effect of two-way interaction between syllable stress and position on spectral balance.

There was a significant three-way interaction between dialect, syllable stress and position, $X^2(4) = 72.77$, $p < .001$. Figure 4.11 plots the interaction. In pairwise comparisons, we examined whether the *difference* in spectral balance between stressed syllables (heavy or light) and unstressed syllables *differs* between dialects across positions. Thus, the dependent variable in this comparison is the *difference* in spectral balance between stressed syllables (heavy or light), and unstressed syllables.

There was a significant difference in heavy vs. unstressed spectral balance between dialects in initial position, $p = .003$, a significant difference in light vs. unstressed spectral balance between dialects in initial position, $p = .003$, and a significant difference in light vs. unstressed spectral balance between dialects in final position, $p = .005$, with Hadari showing greater contrast in all positions. Table 4.4 summarises the statistical test.

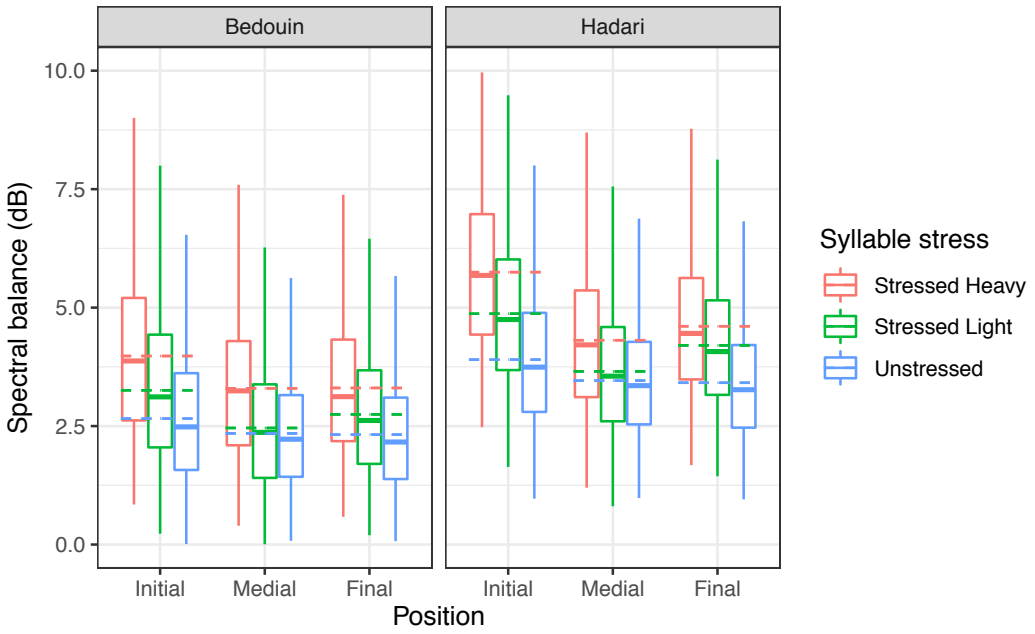


Figure 4.11: The effect of three-way interaction between dialect, syllables stress and position on spectral balance.

Table 4.4: Pairwise comparison of the difference in spectral balance between dialects across stressed and unstressed syllables in initial and final positions.

Initial	Bedouin	Hadari	p-value
Heavy vs Unstressed	3.98-2.65 = **1.33**	5.82-3.90 = **1.92**	*p = .003*
Light vs Unstressed	3.26-2.65 = **0.61**	4.88-3.90 = **0.98**	*p* = .003
Final	Bedouin	Hadari	p-value
Light vs Unstressed	2.75-2.32 = **0.43**	4.20-3.41 = **0.79**	*p* = .005

Thus, Hadari tends to exhibit greater contrast in spectral balance than Bedouin between stressed syllables (heavy and light) and unstressed syllables in phrase-initial position, and greater contrast in spectral balance between light and unstressed syllables in final position. This is reflective of the fact that Hadari contrasts between three prominence levels, as shown earlier, with the difference between dialects in final position confined to the contrast of light vs. unstressed.

4.5 Summary and discussion

The aim of the analysis of spectral balance was to explore the hierarchical metrical structure in the production of Bedouin and Hadari speakers in speech cycling. We found that Hadari

realises three levels of metrical prominence: the strongest metrical position is the initial position and the second strongest is the final position, while the medial position is the weakest. Bedouin, on the other hand, realises two levels of metrical prominence: the stronger metrical position is the initial position, and the weaker positions are medial and final. We may represent the metrical structure in the two dialects with a grid-based representation of relative prominence as in Figure 4.12, below.

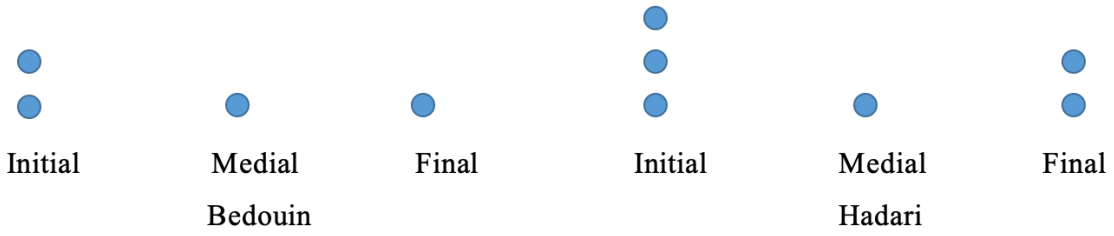


Figure 4.12: Grid-based representation of metrical structure in Bedouin and Hadari.

The difference in spectral balance between dialects in different positions varied, however, based on syllable stress. Hadari showed greater contrast between stressed (heavy or light), and unstressed syllables than Bedouin in phrase-initial position. In final position, however, the difference was only in the degree of contrast between light and stressed syllables, with Hadari showing greater contrast.

Such gradient differences in the phonetic realization of higher-level prosodic effects are important in dialectal differentiation. White et al. (2012) showed in a discrimination task that Welsh English and Orlando English were discriminated by English listeners based on gradient degrees of higher-level prosodic effects. The difference between English accents, however, was due to edge-related cues, i.e., final lengthening, rather than prominence. Also, White et al. (2009) found that speakers of Sicilian Italian demonstrated greater lengthening effects of utterance-final nuclear stressed vowel compared with utterance-medial pre-nuclear stressed vowels than speakers of Venetan Italian.

Given the contrasting metrical structure between Hadari and Bedouin, we can assert that the alignment of vowel onsets at harmonic phase angles not only reflects simple division of the phrase repetition cycle but is associated with a hierarchical metrical structure in terms of relative prominence.

The varying degrees of metrical strength of stress beats in speech production may also reflect a perceptual attribute of stress beats. For instance, we reviewed in section (1.5.4.2) Allen's (1972a,b) work, which showed that stresses with a higher level of prominence attracted more taps than stresses with a lower level of prominence, thus reflecting the perception of hierarchical metrical structure. The notion of metrical structure may explain why studies that studied perceptual isochrony did not find differences between "stress-timed" and "syllable-timed" languages. For instance, we reviewed in section (1.2.2) Scott et al.'s (1985) study, which examined isochronous tapping to stresses and syllables in English and French. Scott et al. found that isochronous tapping to English and French sentences was similar, despite having different rhythm classes, i.e., "stress-timed" vs. "syllable-timed". Possibly, the participants could not react differently to English and French sentences because the task of regular tapping does not capture the gradient metrical nature of speech.

The dissociation between isochrony and the metrical was addressed explicitly in Katz et al.'s (2015) study on the perception of musical meter. Katz et al. examined French listeners' reaction to 6/8 and 3/4 meters, Figure 4.13.

Figure 4.13: 6/8 and 3/4 metrical patterns.

In these meters, the second and fourth beats divide the beats sequence isochronously, at one-third (1/3) and two-thirds (2/3), respectively. Also, in both metrical patterns, the third beat is 1/2 of the beats sequence. Importantly, however, the second and the fourth beats are weak in 6/8 meter, and the third is weak in the 3/4 meter. Thus, for example, if the third beat in the 6/8 pattern attracts a stronger reaction than the third beat in the 3/4 pattern, this would indicate a role for metrical structure in processing the metrical pattern, rather than isochrony and simple divisions. They found that French listeners fall into two groups: a group that reacted slowly to strong beats and a group that reacted fast to strong beats. Crucially, those who reacted slowly to the strong beat in the 6/8 meter (third beat) also reacted slowly to the

strong beats in the 3/4 meter (second and fourth), and those who reacted fast to the strong beat in the 6/8 meter also reacted fast to the strong beats in the 3/4 meter. Consistency in speed in the reaction to strong beats in both types of meter, although they achieve different simple divisions, implies that they were reacting to a hierarchical metrical structure rather than isochronous, simple divisions.

In all, stress beats have a hierarchical metrical structure in speech cycling. The metrical structure of speech is also important in perceiving differences between languages (White et al., 2012).

Chapter 5. Experiment 2: mutual timing influences between stress feet and syllables in Hadari and Bedouin Kuwaiti dialects

5.1 Introduction

In the phase measurements Experiment 1(a), we have shown that vowel onsets of stressed syllables lie at a simple phase, 1/2, within the Phrase Repetition Cycle, reflecting a hierarchical nesting relationship between vowel onsets of stressed syllables and the Phrase Repetition Cycle. Dialectal differences in the organization of vowel onsets within the phrase repetition cycle were also observed, with various alignments of heavy and light syllables vowel onsets around the simple phase of 1/2. In this section, we examine the potential hierarchical timing relation between stress feet and syllables in Hadari and Bedouin in our speech cycling corpus.

Lenneberg (1967) suggested that there is an underlying frequency period that ranges from around 5 Hz (200 ms) to 7 Hz (142 ms) that controls syllables' production. In line with this suggestion, Greenberg et al. (2003) investigated the power spectrum of low-pass filtered amplitude envelope in the Switchboard corpus of American English spontaneous speech. Low-pass filtered amplitude envelope is said to exhibit slow energy fluctuation that corresponds to the alternation between vowels and consonants. Therefore, the power spectrum of the amplitude envelope might reflect linguistic information regarding syllabic temporal distribution (Rosen, 1992). Figure 5.1 represents the power spectrum of 30 minutes' materials from the Switchboard corpus as analysed by Greenberg et al. The spectrum shows a peak at 5 Hz (200 ms), with a broad energy distribution between 2 Hz and 10 Hz. The power spectrum reflects the temporal syllabic distribution in the Switchboard corpus. The mean syllable duration is 200 ms, which corresponds to the frequency peak at 5 Hz. The frequency range between 4 to 6 Hz (250 ms – 160 ms) represents stressed and unstressed syllables durations. Word durations, especially, content words, are represented in the spectrum with a low frequency at 2 Hz.

Tilsen and Johnson (2008) analysed the power spectrum of the low-pass filtered amplitude envelope from the Buckeye corpus (Pitt et al., 2005), which is made up of conversational speech of American English. Figure 5.2 shows the amplitude envelope of a stretch of speech of around 2 seconds, and Figure 5.3 shows the power spectrum of that amplitude envelope.

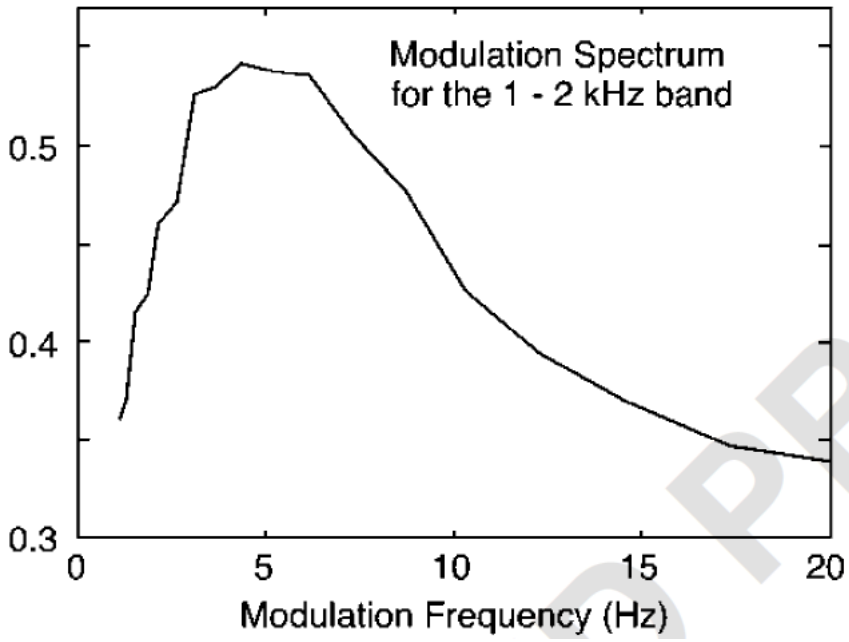


Figure 5.1: Power spectrum of the low-pass filtered amplitude envelope from 30 minutes' material from Switchboard corpus. The x-axis represents the power scale and the y-axis represents frequency peaks in Hz. It can be seen that there is a peak at 5 Hz (200 ms), which corresponds to stressed syllables duration. (Greenberg et al., 2003, p. 7).

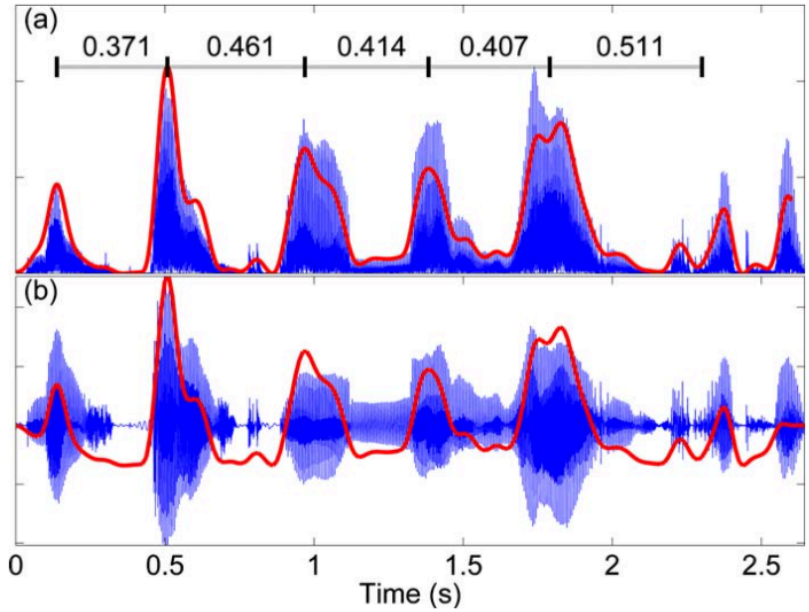


Figure 5. 2: In (a) the energy envelope superimposed over the magnitude of the band-passed signal. In (b), the energy envelope is superimposed over the original signal of a stretch of speech of 2 seconds. Intervals between the highest peaks are also shown. (Tilsen & Johnson, 2008).

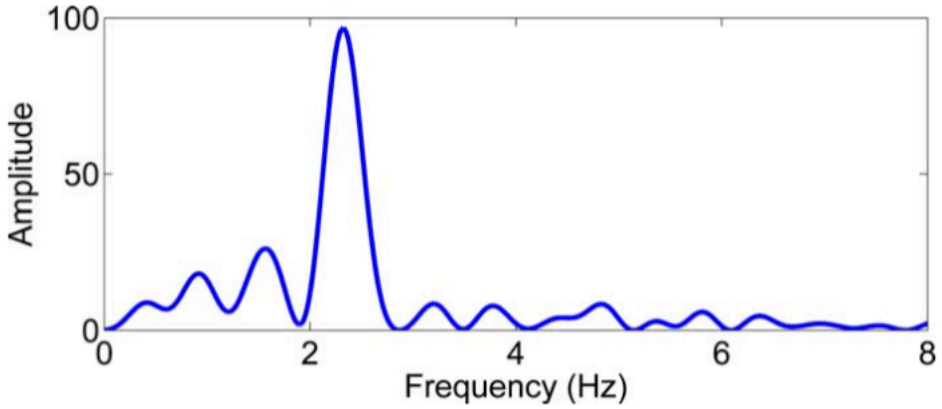


Figure 5.3: spectral representation obtained by a Fast Fourier Transform of low-pass filtered energy envelope. (Tilsen & Johnson, 2008).

It can be seen from Figure 5.3 that there is a spectral peak at ~2 Hz. This peak corresponds to the duration of suprasyllabic units such as stress feet at around 500 ms. Tilsen and Johnson asserted that the power spectrum of the amplitude envelope might be useful in characterising the "rhythmicity" of a certain chunk of speech. They found that in the Buckeye corpus in chunks ranging between 2 to 3 seconds, 23 % had frequency peaks that correspond to the durations of suprasyllabic units such as stress feet. This indicates that in such chunks, there is a tendency for a regular occurrence of stress feet at a certain durational window. It is possible that this proportion of chunks, 23 %, includes repetitions of words that may be due to hesitations, which can be associated with a semi-regular occurrence of stress feet, leading to a peak at ~2 Hz of the envelope's power spectrum (see Tilsen, 2006 on the relationship between regular stress feet and hesitations).

The findings from Greenberg et al. (2003) and Tilsen and Johnson (2008) show that the power spectrum of the energy envelope contains information regarding prosodic structure of speech. While Tilsen and Johnson's findings emphasised semi-regular occurrence of suprasyllabic units, potentially due to repetitions, the analysis of a more spontaneous and fluent speech as in Greenberg et al. reflected multiple temporal distributions of syllabic and suprasyllabic structures in speech. Crucially, however, statistics of the amplitude envelope, as obtained by Greenberg et al. and Tilsen and Johnson, do not reflect mutual timing effects between syllabic and suprasyllabic units, specifically, stress feet. They only provide an "acoustic signature" of the temporal distribution of different prosodic units.

As formulated by the coupled oscillators model (O'Dell & Nieminen, 1999), the interaction between different levels of the prosodic hierarchy implies (a) the dominance of either higher- or lower-level prosodic units, and (b) as a consequence of the dominance of a certain prosodic unit, the natural frequency of the weaker level will be prone to more changes in time (see section 1.5.1 for an overview). Tilsen and Arvaniti (2013) obtained different statistical metrics from the energy envelope to quantify potential mutual timing effects between stress feet and syllables across different languages with different rhythmic characteristics. There are two key aspects of these metrics. First, they quantify the dominance of either stress feet or syllables by quantifying power distribution at frequency rates that correspond to stress feet and syllables. For example, greater power distribution in the power spectrum at frequency ranges from 2 to 3 Hz would indicate the dominance of stress feet. Second, they quantify the change in the "instantaneous frequency" of stress feet and syllables in time. For instance, if stress feet oscillation was more dominant than syllabic oscillation, there would be more variability in the instantaneous frequency of syllabic oscillation than stress feet oscillation. Tilsen and Arvaniti showed that power distribution and variability in instantaneous frequency metrics classified languages in a way that concords, in general, with their rhythmic characteristics. For example, English showed relatively greater variability in syllabic instantaneous frequency than Spanish, and relatively lower variability in stress feet instantaneous frequency than Spanish, thus, potentially reflecting the dominance of stress feet in English. Thus, such metrics are promising in quantifying the potential mutual timing effects between stress feet and syllables across Hadari and Bedouin dialects. These metrics will be used in our analysis in this chapter and will be described in more detail in the methods section below.

5.2 Methods

Each phrase (repetition) in the speech cycling data was extracted into a separate sound file for the amplitude-based analyses. The number of analysed tokens (phrase) for each of the seven statistical metrics we will obtain from the envelope, see below, is provided for Bedouin and Hadari in Tables 5.1 and 5.2, respectively. Cells arrangement is based on the higher level of interaction (dialect*metronome rate*stress pattern); see section 5.4.

Table 5. 1: Total analysed tokens (phrases) in Bedouin's spectro-temporal data: 2788. *This is the same number of phrases analysed in the external phase data. Note that these numbers should be multiplied by 7, corresponding to the number of statistical metrics. Cells arrangement is based on the higher level of interaction (dialect*metronome rate*stress pattern).*

Tokens per rate trial	**Slow**	**Medium**	**Fast**
	1049	1075	1036
Iambic	528	543	509
Trochaic	521	532	527

Table 5. 2: Total analysed tokens in Hadari's spectro-temporal data: 3160.

Tokens per rate trial	**Slow**	**Medium**	**Fast**
	1049	1075	1036
Iambic	528	543	509
Trochaic	521	532	527

5.2.1 Obtaining the amplitude envelope

As we have demonstrated above, low pass filtered amplitude envelope contains frequency modulations that broadly correspond to syllable level and foot level time scale. Thus, we will describe methods to extract the amplitude envelope and relevant processes for our analysis. We used R package *seewave* (Sueur et al., 2008) to obtain the amplitude envelope.

From our speech cycling corpus, each utterance (repetition) was separated into a single sound file. The sound files were downsampled from 44100 Hz to 16000 Hz in order to ensure accurate transmission of the wideband speech signal (Villing et al., 2004). In extracting the amplitude envelope, recall that it represents alternations between consonants and vowels in the signal, with vocalic energy representing local peaks. Thus, when processing the signal, we aim at retaining most of the vocalic energy while de-emphasising consonantal energy so that the envelope is represented with local peaks that correspond to vocalic energy (Tilsen & Arvaniti, 2013, p. 630). Therefore, the signal is bandpass filtered in the range from 500 Hz to 4000 Hz through a second order Butterworth filter. The lower cut-off frequency significantly attenuates the contribution of the fundamental frequency in the signal. This renders voiced consonants similar to voiceless consonants and further distinguishes them from vowels, as vowel formants energy is preserved within the cut-off ranges. The higher cut-off frequency (4000 Hz) attenuates the contribution of high frequency bands energy of fricatives and bursts so that they are not represented with peaks in the envelope. Thus, with these cut-off

frequencies (500 Hz – 4000 Hz), the contribution of vocalic energy is emphasised relative to consonants. Figure 5.4 and Figure 5.5 show the original signal and the bandpass filtered one, respectively, of a single utterance from a single speaker from our speech cycling corpus.

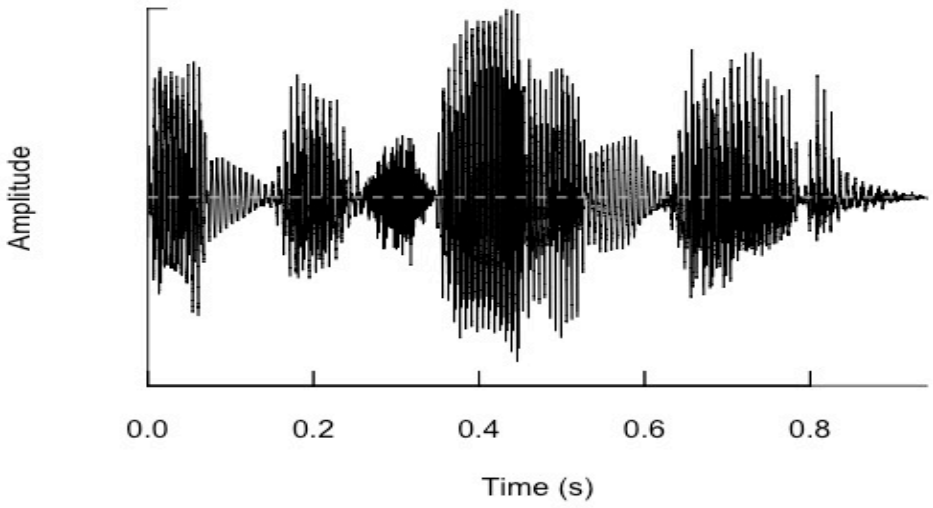


Figure 5.4: Original waveform.

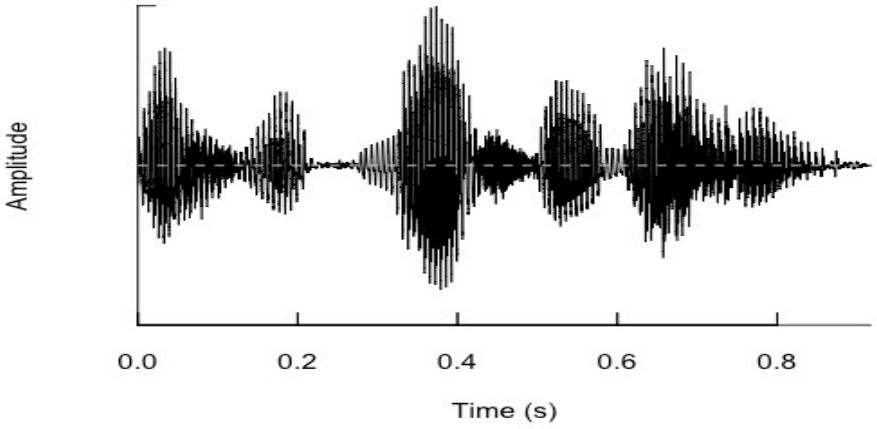


Figure 5.5: Waveform of bandpass filtered signal from 500 Hz to 4000 Hz.

Next, to obtain the envelope which represents smooth alternations between consonants and vowels, we obtain the absolute magnitude of the bandpass filtered signal, and low pass filter it with a 12 Hz cut-off. The cut-off frequency at 12 Hz means that the shortest syllables represented in the envelope have a duration of 83 ms, as this was the lowest bound of

syllables duration in our corpus. After low pass filtering at 12 Hz, there is significant redundancy in the signal; thus, we downsample from 16000 Hz to 100 Hz. This changes the temporal resolution from 0.0000625s to 0.01s and is useful to minimize the computation time needed for different envelope analyses. We also tapered the envelope with a Tukey window (r = 0.1) in order to mitigate edge artefacts from being represented, which usually emerge as a result of the filtering procedure. Figure 5.6 represents the amplitude envelope, and Figures 5.7 and 5.9 show the amplitude envelope superimposed over the bandpass-filtered signal and the absolute magnitude of the bandpass-filtered signal, respectively.

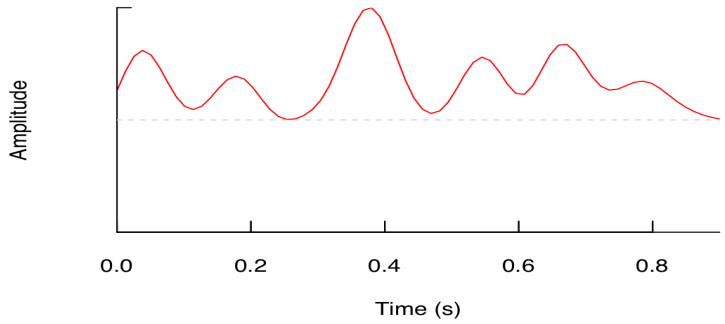


Figure 5. 6: Amplitude envelope low pass filtered at 12 Hz.

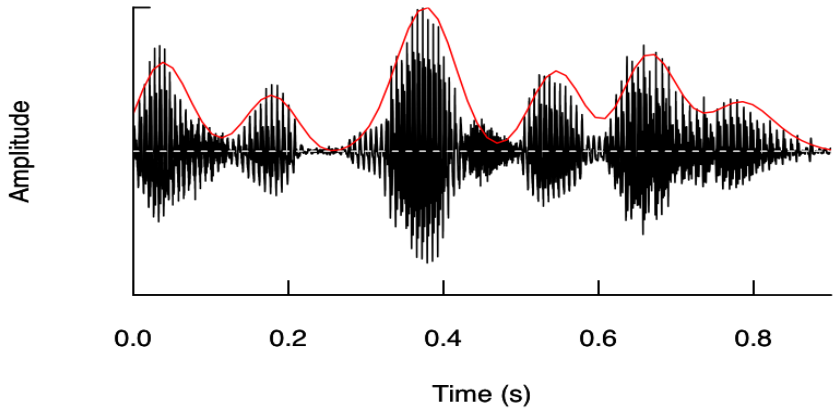


Figure 5.7: The amplitude envelope superimposed over the bandpass-filtered signal.

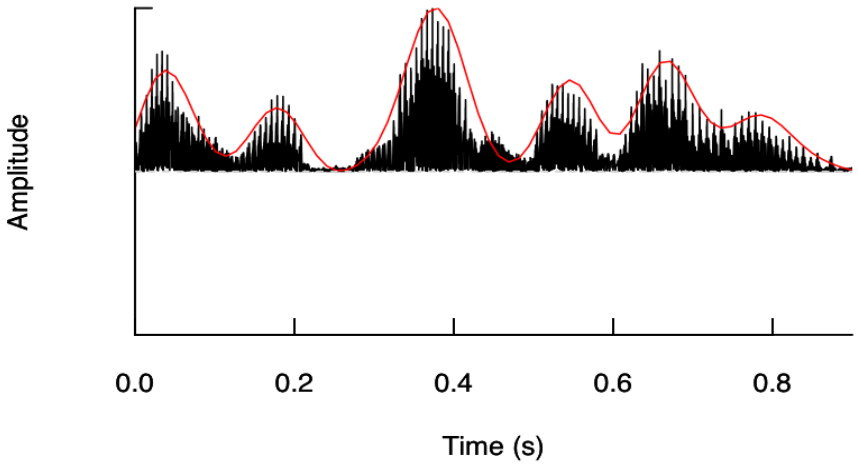


Figure 5.8: *The amplitude envelope superimposed over the absolute magnitude of the bandpass-filtered signal.*

The amplitude envelope is then normalized through mean subtraction and then rescaled so that it varies between -1 and 1 by dividing the data points of the envelope by the maximum value.

5.2.2 Obtaining power spectrum of the amplitude envelope

After obtaining the amplitude envelope in the time domain, we now obtain the envelope frequency components with different amplitudes through a Fourier transform. Recall that our aim from taking frequency representation of the amplitude envelope is to obtain frequency domain information related to foot level and syllable level time scales. We used the R package *stats* (R Core Team, 2019) to compute the frequency domain representation of the amplitude envelope.

We obtained the Fourier transform through the function *Spectrum* of the R package *stats*. Before performing the Fourier transform, we padded the envelope with zeros in order to aid the spectral analysis with greater frequency resolution. The magnitude of the Fourier transform was squared to obtain a power spectrum. Squaring the magnitude of the Fourier transform is important so that energy variation of the time domain signal is preserved in the frequency domain. This follows the Parseval's theorem, which states that the sum of the

square of the time domain signal is equal to the sum of the square of its transform (Bloomfield, 2000). Figure 5.9 shows the power spectrum of the normalized amplitude envelope.

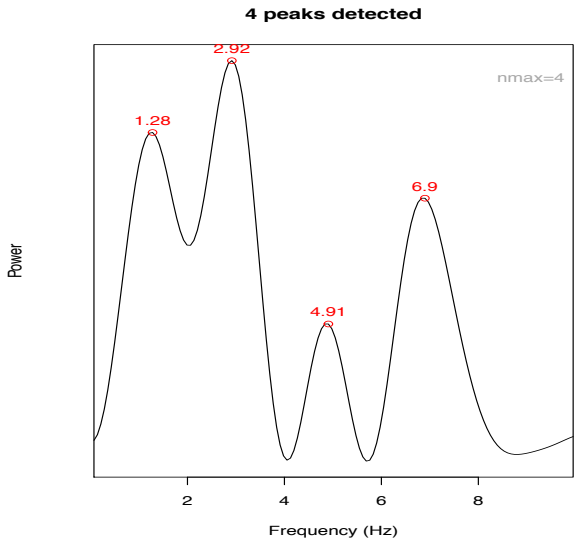


Figure 5.9: Power spectrum of normalised amplitude envelope. Red dots indicate different frequency peaks.

We can see in Figure 5.9 four frequency peaks represented in the power spectrum. There is a peak at 1.28 Hz, which probably corresponds to phrasal accent time scale (781 ms). The stress feet duration in our corpus ranges from 240 ms to 412 ms. This corresponds to frequencies that range from ~ 2.5 Hz to ~4 Hz. Thus, from the power spectrum in Figure 9, a peak at 2.92 Hz corresponds to the stress feet time scale at around 342 ms. As for syllabic time scales, they range in our corpus from 80 ms to 222 ms, corresponding to frequencies from 4.5 Hz to 12 Hz. Thus, peaks at 4.91 Hz and 6.9 Hz correspond to syllable level time scales at 203 ms and 144 ms, respectively. Since the power spectrum provides meaningful information regarding the occurrences of stress feet and syllables, multiple metrics can be extracted from the power spectrum that could quantify the dominance of either stress feet or syllables in the signal. The first is spectral band power ratio (SBPr). It is computed by dividing the sum of the power at frequency ranges corresponding to stress feet (2.5 Hz to 4 Hz) by the sum of the power at frequency ranges corresponding to syllable level (4.5 Hz to 12 Hz). As such, SBPr quantifies the power concentration at the frequency range of stress feet. The higher the SBPr value, the more power concentration at the stress feet frequency range. The second metrics that can be extracted is the Centroid, which is the spectral centre of gravity. Centroid is computed by taking the sum of all frequencies from 2.5 Hz to 12 Hz

multiplied by their powers, then dividing by the sum of powers. The Centroid in Hz shows at which frequency the power is concentrated. An advantage of the Centroid over SBPr is that the former is not dependent on the arbitrary division of the spectrum to certain bands that may be motivated by theoretical accounts. A Centroid value at, for example, 3 Hz would indicate that power is concentrated at the foot level time scale, while a Centroid value at 5 Hz would indicate power concentration at the syllable level time scale.

5.2.3 Empirical Mode Decomposition of the amplitude envelope

The Fourier transform is useful in providing information regarding global frequency components in the speech signal. However, one of the drawbacks of the use of the Fourier transform is that it does not provide information regarding the change of frequencies and amplitudes with time. Put more formally, the fast Fourier transform holds the assumption that the data is stationary. Stationarity refers to the stability of the data statistics, such as the mean and the variance. For example, when the speech signal is represented in the Fourier spectrum with N frequencies and amplitudes, it assumes that the signal has frequencies and amplitudes that do not change throughout time. This is, however, not the case for speech signal or for most physical processes, as their frequencies and associated amplitudes vary in the time domain. Thus, it is important to account for changes in frequency and amplitude throughout time. We will demonstrate here Empirical Mode Decomposition (EMD) (Huang et al., 1998), a method developed to deal with non-stationary data, which is an essential signal processing step to account for changes in amplitude and frequency of time scales representing syllables and stress feet.

The EMD method assumes that any non-stationary/non-linear time series data is made up of simple oscillatory modes, which are called intrinsic mode functions. Therefore, the result of the EMD algorithm is a set of intrinsic mode functions (IMFs) that represent simple oscillations in the signal at different time scales. These IMFs can be assigned instantaneous frequencies and amplitudes that describe the amount of change in time. An example simple oscillation in the signal can be seen in Figure 5.10, where the blue dotted line represents a full cycle of a simple oscillation that starts with a local maximum and ends at the following local maximum, crossing two zeros and passing a local minimum. Huang et al. (1998) provide two conditions for a signal to be considered an intrinsic mode function. First, the number of local extrema and zero crossing points should be equal or differ by one. Second, the local mean of

maximum and minimum points at any given time point should be zero. Signals that do not meet these conditions are considered complex waves and do not qualify as intrinsic mode functions.

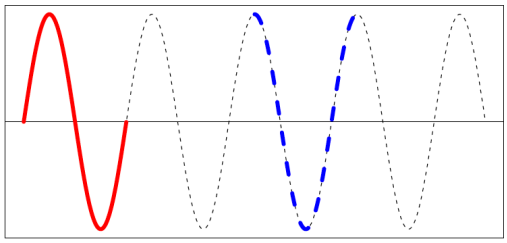


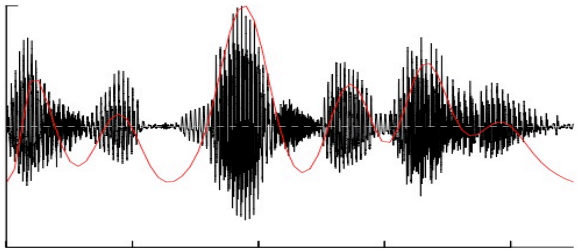
Figure 5.10: A signal with simple oscillations. The red line indicates the minimum period of an oscillation, which starts with a maximum and ends with a minimum, crossing one zero. The blue dotted line shows a full cycle of a simple oscillation, which starts with a maximum and ends with the following maximum, crossing two zeros and a local minimum. (Kim & Oh, 2008, p. 40).

The EMD works as an iterative filter: it first calculates the intrinsic mode with the highest frequency, then this first IMF is subtracted from the original signal, producing a residual, which then acts as a new signal, and the EMD is applied to it again to extract the next highest frequency IMF. The algorithm repeats until there are no oscillations left in the residual. The original signal can be reconstructed again by summing all extracted IMFs and the residual.

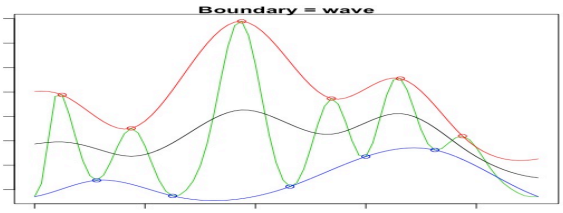
The extraction of IMFs is made through a process called sifting. The process of sifting starts first by identifying local maxima and minima in the signal, and then an upper envelope is created by interpolating the local maxima with a cubic spline line, and another lower envelope is created for the local minima by the same procedure. Second, the difference between the original signal and the average of the upper and lower envelopes is calculated by subtracting the average from the signal, producing the first sift ($sift_1$). If $sift_1$ does not meet the two conditions of intrinsic mode functions, another sifting process is applied to $sift_1$ and $sift_2$ will be assessed if it meets intrinsic mode functions conditions. The sifting process repeats until the first IMF is produced and then subtracted from the original signal, producing a residual, from which the second IMF is extracted.

We used R package *EMD* (Kim & Oh, 2008) to compute IMFs of the amplitude envelope to represent simple oscillations of syllables and stress feet time scales. The first IMF with the fastest oscillation is said to correspond to syllables, while the second fast IMF is said to correspond to stress feet (Tilsen & Arvaniti, 2013). The sifting process to obtain IMF1 is exemplified in Figure 5.11, with the normalised amplitude envelope of an utterance from our speech cycling corpus acting as the original signal. Note that if the sifting process continues to the extreme, we will end up with frequency modulated signal with constant amplitude. Thus, there has to be a stopping rule to the sifting process to retain the physical attributes of amplitude and frequency modulation of the signal. The stopping rule was set in a way that the standard deviation of amplitudes of consecutive sifts, e.g., $Sift_1$ and $Sift_2$ is within 0.2 and 0.3 range. As can be seen from Figure 5.11, the amplitude envelope goes through the first iteration process to obtain the average envelope maxima and minima, and the average envelope is subtracted from the original signal, i.e., the amplitude envelope, to produce $Sift_1$. Iteration continues on $Sift_1$ since it does not meet IMF conditions, until IMF conditions are satisfied. Figure 5.12 shows IMF1 and IMF2 together with the original signal. IMF1 has six peaks, thus it is likely that it corresponds to the syllables time scale, and IMF2 has three peaks which makes it plausible to represent the stress foot time scale.

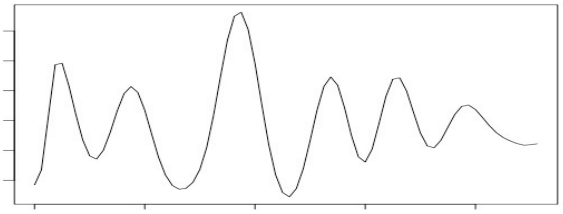
As pointed out earlier, the combination of IMFs with the residue can be used to reconstruct the signal. Thus, IMFs have consistent physical attributes of the amplitude envelope and contain energy that belongs to the amplitude envelope at different time scales. Tilsen and Arvaniti (2013) pointed out that this property of IMFs may be used to quantify the dominance of stress feet relative to syllables. By obtaining the power spectrum of IMF1 and IMF2, the ratio of the sum of powers of IMF2 to IMF1 may be computed (Ratio21). This measure quantifies the energy contribution to the signal by stress foot oscillation relative to syllable level oscillation. The higher the ratio, the greater the energy contribution by stress foot oscillation to the signal.



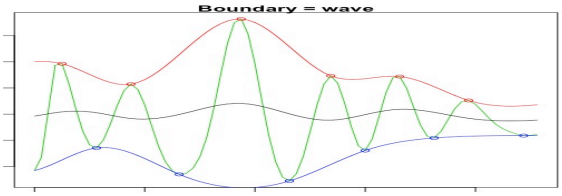
(a)Normalized envelope over band passed signal



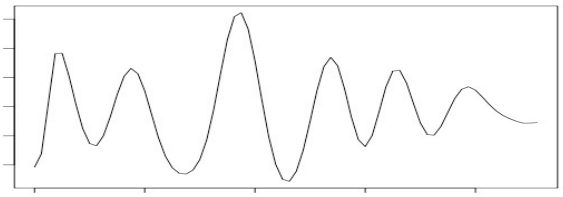
(b)First iteration



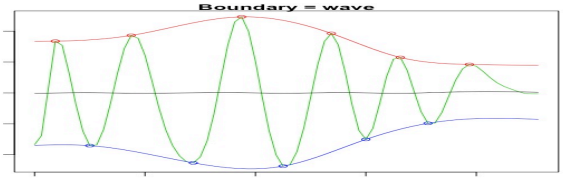
(c)Sift$_1$



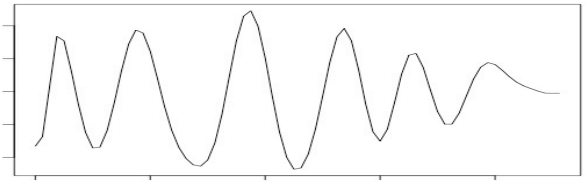
(d)Second iteration



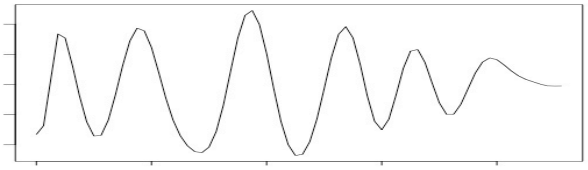
(e)Sift$_2$



(f)Third iteration



(g)Sift$_3$



(h)IMF1

Figure 5.11: Sifting process on the amplitude envelope shown in red in (a). In (b) first iteration to obtain the average of maxima and minima shown in black. Sift$_1$ in (c) is obtained after subtracting the average envelope from the original signal. In (d) second iteration is applied on Sift$_1$. (f) shows the final iteration where it is clear that the average of maxima and minima is zero. Thus, Sift$_3$ in (g) matches the conditions of an IMF, and can be regarded as the first IMF as in (h).

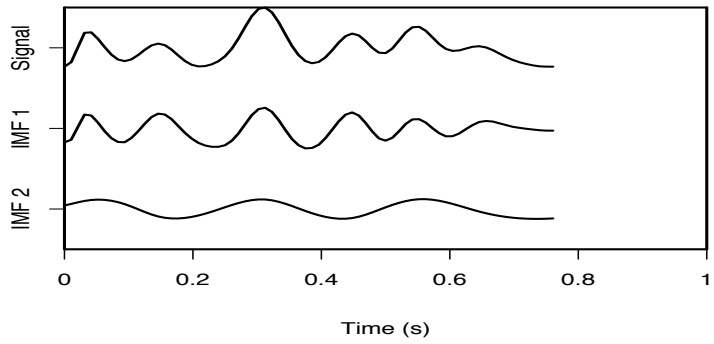


Figure 5.12: IMF1 and IMF2 extracted from the original signal, the amplitude envelope. It can be seen that there are six peaks in IMF1 representing syllable level oscillation and three peaks in IMF2 representing stress foot level oscillation.

5.2.4 The Hilbert-Huang transform

After obtaining IMF1 and IMF2, the next step is to compute changes in frequency and amplitude in time of IMF1 and IMF2, which should reflect changes in syllable level and stress foot level oscillation. The extracted simple oscillations in the signal, i.e., IMF1 and IMF2, have meaningful instantaneous frequency, and amplitude in the Hilbert transform. Through the Hilbert transform, the instantaneous phase is first computed, and then the time derivative of the instantaneous phase is computed to obtain the instantaneous frequency of IMFs. The instantaneous amplitude is the absolute magnitude of the Hilbert transform. We used R package *EMD* (Kim & Oh, 2009) to compute the Hilbert transform of IMFs and package *hht* (Bowman & Lees, 2013) to plot the Hilbert transform. Figure 5.13 plots the change in frequency and amplitude in the time domain of IMF1 and IMF2. The spectrogram used to represent the instantaneous frequency and amplitude of intrinsic mode functions is called the Hilbert-Huang transform. In Figure 5.13, the y-axis shows the frequency levels of IMF1 and IMF2, the x-axis represents the change throughout time, and the colour change represents the change in amplitude. IMF1 exhibits greater changes in instantaneous frequency throughout time than IMF2. Since IMF1 corresponds to the syllable level time scale and IMF2 corresponds to the stress foot time scale, then the greater changes in instantaneous frequency in IMF1 means greater changes at syllable level oscillation than at stress foot oscillations.

Thus, a useful metric that may be obtained to assess the stability of syllable level and stress foot level oscillation is the variance in instantaneous frequency of IMF1 and IMF2 (vIMF1, and vIMF2, respectively). The rate of oscillation of IMFs may also be inferred by taking the average of the instantaneous frequency of IMFs. For instance, the rate of oscillation of IMF1, i.e., syllable level oscillation, in the example provided in Figure 5.13, is faster than the rate of IMF2, i.e., stress foot oscillation, at 6.1 Hz (164 ms) and 3 Hz (333 ms), respectively. Thus, another set of metrics that may be obtained from the Hilbert-Huang transform is rate metrics, i.e., the average instantaneous frequency of IMF1 and IMF2 (mIMF1, and mIMF2, respectively).

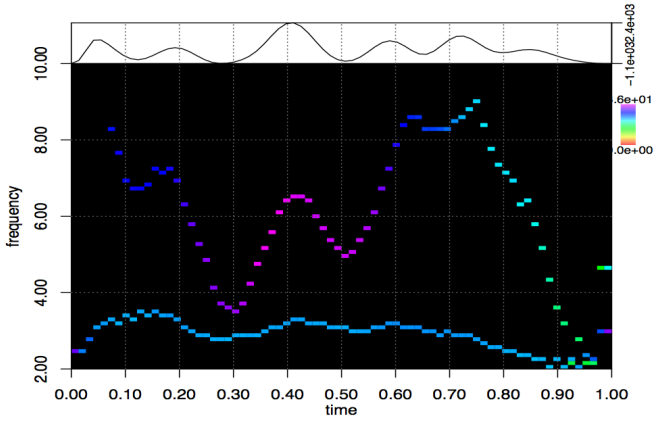


Figure 5.13: A spectrogram showing the instantaneous frequency and amplitude of IMF1 and IMF2. The y-axis represents frequency in Hz, and the x-axis represents time in seconds. Variation in colour indicates variation in amplitude. IMF1 is represented with higher frequencies and greater throughout time, while IMF2 is represented with lower and lesser changes in frequencies. The original amplitude envelope is shown on the top panel in black.

5.2.5 Summary of statistical metrics

Table 5.1 summarises statistical metrics from different processes of the amplitude envelope, which may be grouped into three different sets of metrics based on the concept they convey: rate metrics, power distribution metrics, and rhythmic stability metrics. Rate metrics, mIMF1 and mIMF2, simply describe the rate of syllable level and stress foot level oscillation. Power distribution metrics, Ratio21, SBPr, and Centroid describe the relative dominance of either stress foot level power or syllable level power. Rhythmic stability metrics, vIMF1, and vIMF2 describe the relative stability of syllable level oscillation or stress foot level oscillation, respectively. Less variance at syllable level oscillation or stress foot level

oscillation would indicate greater dominance of syllable level or stress foot level oscillation in the signal.

Table 5.3: Statistical metrics obtained from the processing of the amplitude envelope. Adapted from Tilsen and Arvaniti (2013, p. 634).

Type	*Metric*	*Description*	*Interpretation*
Rate metrics	mIMF1	Mean instantaneous frequency of IMF1	Rate of syllable level oscillation
	mIMF2	Mean instantaneous frequency of IMF1	Rate of stress foot level oscillation
Power distribution metrics	Ratio21	Ratio of the power of IMF2 to the power of IMF1	Amount of power in stress foot oscillation relative to syllable level
	SBPr	Ratio between spectral bands powers: from 2.5 Hz to 4 Hz and from 4.5 Hz to 12 Hz	Amount of power in stress foot level spectral band relative to syllable level band
	Centroid	Power spectrum centroid computed over the range from 2.5 Hz to 12 Hz	Power concentration at either stress foot level or syllable level
Rhythmic stability metrics	vIMF1	Variance of instantaneous frequency of IMF1	Stability of syllable level oscillation
	vIMF2	Variance of instantaneous frequency of IMF2	Stability of stress foot level oscillation

5.3 Predictions

We have described how power distribution metrics and rhythm-stability metrics may correspond to the dominance of either syllables or stress feet time scales in the signal. In our speech cycling corpus, there is a regular alternation between stressed and unstressed syllables, thus there is a regular occurrence of stress feet. On the other hand, there is more variability in syllable structure. Accordingly, we expect greater power concentration at the stress feet time scale and greater stability, i.e., less variance, in stress feet instantaneous frequency. Regular occurrence of stress feet is similar in Hadari and Bedouin dialects, thus we do not expect categorical differences. For example, in Centroid, we do not expect Hadari to exhibit power concentration at stress feet range (2.5 Hz to 4 Hz) and Bedouin to exhibit power concentration at syllables range (4.5 Hz to 12 Hz). However, dialectal differences in syllable duration would affect the degree of stress feet dominance, i.e., the degree of regularity of stress feet. From our analyses of syllabic durational profiles in Experiment 1 (b), two factors that are predicted to influence the degree of stress feet dominance between Hadari and Bedouin dialects. They are metronome period and trochaic and iambic stress patterns.

As for metronome period effects, we have already seen when analysing syllabic durations that Hadari exhibited greater unstressed syllable reduction than Bedouin across metronome periods in phrase-initial position. Greater unstressed syllable reduction in Hadari could minimize the temporal differences between stress feet across metronome periods in Hadari. Thus, we predict that Hadari would exhibit less changes in power distribution at stress feet time across different metronome periods than Bedouin, and Hadari would exhibit less variance in stress feet instantaneous frequency across different metronome periods.

The potential effect of trochaic and iambic patterns on the degree of foot time scale dominance between dialects relates to the different syllable structure between the two sets of sentences and the tolerance of unstressed syllable reduction between the two dialects. The syllable structure of unstressed syllables in the iambic sentences, CV, is simpler than in the trochaic sentences, CVC. Since Hadari tends to exhibit greater unstressed syllable reduction than Bedouin, the former dialect may exhibit greater reduction of complex unstressed syllables, CVC, in trochaic sentences, which may minimize the temporal differences between stress feet in the trochaic and iambic sentences. In particular, we found in Experiment 1 (b) that Hadari tended to reduce unstressed syllables to a greater degree than Bedouin, most notably, in phrase-initial position and at the shortest metronome period. A post-hoc examination shows that Hadari tends to have shorter unstressed syllables in the trochaic sentences and in the iambic sentences, 147 ms vs 104 ms, than Bedouin, 170 ms vs 114 ms, in phrase-initial position at the shortest metronome period. Greater unstressed syllable reduction in Hadari leads to smaller differences between unstressed syllables across trochaic sentences and iambic sentences, 147 ms – 104 ms = 43 ms, than in Bedouin, 170 ms – 114 ms = 56 ms. We predict that smaller differences between unstressed syllables across the trochaic and iambic sentences would lead to relatively more regular stress feet durations across trochaic and iambic sentences in Hadari than in Bedouin. Thus, power distribution metrics are predicted to be more similar across iambic and trochaic sentences in Hadari than in Bedouin, and there would be less variability in the instantaneous frequency of stress feet time scale across iambic and trochaic sentences in Hadari than in Bedouin.

5.4 Analysis

Each metric demonstrated in Table 5.1, was the dependent variable in a linear regression mixed-effects model. Predictors in each model were dialect (Bedouin and Hadari), stress

pattern (iambic and trochaic) and metronome period which was included as a continuous predictor.

Two-way interactions between dialect and metronome period and dialect and stress pattern were included in the models, as well as three-way interaction between dialect, stress pattern and metronome period.

The random structure in all models included speaker and sentence as random intercepts. Random slopes for metronome period and stress pattern by speaker, as well as random slopes for metronome period and dialect by sentence were included in the model.

All predictors were centred and likelihood ratio tests were conducted for significance testing. Models and significance testing were done through the use of package *afex* in R software. Pairwise comparisons for interactions levels were conducted through package *phia.*

5.5 Results

5.5.1 Rate metrics

We will start by reporting results from models that included rate metrics, mIMF1 and mIMF2 as the dependent variables. Recall that mIMF1 corresponds to the oscillation rate of syllables and mIMF2 corresponds to the oscillation rate of stress feet.

5.5.1.1 mIMF1 metric

The intercept value, which corresponds to the mean of all predictors is 6.4 Hz. This corresponds to a mean rate of 156 ms of syllabic intervals. Note that the intercept value does not correspond precisely to the mean duration of syllables, extracted from our data, Experiment 1(b), which was 182 ms. This is potentially due to a general shortcoming of the interpolation process in the EMD algorithm, which may include ultra-frequency and low amplitude data points in the IMF, thus distorts the physical representation of the original signal (Huang et al., 1998, p. 921). We will discuss later potential improvements in the computation of IMFs to capture the actual syllabic durational patterns.

Figure 5.14 shows dialectal differences in syllable oscillation rate. There was no effect for dialect on syllabic oscillation rate, $X^2(1) = 1.44$, $p = .2$.

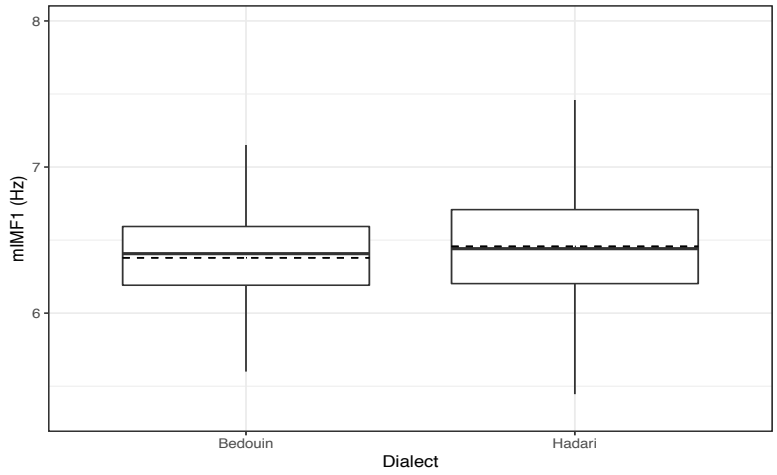


Figure 5.14: Dialectal differences in rate of syllabic oscillation.

Figure 5.15 illustrates the difference in syllabic oscillation rate between iambic and trochaic patterns, where the iambic pattern has higher oscillation rate than the trochaic pattern. There was a significant effect for stress pattern, $X^2(1) = 5.27$, $p = .02$. For the iambic pattern, β = 0.17 Hz, and SE = 0.06 Hz, with the trochaic pattern as the reference level. As β represents the change around the intercept, predictions for the iambic pattern is 6.4+0.17 = **6.57 Hz** (152 ms), and for the trochaic pattern 6.4+(-0.17) = **6.23 Hz** (158 ms). Thus, the differences in mIMF1 between the iambic and trochaic sentences in mIMF1 are very small.

Figure 5.16 shows metronome period effect on syllabic oscillation rate. Not surprisingly, syllabic oscillation rate is faster at shorter metronome periods. There was a significant effect for metronome period on syllable rate, $X^2(1) = 4.35$, $p = .03$, with β = 0.07 Hz, and SE = 0.03 Hz. The slope, β, represents the change around the intercept at the shortest metronome period, which was assigned a value of 1 after treating metronome period as a continuous variable. Thus, predictions for the shortest metronome period is, 6.4+0.07 * 1 = **6.47 Hz** (154 ms), at the medium period, 6.4+0.07 * 0 = **6.42 Hz** (156 ms) and at the longest period, 6.4+0.07 * -1 = **6.34 Hz** (158 ms).

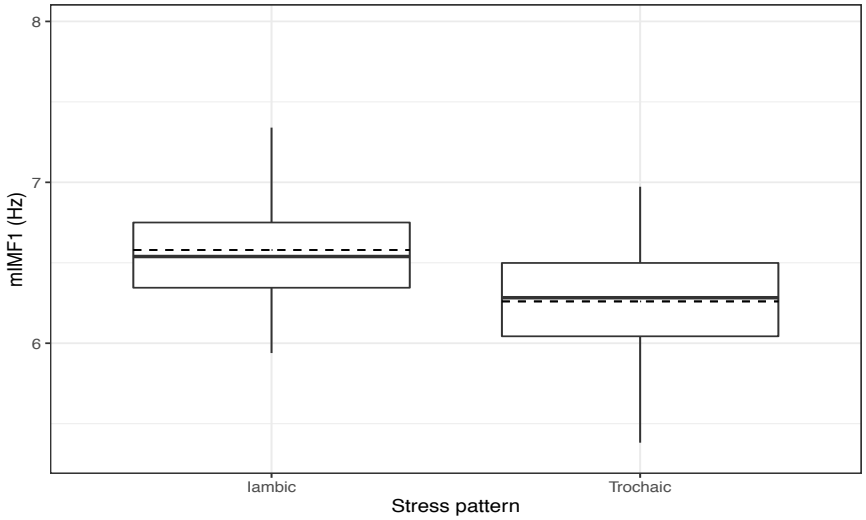


Figure 5.15: Syllabic oscillation rate in iambic and trochaic sentences.

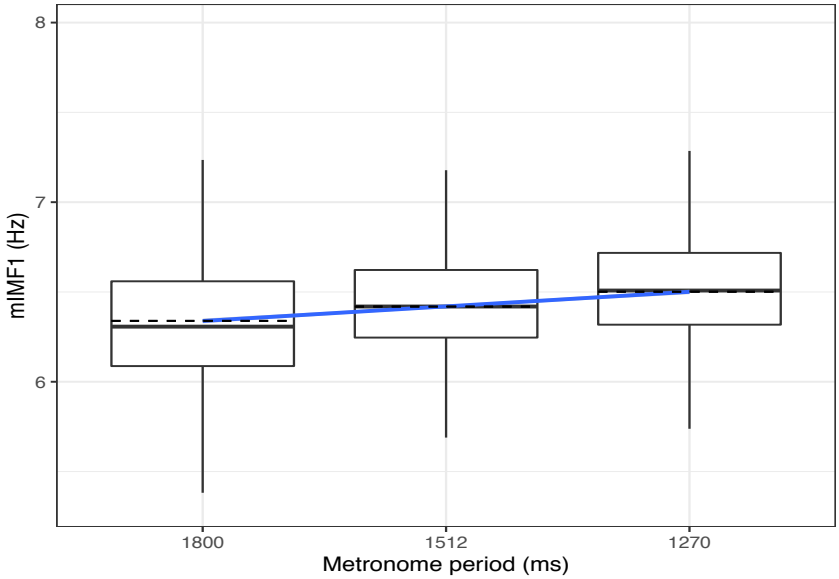


Figure 5.16: *Metronome period effect on syllabic oscillation rate.*

The two significant main effects simply describe the rate of syllabic oscillation, which was faster at the iambic sentences than in the trochaic sentences, and faster at shorter metronome periods. However, note that the difference between the iambic and trochaic patterns, and the difference between longer and shorter metronome periods in syllables' rate did not reach the perceptual threshold for rate differences, 5% (Quenè, 2007). As mentioned, this might be due to the distorted representation of the signal caused by the interpolation process in the EMD algorithm.

There was no two-way interaction between dialect and stress pattern, $X^2(1) = 1.45$, $p = .2$, or between dialect and metronome period, $X^2(1) = 2.54$, $p = .1$, or between metronome period and stress pattern, $X^2(1) = 3.16$, $p = .07$. There was a significant three-way interaction between dialect, stress pattern and metronome period, $X^2(5.17) = 11.08$, $p = .02$. We used R function *predict* from package *stats* (R Core Team, 2013) to generate predictions of the interaction levels. Figure 5.17 plots the interaction.

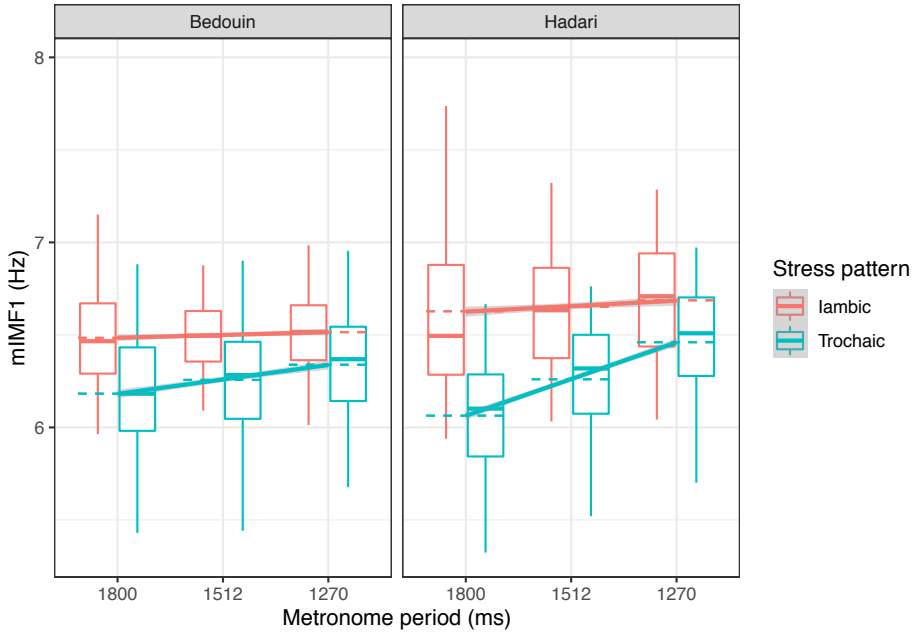


Figure 5.17: The effect of three-way interaction between dialect, stress pattern and metronome period on syllabic oscillation rate.

As can be seen from Figure 5.17, the largest difference between dialects in syllabic oscillation rate across iambic and trochaic sentences is at the longest metronome period. Post-hoc pairwise comparison showed no differences in syllabic oscillation rate between iambic and trochaic pattern at the longest metronome period in Bedouin, $p = .08$, whereas in Hadari there were significant differences, $p = .03$, with the trochaic sentences tending to have slower syllabic oscillation rate, **6.1 Hz** (164 ms), than the iambic sentences, **6.6 Hz** (151 ms), and the difference in syllables' rate is above the perceptual threshold for rate differences, > 5%. This suggests that syllable duration in Hadari between the iambic and the trochaic patterns in the longest metronome period is more variable than in Bedouin.

5.5.1.2 mIMF2 results

Recall that mIMF2 corresponds to the rate of stress feet oscillation. The intercept value of the model is 3.23 Hz, which corresponds to a duration of 310 ms. This rate does not correspond precisely to the mean foot duration in our experimental materials, which is 335 ms. As explained earlier, the distorted representation of the original signal in the computation of IMFs can be due to the interpolation process involved in the EMD algorithm, which can include ultra-frequency data points.

Figure 5.18 shows dialectal differences in stress feet oscillation rate. There was no effect of dialect on mIMF2, $X^2(1) = 0.16$, $p = .6$.

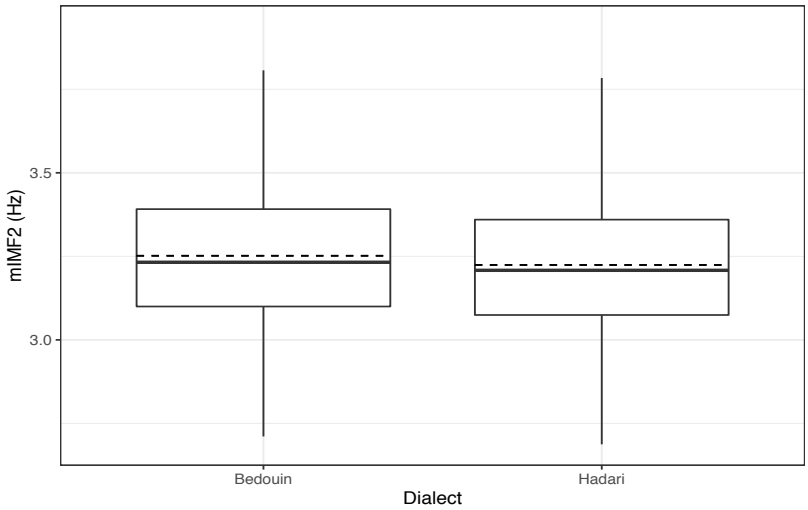


Figure 5.18: The rate of stress feet oscillation in Bedouin and Hadari.

Figure 5.19 shows differences between trochaic and iambic sentences in mIMF2. There was no effect of stress pattern on mIMF2, $X^2(1) = 1.2$, $p = .2$.

Figure 5.20 shows the effect of metronome period on mIMF2. It can be seen that, as expected, the rate of stress feet oscillation is faster at shorter metronome periods. There was a significant effect of metronome period on the rate of stress feet oscillation, $X^2(1) = 22.66$, $p = .001$, with $\beta = 0.1$ Hz, and SE = 0.002 Hz. The slope, β, represents the change around the intercept at the shortest metronome period, which was assigned a value of 1 after treating metronome period as continuous variable. Thus, prediction for the shortest metronome period

is 3.23+0.1 = **3.33 Hz** (300 ms), for the medium period, 3.23+0.1 * 0 = **3.24 Hz** (308 ms), and for the longest period, 3.23+0.1 * -1 = **3.13 Hz** (320 ms).

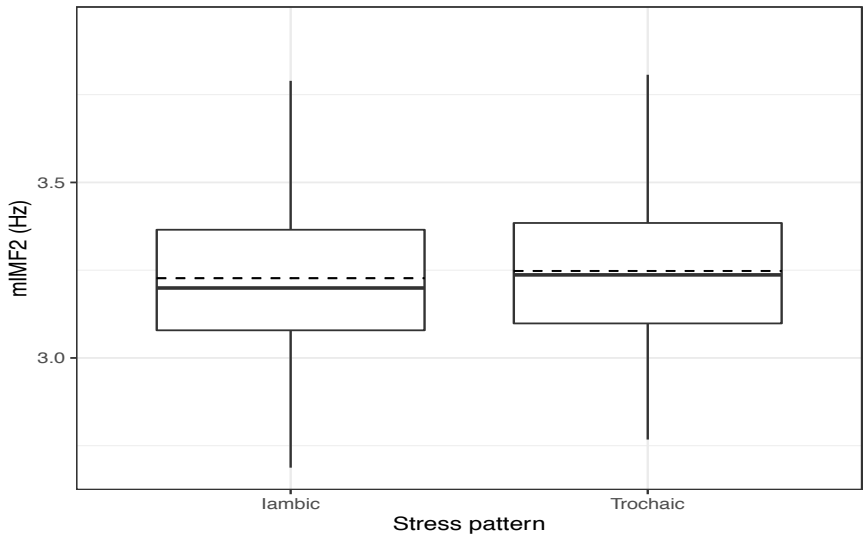


Figure 5.19: The rate of stress feet oscillation in iambic and trochaic sentences

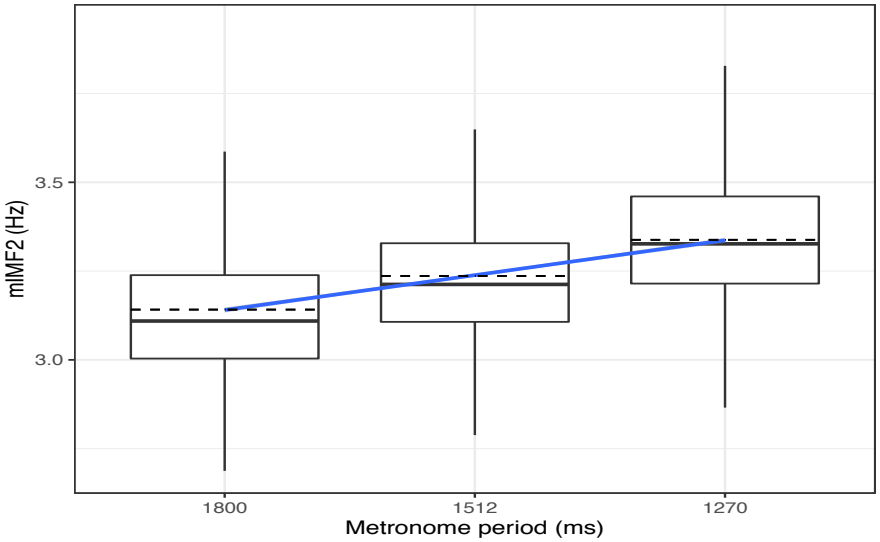


Figure 5.20: The effect of metronome period on stress feet oscillation rate.

Only a single main effect had a significant effect on stress feet oscillation rate, that is metronome period. Not surprisingly, stress feet oscillation rate is faster at shorter metronome periods, reflecting compression in stress feet duration at shorter metronome periods. Note, however, that the only difference that exceeded the perceptual threshold for rate difference is that between the longest and the shortest metronome periods, > 5%.

There was no two-way interaction between dialect and stress pattern, $X^2(1) = 1.17$, $p = .2$. There was a significant two-way interaction between dialect and metronome period, $X^2(1) = 5.0$, $p = .02$. Figure 5.21 plots the interaction. As it is illustrated in Figure 21, the largest difference in mIMF2 between dialects is at the longest metronome period. Hadari exhibited slower stress feet oscillation than Bedouin at 3.10 Hz (322 ms) and 3.18 Hz (314 ms) respectively. Pairwise comparison confirmed this statistical trend, $p = .04$. However, the difference in the oscillation rate of stress feet between Hadari and Bedouin at the slowest metronome period does exceed the perceptual threshold for rate differences, < 5%.

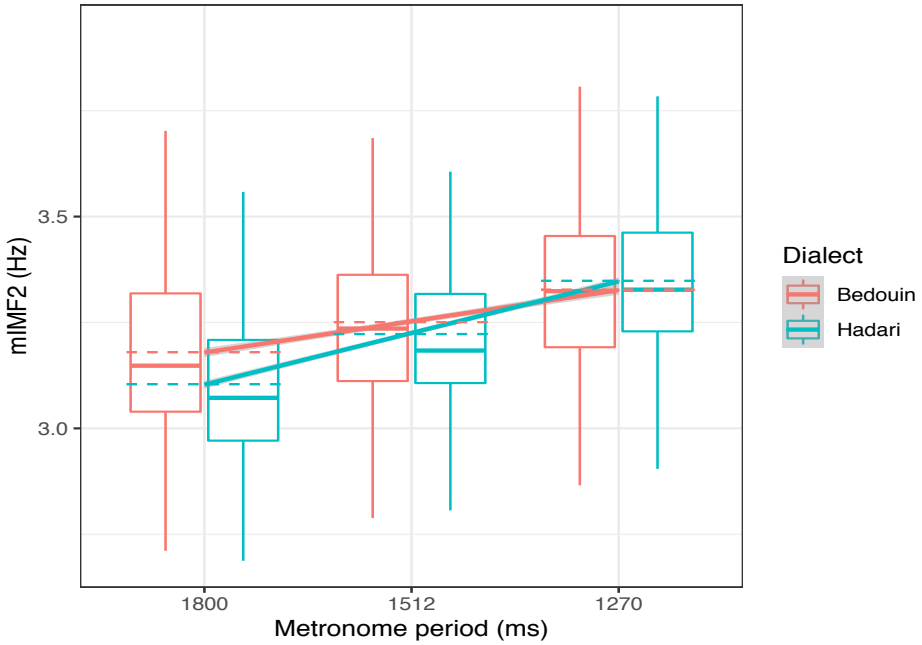


Figure 5.21: The effect of two-way interaction between dialect and metronome period on stress feet oscillation rate.

There was no interaction between metronome period and stress pattern, $X^2(1) = 0.70$, $p = .4$, and there was no three-way interaction between dialect, stress pattern and metronome period, $X^2(1) = 1.5$, $p = .2$.

The analysis of mIMF1 and mIMF2 was meant to quantify the rate of syllabic and stress feet oscillation, respectively. A general shortcoming was that IMFs representation of the original acoustic signal was not accurate, as the means of mIMF1 and mIMF2 did not correspond precisely to the mean durations of syllables and stress feet in the original signal. We will discuss later possible improvements in the computation of IMFs. Beyond this shortcoming,

there was a dialectal difference as Hadari exhibited greater difference in mIMF1 between the iambic and trochaic sentences, with trochaic sentences having slower syllabic oscillation rate than the iambic sentences. Also, in mIMF2, Hadari showed slower stress feet oscillation rate than Bedouin at the longest metronome period, however, the difference in the oscillation rate did not exceed the perceptual threshold for rate differences. As mIMF1 and mIMF2 only reflect the rate of syllabic and stress feet oscillation, they are not related directly to our question regarding the relative dominance of stress feet time scale between Hadari and Bedouin. Our analysis of rhythm-stability metrics and power distribution metrics would be more closely related to answering our questions.

5.5.2 Rhythmic stability metrics

Recall that rhythm-stability metrics quantify the variability in syllabic and stress feet instantaneous frequency. Thus, for example, less variability in stress feet instantaneous frequency means greater dominance of the stress feet time scale relative to the syllabic time scale.

5.5.2.1 vIMF1 (stability of syllabic instantaneous frequency)

The model's intercept value is 2.83. Figure 5.22 plots dialectal differences in vIMF1. There was no effect for dialect, $X^2(1) = 2.3$, $p = .1$.

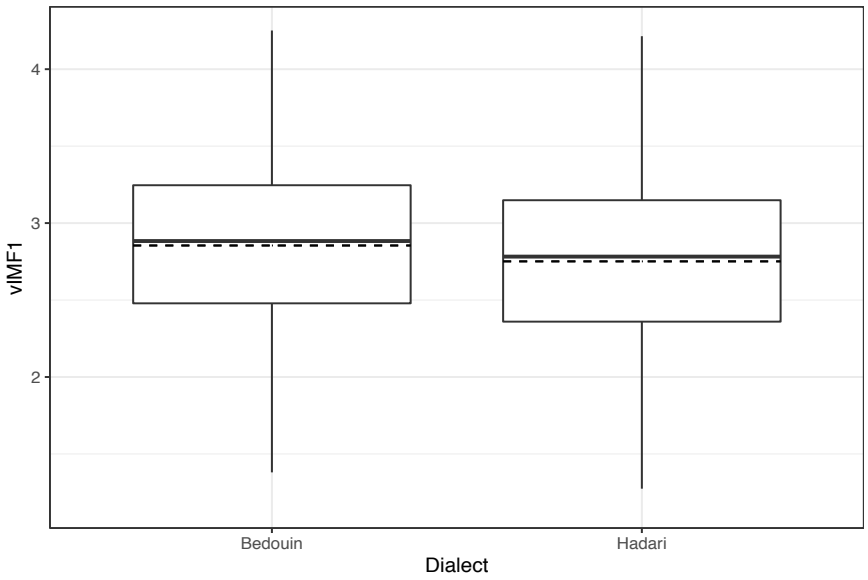


Figure 5.22: Dialectal differences in vIMF1.

Figure 5.23 plots the difference between iambic and trochaic sentences in vIMF1 values. There was no significant effect for stress pattern on vIMF1, $X^2(1) = 0.02$, $p = .8$.

Figure 5.24 shows the effect of metronome period on vIMF1. As can be seen, the variability of syllabic instantaneous frequency decreases at shorter metronome periods.

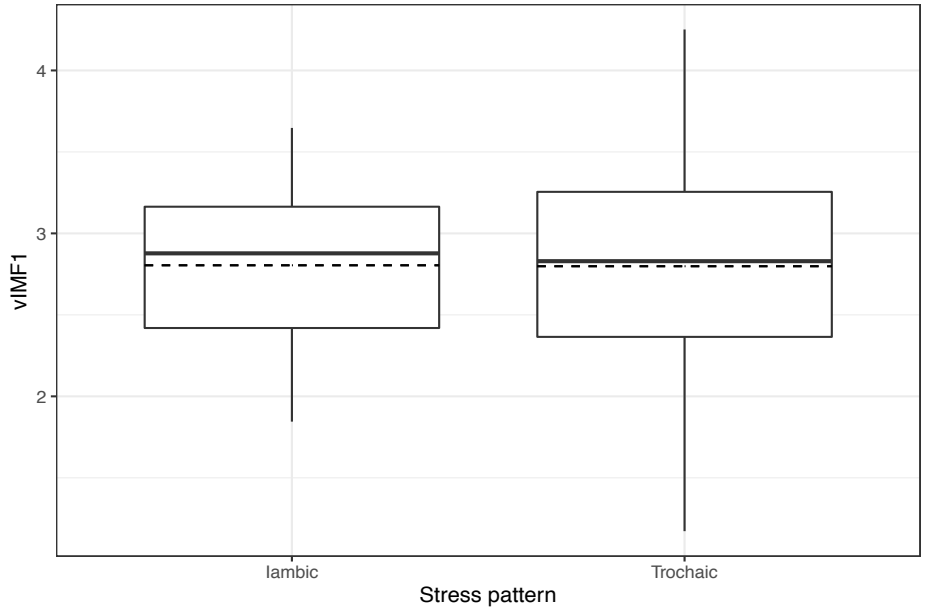


Figure 5.23: vIMF1 in the iambic and trochaic sentences.

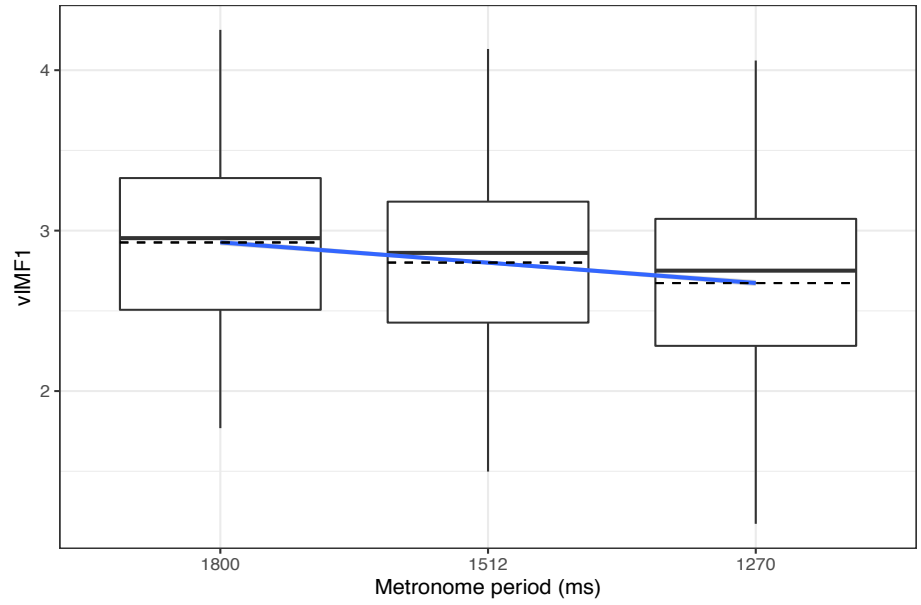


Figure 5.24: The effect of metronome period on vIMF1.

There was a significant effect of metronome period on vIMF1, $X^2(1) = 7.7$, $p = .006$, with β = -0.12, and SE = 0.04. The slope, β, represents the change around the intercept for the shortest metronome period, which was assigned of 1 after treating metronome period as a continuous variable. Thus, predictions for the shortest metronome period is 2.83+(-0.12) = **2.71**, for the medium period, 2.83+(-0.12) * 0 = **2.83**, and for the longest period, 2.83+(-0.12) * -1 = **2.95**. The decrease in vIMF1 at shorter periods reflect less variability in syllabic durational intervals at shorter metronome periods.

There was no two-way interaction between dialect and stress pattern, $X^2(1) = 0.15$, $p = .6$, or between dialect and metronome period, $X^2(1) = 0.1$, $p = .9$. There was a significant two-way interaction between stress pattern and metronome period, $X^2(1) = 7.2$, $p = .007$. Figure 5.25 plots the interaction. As can be seen from Figure 5.25, variance of syllabic instantaneous frequency decreases from longer to shorter metronome periods in the trochaic sentences, whereas it is similar across different metronome periods in the iambic sentences.

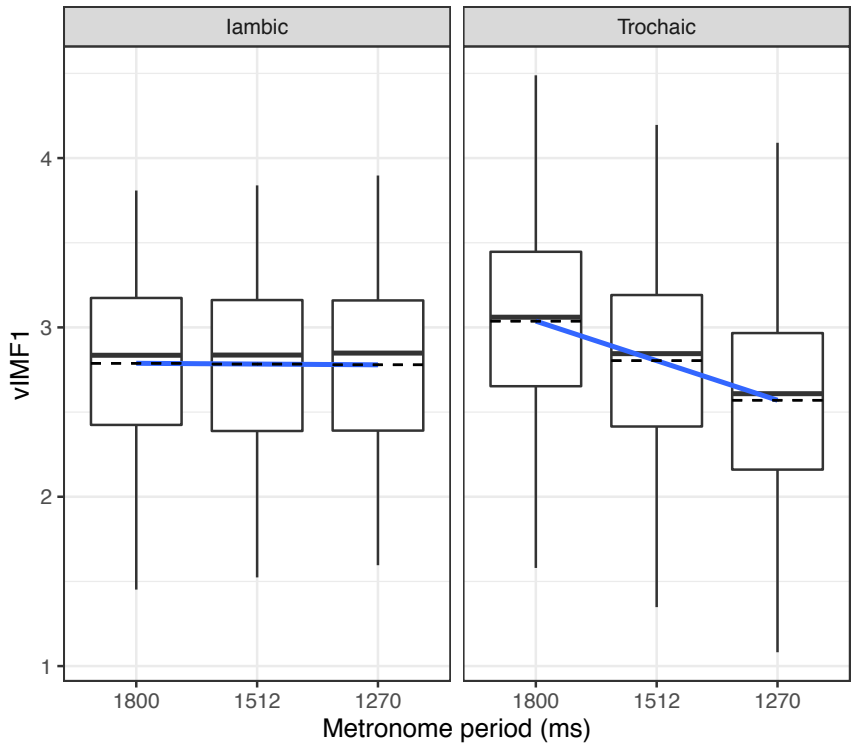


Figure 5.25: The effect of two-way interaction between stress pattern and metronome period on vIMF1.

Pairwise comparison showed significant differences in vIMF1 between the longest and the medium metronome periods in the trochaic sentences, $p < .0001$, and between the medium

and the shortest periods in the trochaic sentences, $p < .0001$, while there were no differences in vIMF1 across different metronome periods in the iambic sentences, in all comparisons $p > .05$. This indicates more variability in syllabic durational intervals from longest to shortest metronome periods in the trochaic sentences, which is also associated with greater syllabic compression. As unstressed syllables in the trochaic pattern are more complex, CVC, than in the iambic pattern, CV, unstressed syllables are more prone to greater compression effects in the trochaic sentences than in the iambic sentences from the longest to the shortest metronome periods.

There was no significant three-way interaction between dialect, stress pattern and rate, $X^2(1) = 0.3$, $p = .5$.

5.5.2.2 vIMF2 (variability of stress feet oscillation)

The model's intercept value of vIMF2 is 0.41. Figure 5.26 illustrates dialectal differences in vIMF2. There was no effect of dialect on vIMF2 values, $X^2(1) = 0.07$, $p = .7$.

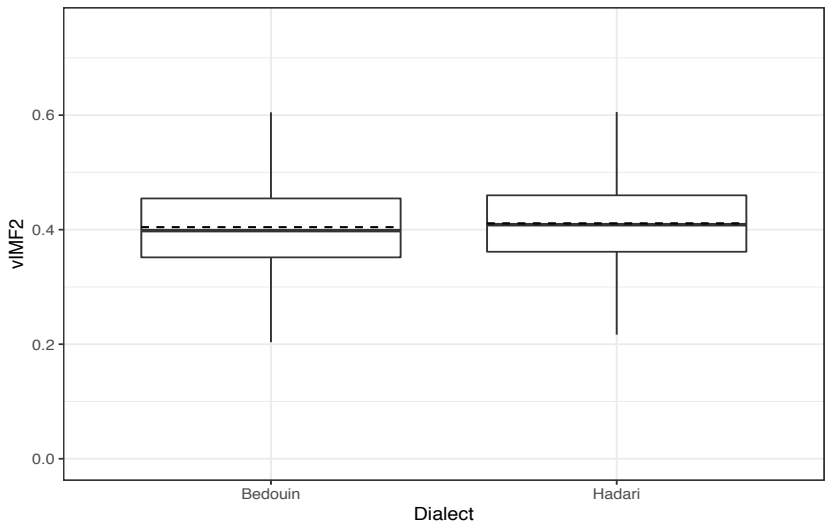


Figure 5.26: Dialectal differences in vIMF2.

Figure 5.27 shows mean vIMF2 values between iambic and trochaic sentences. There was no effect of stress pattern on vIMF2, $X^2(1) = 1.4$, $p = .2$.

Figure 5.28 shows the effect of metronome period on vIMF2. As can be seen, the variability of stress feet instantaneous frequency decreases at shorter metronome periods. There was a

significant effect for metronome period on vIMF2, $X^2(1) = 17.02$, $p = .001$, with β = -0.04, and SE = 0.007. As β represents the change around the intercept at the shortest metronome period, prediction for the shortest metronome period is 0.4+(-0.04) = **0.36**. For the medium period it is 0.4+(-0.04) * 0 = **0.4**, and for the longest period it is 0.4+(-0.04) * -1 = **0.44**. The decrease in the variability of stress feet instantaneous frequency at shorter metronome periods indicates less variability is stress feet durations at shorter metronome periods.

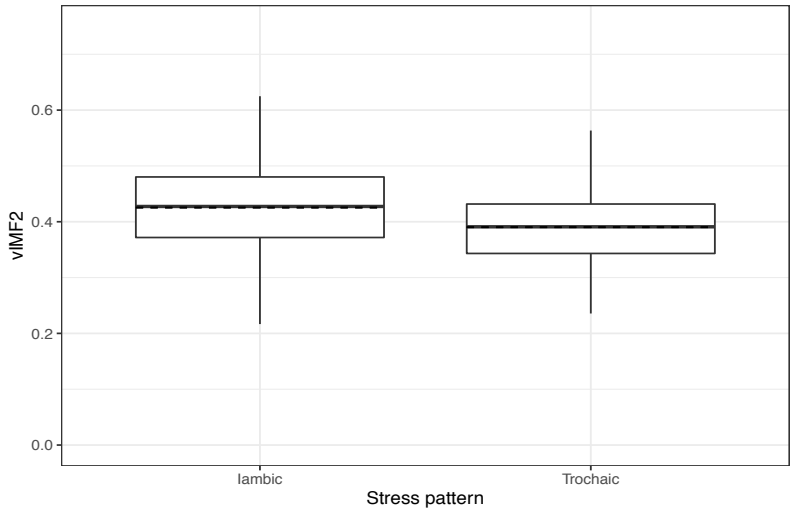


Figure 5.27: vIMF2 in iambic and trochaic sentences.

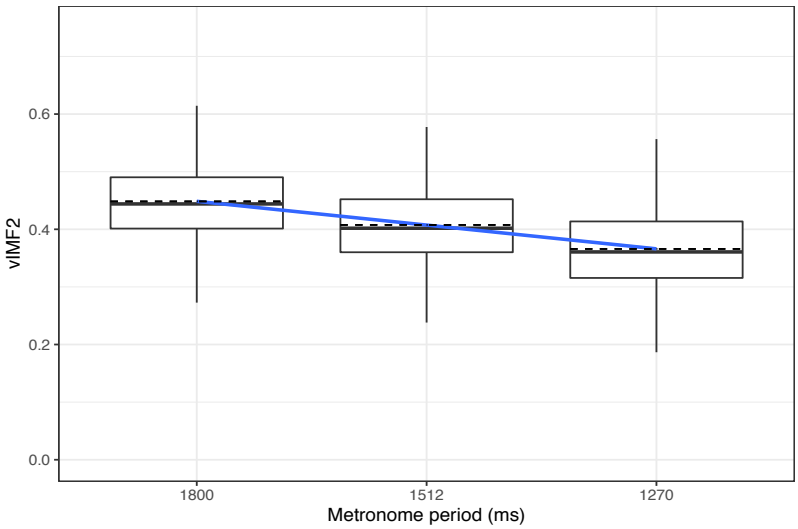


Figure 5.28: The effect of metronome period on vIMF2.

There were no two-way interactions between dialect and stress pattern, $X^2(1) = 0.04$, $p = .8$, or between dialect and metronome period, $X^2(1) = 0.4$, $p = .5$, or between stress pattern and

metronome period, $X^2(1) = 1.4$, $p = .2$. There was no three-way interaction between dialect, stress pattern and metronome period, $X^2(2) = 0.01$, $p = .9$.

Variability metrics, vIMF1 and vIMF2, did not reveal any differences between dialects. Despite the lack of discriminatory power in variability metrics, it is plausible to assert that variability metrics may reveal the degree of variability in stress distribution. For example, a regular alternation between stressed and unstressed syllables may lead to low variability in the instantaneous frequency of stress feet, while sparse distribution of stresses may lead to higher variability. As the distribution of stresses in our speech cycling text materials is similar across dialects, no difference in vIMF2 between dialects was detected. The variability in the instantaneous frequency of syllabic intervals may be sensitive to complexity of syllable structure. The latter assertion may be supported by the two-way interaction between metronome period and stress pattern in vIMF1. The greater decrease in vIMF1 across different metronome periods in the trochaic pattern can be attributed to the more complex unstressed syllables in the trochaic pattern (CVC) than in the iambic pattern (CV); more complex syllables are prone to greater compression from longer to shorter metronome periods, hence there are greater variability in syllabic instantaneous frequency across different metronome periods.

5.5.3.1 Power distribution metrics

Power distribution metrics quantify energy concentration at specific frequency ranges, which may reveal degree of dominance of either syllabic or stress feet time scales in the signal.

5.5.3.1 Centroid

The Centroid metric is a weighted mean of frequencies in the range from 2.5 Hz to 12 Hz. The model intercept is 5.25 Hz. This value corresponds to syllable level time scale rather than stress feet. More specifically, it corresponds to the period of stressed syllables (5.2 Hz = 192 ms), which indicates the dominance of stressed syllables time scale in the speech signal.

Figure 5.29 shows dialectal differences in centroid values. There was no effect for dialect, $X^2(1) = 0.19$, $p = .6$.

Figure 5.30 shows differences between iambic and trochaic sentences in centroid value. There was no effect for stress pattern on centroid value, $X^2(1) = 0.144$, $p = .2$.

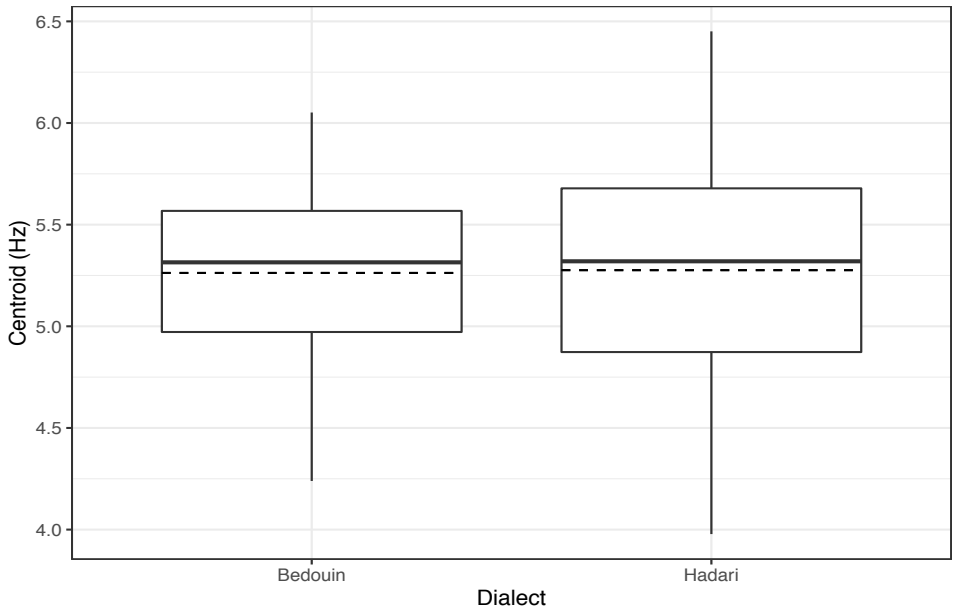


Figure 5.29: Dialectal differences in centroid value.

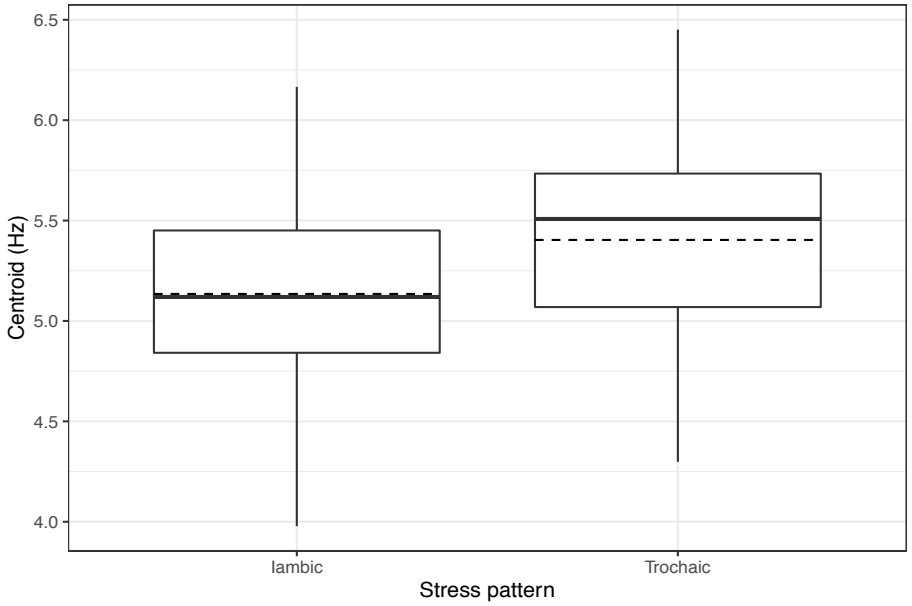


Figure 5.30: Centroid values in the iambic and trochaic sentences.

Figure 5.31 illustrates the effect of metronome period on Centroid value. Centroid value increases at shorter metronome periods. There was a significant effect for metronome period on Centroid value, $X^2(1) = 21.11$, $p < .001$, with $\beta = 0.13$ Hz, and SE = 0.02 Hz. As β

represents the change around the intercept at the shortest metronome period, prediction for the shortest metronome period is 5.23+0.13 = **5.38 Hz** (185 ms). For the medium metronome period the prediction is 5.23+0.13 * 0 = **5.23 Hz** (192 ms), and for the shortest metronome period it is 5.23+0.13 * -1 = **5.1 Hz** (196 ms). The Centroid value at all periods is within the range of syllabic intervals and in particular stressed syllables, thus the increase in Centroid value at shorter periods reflects compression in stressed syllables at shorter metronome periods.

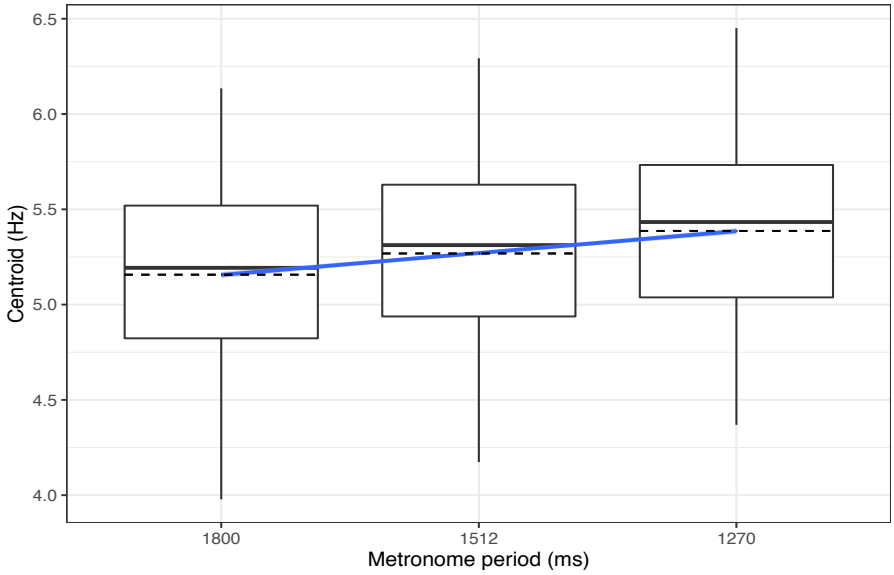


Figure 5.31: Metronome period effects on centroid value.

There were no two-way interactions between dialect and stress pattern, $X^2(1) = 0.67$, $p = .4$, or between dialect and metronome period, $X^2(1) = 0.01$, $p = .9$. There was a significant two-way interaction between stress pattern and metronome period, $X^2(1) = 4.6$, $p = .03$. Figure 5.32 plots the interaction. It can be seen that the difference between iambic and trochaic sentences in the Centroid value decreases at shorter metronome periods. In pairwise comparison, we tested whether the difference between the iambic and the trochaic sentences differs significantly across the metronome periods. The difference between iambic and trochaic sentences in the Centroid value was greater in the longest period than in the medium period, $p = .003$, and the difference between the iambic and trochaic sentences was greater at the medium period than at the shortest periods significant, $p = .005$. The smaller differences in the Centroid value between the iambic and trochaic sentences at shorter metronome

periods indicate that there is a compression in syllables duration, particularly, stressed syllables, from the longest to the shortest periods.

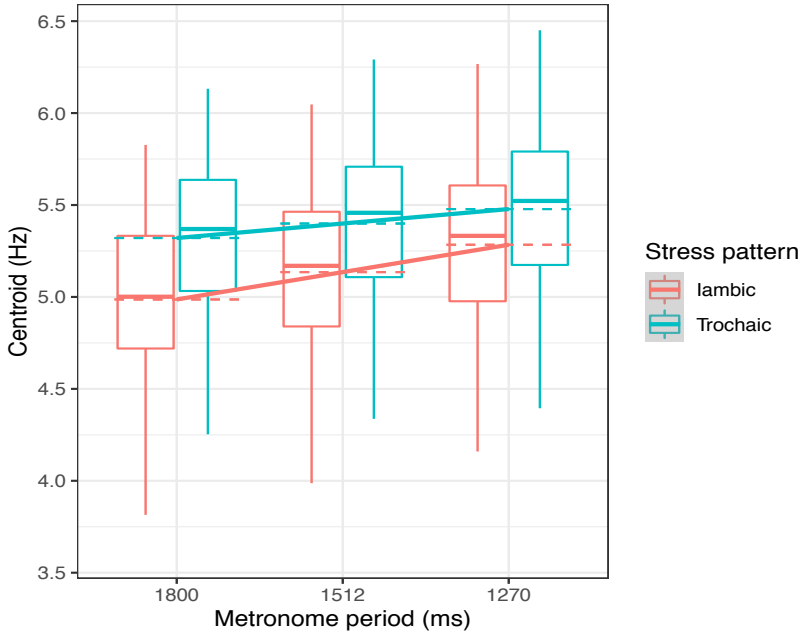


Figure 5.32: The effect of two-way interaction between stress pattern and metronome period on Centroid value.

There was no three-way interaction between dialect, stress pattern and metronome period, $X^2(2) = 0.6$, $p = .4$.

The fact that the Centroid was at a frequency value that corresponds to stressed syllables, ~ 5 Hz, suggests that it may be sensitive to acoustic cues that signal prominence. That is, the Centroid may reflect the greater lengthening degree and greater spectral strengthening associated with stressed syllables, and that this acoustic strengthening pattern is to some degree regular throughout the signal. As such, the Centroid metric may be useful in uncovering the gradient acoustic strengthening of stressed syllables between languages, and how this may consistent throughout the speech signal.

5.5.3.2 SBPr

The SBPr metric is computed by dividing the sum of powers in the frequency range that corresponds to stress feet time scale (2.5 Hz – 4 Hz) by the sum of powers in the frequency range that corresponds to syllables time scale (4.5 Hz – 12 Hz). Thus, SBPr quantifies the

relative amount of power concentration in stress feet time scale. The higher the value of SBPr, the greater power concentration at frequency ranges that corresponds to stress feet time scale.

The model's intercept value is 1. Figure 5.33 illustrates dialectal differences in SBPr value. There was no effect of dialect on SBPr value, $X^2(1) = 0.01$, $p = .9$.

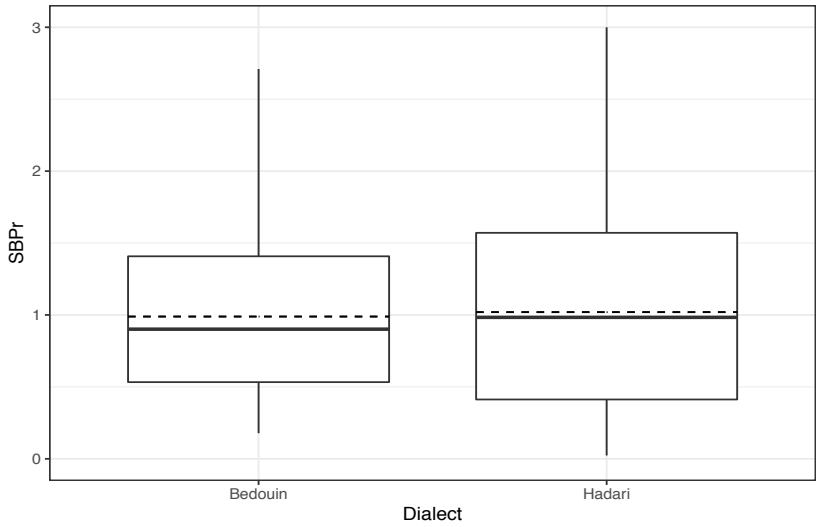


Figure 5.33: Dialectal differences in SBPr value.

Figure 5.34 illustrates the difference between iambic and trochaic sentences in SBPr value. There was no effect for stress pattern on SBPr value, $X^2(1) = 2.64$, $p = .1$.

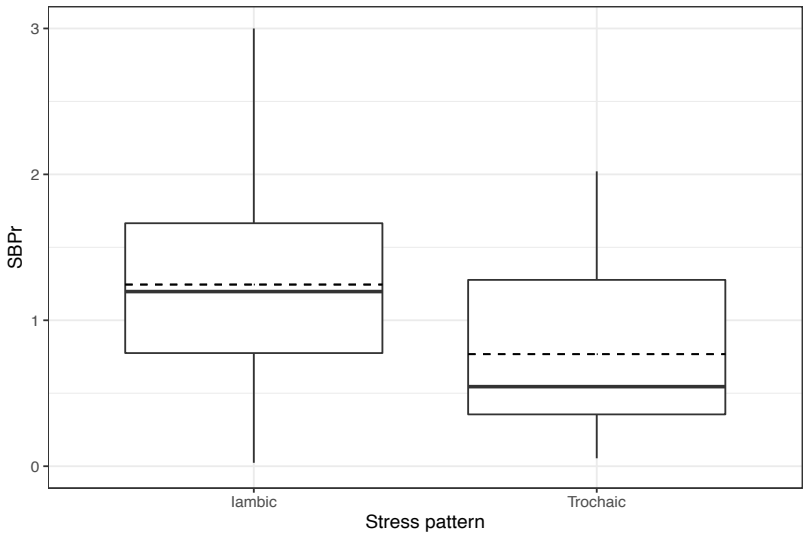


Figure 5.34: SBPr value in the iambic and trochaic sentences.

Figure 5.35 shows the effect of metronome period on SBPr value. As can be seen, the value of SBPr decreases at shorter metronome periods. There was a significant effect of metronome period on SBPr value, $X^2(1) = 7.43$, $p = .006$, with $\beta = -0.06$, and SE = 0.02. As β represents the change around the intercept at the shortest metronome period, prediction for the shortest metronome period is 1+(-0.06) = **0.94**. At the medium metronome period, the prediction is 1+(-0.06) * 0 = **1**, and at the longest period the prediction is 1+(-0.06) * -1 = **1.06**.

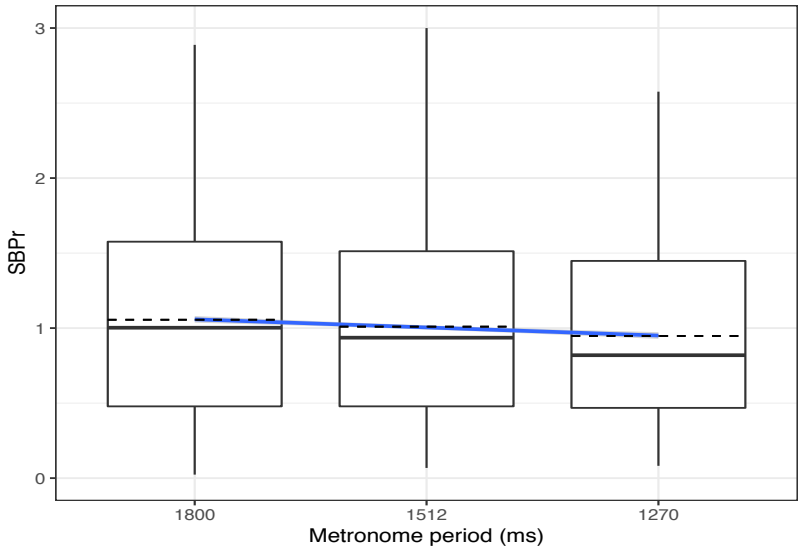


Figure 5.35: The effect of metronome period on SBPr value.

Since SBPr quantifies the relative power concentration in stress feet time scale, the decrease in its value at shorter metronome periods may reflect compression in stress feet intervals from the longest to the shortest periods.

There were no two-way interactions between dialect and stress pattern, $X^2(1) = 0.37$, $p = .5$, or between dialect and metronome period, $X^2(1) = 1.30$, $p = .2$. There was a significant two-way interaction between stress pattern and metronome period, $X^2(1) = 10.52$, $p = .001$. Figure 5.36 plots the interaction. It can be seen that there is a greater decrease in SBPr from the longest to the shortest metronome periods in the iambic sentences, while SBPr is similar across the different metronome periods in the trochaic sentences. Pairwise comparison showed that SBPr was larger in the longest period than in the medium period in the iambic sentences, $p < .0001$, but there was no difference between the medium and the shortest periods in the iambic sentences, $p = .1$. In trochaic sentences there was no difference in SBPr between the longest and the medium periods, $p = .4$, and between the medium and the shortest periods, $p = .2$.

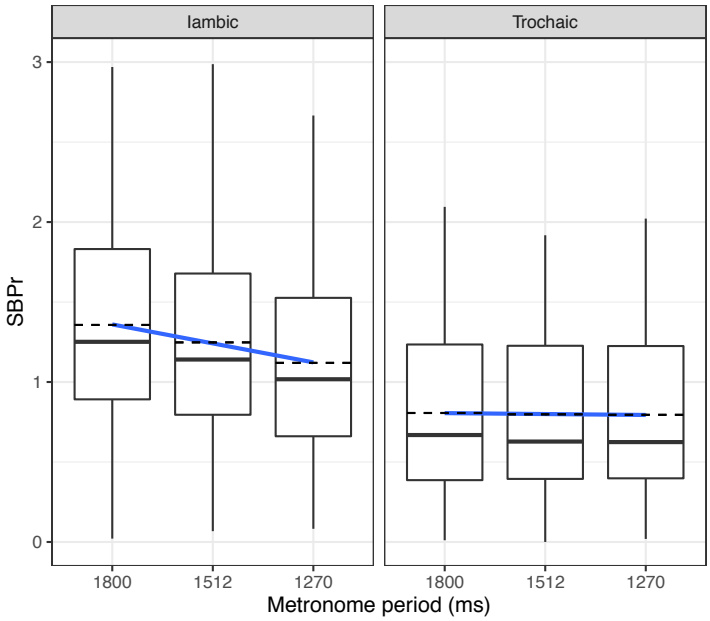


Figure 5.36: The effect of two-way interaction between stress pattern and metronome period on SBPr.

Since SBPr captures relative power concentration in stress feet time scale, then greater decrease in its value at shorter metronome periods in the iambic pattern indicates greater compression in stress feet intervals in the iambic pattern from the longest to the shortest periods. On the other hand, there is a greater tendency to regularize stress feet across different metronome periods in the trochaic pattern. However, another possible interpretation of this interaction may be made with reference to stressed syllable timing. In particular, it may be that within stress feet, stressed syllable durations vary between the iambic and trochaic sentences across different metronome periods. That is, stressed syllables compress more from the longest to the shortest metronome periods in the iambic sentences than in the trochaic sentences. Indeed, this interpretation accords with our earlier findings with regard to syllables duration, that stressed syllables in the iambic sentences were longer than in the trochaic sentences, which makes stressed syllables in the iambic sentences more prone to compressibility effects as a function of metronome period. Also, this interpretation may be supported by the fact that the Centroid value was at ~ 5 Hz which corresponds to stressed syllable timing.

There was no three-way interaction between dialect, stress pattern and metronome period, $X^2(1) = 0.3$, $p = .5$

5.5.3.3 Ratio21

Ratio21 metric computes the relative power concentration in the power spectrum of IMF2 which corresponds to stress feet oscillation, by dividing its sum of powers by the sum of powers of IMF1 spectrum, which corresponds to the syllable level oscillation.

The model intercept value is 1.04. Figure 5.37 shows dialectal differences in Ratio21. There was no effect of dialect, $X^2(1) = 0.01$, $p = .9$.

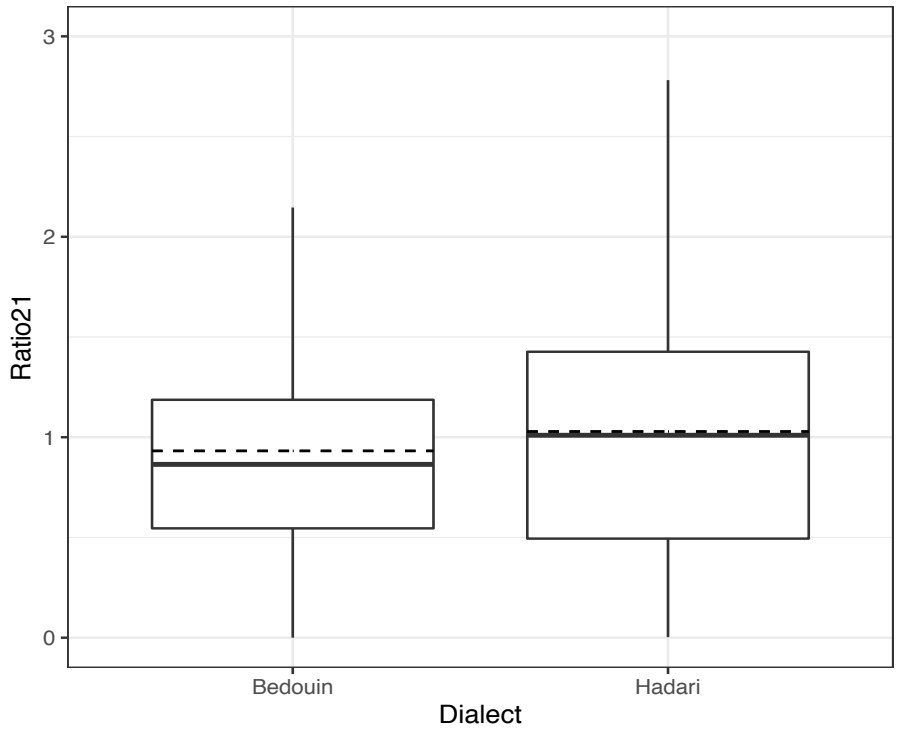


Figure 5.37: Dialectal difference in Ratio21 value.

Figure 5.38 shows the difference between iambic and trochaic sentences in Ratio21 value. There was no effect of stress pattern, $X^2(1) = 3.34$, $p = .06$.

Figure 5.39 shows Ratio21 value across different metronome periods. There was no effect of metronome period on Ratio21, $X^2(1) = 1.77$, $p = .1$.

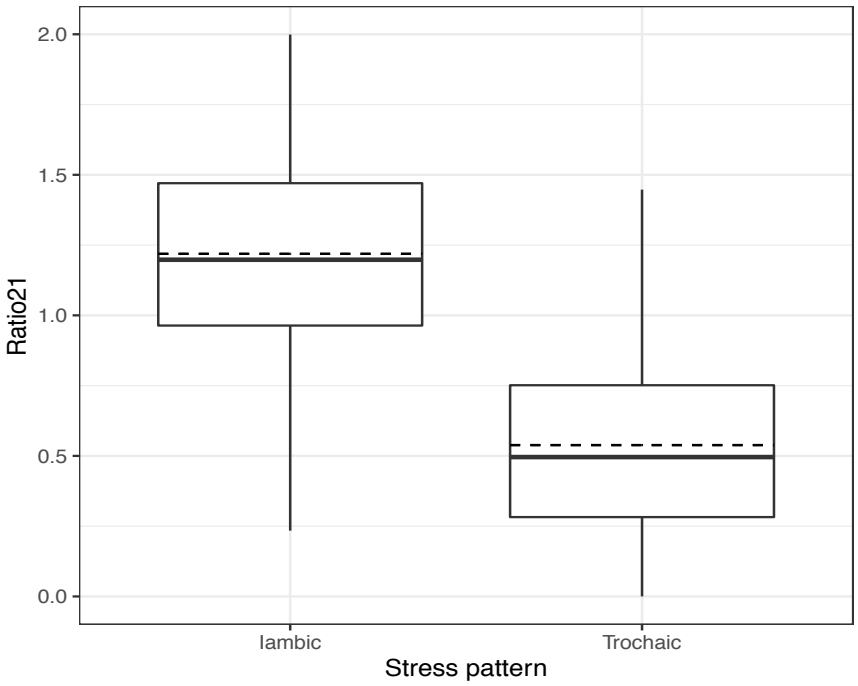


Figure 5.38: Difference between iambic and trochaic sentences in Ratio21 value.

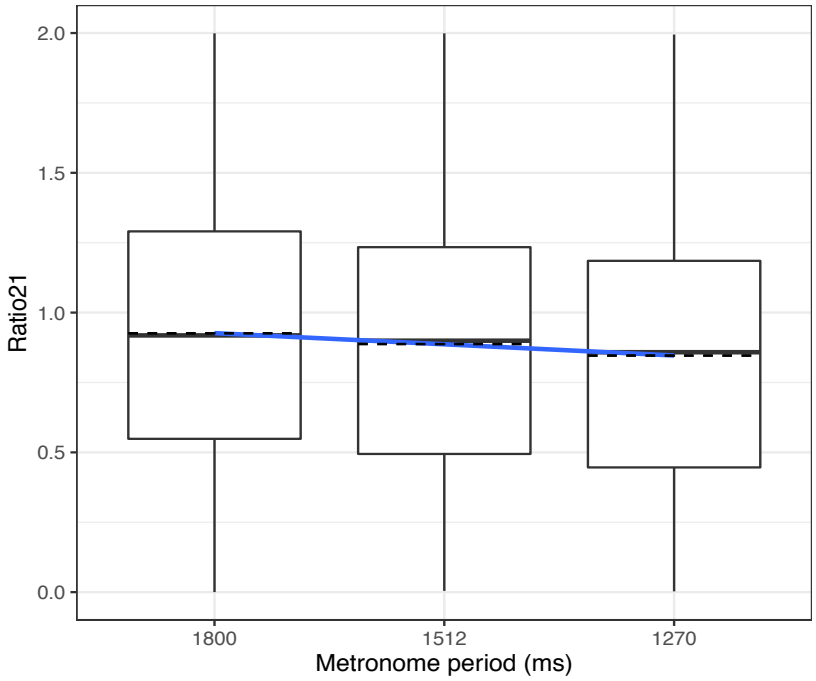


Figure 5.39: Ratio21 values across different metronome periods.

There was no two-way interaction between dialect and metronome period, $X^2(1) = 1.4$, $p = .2$. There was a significant two-way interaction between dialect and stress pattern, $X^2(1) = 57.43$, $p < .001$. Figure 5.40 plots the interaction.

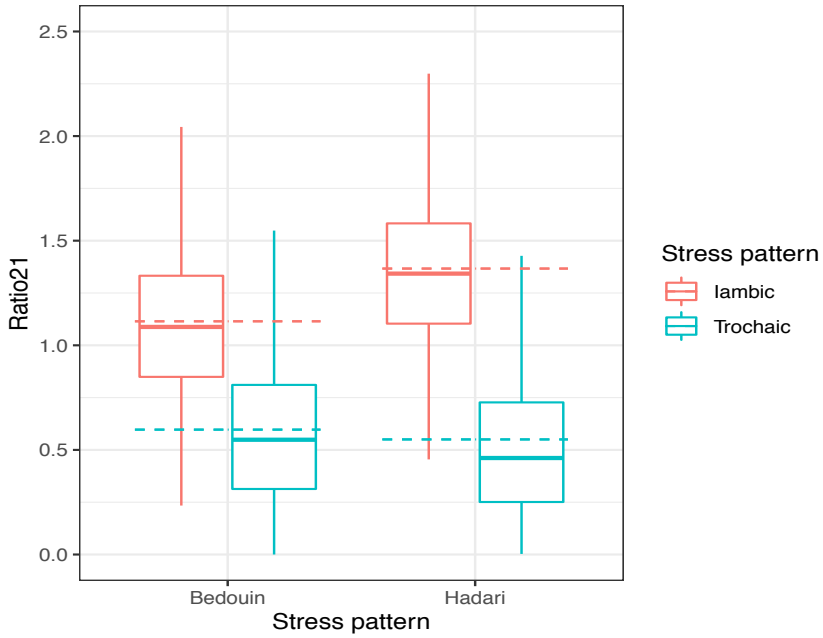


Figure 5.40: The effect of two-way interaction between dialect and stress pattern on Ratio21 value.

There are larger differences in Ratio21 between iambic and trochaic sentences in Hadari than in Bedouin. Pairwise comparison showed that the difference in Ratio21 between iambic and trochaic sentences was significant in Hadari, p = .02, but not in Bedouin, p = .1.

As Ratio21 quantifies power concentration in stress feet frequency range, the higher value in Ratio21 in the iambic sentences in Hadari than in Bedouin may indicate greater regularity of stress feet in Hadari than in Bedouin in the iambic sentences. However, another explanation that refers to stressed syllable timing might be plausible. The higher Ratio21 in the iambic sentences than in the trochaic sentences, especially in Hadari, may reflect greater durational ratio of stressed syllables relative to unstressed syllables in the iambic sentences. To further explore the possibility of this explanation, we computed the durational ratio between stressed and unstressed syllables in the iambic and trochaic sentences in both dialects. We found that the durational ratio between stressed and unstressed syllables in the iambic sentences is larger in Hadari, 1.9, than in Bedouin, 1.7, and similar between dialects in the trochaic sentences, 1.1. This pattern, i.e., the larger temporal stress ratio in the iambic sentences in Hadari, may lend support to an explanation of Ratio21 results that refers to syllabic temporal stress contrast. Also, the Centroid value, ~5 Hz, which corresponds to stressed syllable time scale may support the idea that power distribution metrics, including Ratio21, are sensitive to stressed syllables durational pattern.

There was no two-way interaction between stress pattern and metronome period, $X^2(1) = 0.01$, $p = .9$, and there was no three-way interaction between dialect, stress pattern and metronome period, $X^2(2) = 0.8$, $p = 0.6$.

5.6 Summary and discussion

We analysed multiple statistical metrics derived from the power spectrum, and the instantaneous frequency of low pass filtered amplitude envelope which contains frequency modulations at syllable-level and foot-level time scales.

Rates of syllabic oscillation and stress foot oscillation, mIMF1 and mIMF2, were 6.4 Hz and 3.22 Hz, respectively, which corresponds to a mean syllable duration of 158 ms and a mean stress feet duration of 310 ms. However, as demonstrated earlier, mIMF1 and mIMF2 did not correspond precisely to the mean duration of syllables and stress feet in the original signal, which was 181 ms and 335 ms, respectively. We explained the distorted representation of the signal in mIMF1 and mIMf2 as potentially arising from the interpolation process when constructing the upper and lower envelopes during the sifting process (Figure 5.11). Interpolation may result in the inclusion of ultra-frequency and low amplitude data points, thus leading to the distortion of the physical representation of the original signal (Huang et al., 1998, p. 921). Kim et al. (2012) suggested that smoothing, rather than interpolation, in the sifting process may be useful to mitigate against the effect of ultra-frequency data points from distorting signal representation. One of the consequences of the distorted signal representation is that the difference in syllabic oscillation and stress feet oscillation rate between different metronome periods did not exceed the perceptual threshold for rate differences. Thus, it is advised for future work to use a smoothing process when obtaining IMFs.

Despite these shortcomings in the computation of IMFs, findings from Tilsen and Arvaniti (2013) showed that mIMF1 and mIMF2 are useful in describing syllables and stress feet rate. Tilsen and Arvaniti (2013) showed reliable between-language differences in syllables and stress feet rate, that are consistent with cross-language rhythmic characterisation. For instance, English demonstrated slower mIMF1 oscillation rate than Greek and Spanish, due to the greater syllabic complexity in English than in Spanish and Greek. Thus, rate metrics

can be useful in describing syllable rate and can reduce segmentation time in larger speech corpora.

Rhythmic stability metrics compute variability in syllable level and stress foot level instantaneous frequency, vIMF1 and vIMF2. In both metrics, variability drops at shorter metronome periods, which is not surprising, as syllables and stress feet temporal variation decreases shorter metronome periods which are associated with faster speaking rates.

While rhythmic stability metrics are meant to capture between-dialect relative dominance of either syllable level or stress foot level time scales, there were no detectable differences between dialects. Possibly, the weak discriminatory power between dialects in vIMF1 and vIMF2 can be attributed to what these metrics capture. vIMF1 and vIMF2 seem to reflect complexity in syllable structure and regularity in stress distribution, respectively. Since syllabic complexity and stress distribution are similar between dialects in our speech cycling corpus, no dialectal difference was detectable. Findings from Tilsen and Arvaniti (2013) support this conclusion. English and German demonstrated higher vIMF1 than Korean, Spanish and Italian, which concords with the fact that the former languages have greater degree of syllabic complexity than the latter group of languages. English had lower vIMF2 than Italian and Spanish, since English is known to have relatively more regular alternation between stressed and unstressed syllables than Spanish and Italian. An unexpected finding was the lower vIMF2 in Korean than in English, although Korean does not have lexical stress, and thus does not regulate the distribution of stresses. The lower vIMF2 in Korean than in English may be due to the distribution of syllables within accentual phrases in Korean, thus reflecting language-specific prominence constituency. In all, the quantification of syllabic complexity and stress distribution in large speech corpora could benefit from vIMF1 and vIMF2 metrics.

In the power distribution metric, Centroid, the intercept value was at 5.2 Hz, which corresponds to the stressed syllable period at 192 ms. Therefore, the Centroid metric may be sensitive to acoustic prominence, and it is particularly sensitive to stressed syllable duration. This finding is important, as it aids our interpretation of SBPr and Ratio21 metrics. These latter metrics quantify relative power concentration in stress feet time scale. However, an interpretation that refers to stressed syllables, within stress feet, seems to better reflect the durational patterns in our data which were investigated in Experiment 1 (b). In particular, we

found in SBPr measure two-way interaction between stress pattern and metronome period, where SBPr value decreased more noticeably from the longest to the shortest metronome period in the iambic sentences than in the trochaic sentences. An interpretation that refers to stressed syllables' duration accords with the fact that stressed syllables in the iambic patterns were longer than in the trochaic pattern, thus stressed syllables are more prone to compression effects in the iambic pattern, from the longest to the shortest metronome periods. As for Ratio21, we found two-way interaction between dialect and stress pattern, where Hadari showed greater contrast in Ratio21 between the iambic and trochaic sentences than Bedouin. We interpreted this finding as reflecting the temporal ratio between stressed and unstressed syllables across the iambic and trochaic sentences between the two dialects. To examine this interpretation, we measured the temporal ratio between stressed and unstressed syllables across the iambic and trochaic sentences between the two dialects and found that Hadari had greater contrast between iambic and trochaic sentences than Bedouin.

Thus, power distribution metrics seem to be sensitive to stressed syllable timing, and the temporal ratio of stressed to unstressed syllables between dialects across the iambic and trochaic patterns. As a result, our analyses do not support positing a role of stress feet as a timing unit in speech production in speech cycling.

Findings from Tilsen and Arvaniti (2013) also suggest that interpreting power distribution metrics with reference to stressed syllable timing provides a better account of cross-language rhythmic variation than referring to the stress feet timing. In Tilsen and Arvaniti, Centroid did not appear at stress feet time scale (wherein the cut-off for stress feet was ≤ 3.25 Hz), rather, it was at the syllabic time scale, and varied between languages. For instance, "stress timed" English had the lowest Centroid, at stressed syllables time scale (ranging between 3.4 Hz and 4 Hz for different elicitation methods) while "syllable timed" Spanish and Italian had higher Centroid, indicating that stressed syllables in English are longer and more complex than in Spanish and Italian. English had also higher SPBr and Ratio21 values than Spanish and Italian. If ratio metrics are interpreted as reflecting different degrees of temporal stress contrast, then the higher ratio scores in English than in Spanish and Italian indicate greater temporal stress contrast in English than in Spanish and Italian, which accords with the rhythmic chrematistics of these languages (cf. Dauer, 1983; White & Mattys, 2007a).

Chapter 6. General discussion and conclusion

6.1 Temporal coordination in speech cycling in Hadari and Bedouin

In speech cycling tasks, we aimed to explore differences between Hadari and Bedouin Kuwaiti Arabic dialects in aligning vowel onsets of stressed syllables to simple harmonic phases. In the external phase measure, in a four-way interaction, we found in the shortest metronome period in the trochaic pattern that Bedouin tended to align heavy and light syllables similarly, close to the harmonic phase 0.5. In contrast, Hadari tended to align light syllables earlier than heavy syllables.

There are two processes that may explain the variable phase alignment between dialects in the external phase measure. First, the close alignment of heavy syllables in the shortest metronome period in the trochaic pattern to the harmonic phase of 0.5 in both dialects may be attributed to a top-down effect. As heavy syllables have strong contrast with unstressed syllables, due to their phonological length, they attract closer alignment with the harmonic phase of 0.5 (see section 6.2 below for further discussion). As for light syllables in the shortest metronome period in the trochaic pattern, the tendency in Hadari to align light syllables earlier than heavy syllables may be due to greater unstressed syllable compression, leading to earlier phase alignment.

In the internal phase measure, we found that both dialects tended to align heavy syllables similarly, close to 0.5 phase; however, Hadari aligned light syllables closer to 0.5 phase and more similarly to heavy syllables than Bedouin. Similar to the explanation in the external phase, the closer internal phase alignment of heavy syllables, in both dialects, to a harmonic phase of 0.5 is because heavy syllables have strong contrast with unstressed syllables, reflecting a top-down effect. The earlier alignment of light syllables in Hadari, and closer to 0.5 than in Bedouin, is probably due to greater unstressed syllable reduction in Hadari, which leads to earlier alignment of light syllables.

We also found that metronome period mediates temporal coordination with simple harmonic phases. At the shortest metronome periods, there was closer alignment to the harmonic phase of 0.5, in the internal phase and in the external phase measures. As shorter metronome periods are associated with faster speaking rates, it is possible that there is a preferable

speaking rate, i.e., faster speaking rates, for temporal coordination with the harmonic 0.5 phase (see section (2.11) for further discussion).

To sum up, three potential factors may influence phase alignment in Hadari and Bedouin dialects. The first is a top-down effect. The greater contrast of heavy syllables with unstressed syllables, due to heavy syllables' phonological length, prompts closer alignment to the harmonic phase of 0.5. The second is a bottom-up effect. The tendency in Hadari to exhibit greater unstressed syllable reduction than in Bedouin could potentially lead to an earlier alignment of light syllables in Hadari than in Bedouin. The third is speaking rate, as faster speaking rates were associated with closer alignment to a harmonic phase angle.

We also examined syllable duration to explain the variable phase alignment between dialects. We found that Hadari exhibited greater unstressed syllable reduction than Bedouin, in the phrase-initial position, in the longest and the shortest metronome periods. This finding may support our assertion that greater unstressed syllable reduction in Hadari leads to an earlier alignment of light syllables. A schematic for the potential effect of unstressed syllable reduction in Hadari on earlier phase alignment is provided in Figure 6.1.

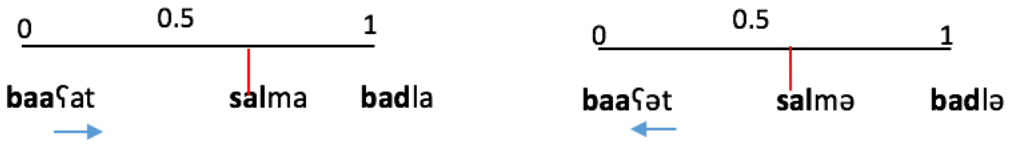


Figure 6. 1: Schematic illustration of the potential effect of unstressed syllable reduction in Hadari, right, on earlier phase alignment, as indicated by the blue arrows.

6.2 Top-down and bottom-up effects in other speech cycling experiments

There are examples in other speech cycling experiments that may support the relevance of the effects mentioned above, top-down, bottom-up, and speaking rate, on temporal coordination in speech cycling. We discussed earlier in section (1.6.1) the difference between English and Spanish, and Italian in constrained speech cycling (Cummins, 2002), in which speakers were asked to align stresses with high-tone and low-tone metronome targets. English speakers

could perform the task easily, while Italian and Spanish found it difficult as they needed more than 30 minutes of training before they could perform the task. Our explanation of this difference refers to the greater stress contrast in English than in Italian and Spanish. English contrasts stressed and unstressed syllables through substantial lengthening of stressed syllables and unstressed syllable reduction; on the other hand, there is less stressed syllable lengthening, especially in Spanish, and less unstressed syllable vowel reduction in Italian and Spanish (Grabe & Low, 2002; White & Mattys, 2007a). Greater stress contrast in English affords emphasising stressed syllable beats through close alignment to metrically important points in speech cycling. Lower stress contrast in Italian and Spanish could make the task unnatural, as a high degree of stress beat emphasis is less affordable by the degree of stress contrast in Italian and Spanish.

As for bottom-up effects, the differences in speech cycling between English and Japanese (Tajima, 1999) and English and Jordanian Arabic (Zawaydeh et al., 2002) provide useful examples. For instance, when the number of unstressed syllables preceding the medial stressed syllable increased, there were later internal phase alignments in Japanese and Jordanian Arabic than in English. This is, possibly, because English allows compression of unstressed syllables to a greater degree than Japanese and Jordanian Arabic, which allows for an earlier internal phase in English, and closer to a harmonic phase angle.

The mediation of speaking rate to temporal coordination was found in Tajima's (1999) speech cycling. In Japanese, slower rates were associated with 2/3 simple phase, while faster rates were associated with 1/2 phase alignment. However, in our data, there was no discrete change in the rhythmic mode across different metronome periods. Rather, the fast rate was associated with closer alignment to 0.5 phase than the slower rates. Perhaps, the difference between our data and Tajima's is that in our case, there are fewer rates than in Tajima's study, which involved ten rates.

6.3 The potential mutual timing influences between stress feet and syllables

Experiment 2 investigated the potential hierarchical temporal relation between stress feet and syllables in speech cycling. Our hypothesis regarding the temporal relation between stress feet and syllables was based on the coupled oscillators model (O'Dell & Nieminen, 1999). The coupled oscillators model asserts mutual timing influences between higher-level and

lower-level prosodic units. We analysed statistics of the amplitude envelope at stress foot energy time scale, 2.5 Hz – 4 Hz, and syllable energy time scale, 4.5 Hz – 12 Hz, to investigate the potential interaction in timing between stress feet and syllables. We found no evidence for the effect of stress feet on the timing of syllables. In particular, the power distribution metric, Centroid, reflected the dominance of stressed syllables in the speech signal, and reference to stressed syllable timing better captures the durational patterns in our data, which were investigated in Experiment 1 (b). For instance, in the SBPr metric, which quantifies relative power concentration in stress feet time scale, we found a greater decrease in SBPr across metronome periods in the iambic pattern than in the trochaic pattern. Interpreting the greater decrease in the iambic pattern in SBPr across metronome rates with reference to stressed syllables in iambic feet is in line with our findings in Experiment 1 (b). There, we found that stressed syllables are longer in the iambic feet than in the trochaic feet. Thus, it is not surprising that longer stressed syllables in the iambic feet exhibit greater compression across metronome rates, as reflected in the greater SBPr decrease across metronome rates in the iambic pattern. With regard to dialectal differences, in the Ratio21 metric, which quantifies power concentration in stress feet instantaneous frequency, there was a two-way interaction between dialect and stress pattern. Hadari exhibited greater contrast in the Ratio21 metric between the iambic and the trochaic patterns, with a higher Ratio21 value in the iambic sentences. Our interpretation referred to differences in the temporal ratio between stressed and unstressed syllables in the iambic feet and the trochaic feet between dialects. We measured the temporal ratio between stressed and unstressed syllables and found that Hadari exhibited greater contrast in temporal ratio across the iambic and trochaic sentences, supporting the notion that Ratio21 reflects temporal stress ratio. Thus, interpreting the power distribution metrics with reference to stressed syllables captures the durational patterns in our data (Experiment 1 (b)). Accordingly, there seems to be no evidence for an effect of stress feet on the timing of syllables. As for rate metrics, mIMF1 and mIMF2, and rhythmic-stability metrics, vIMF1 and vIMF2, we concluded that they might be useful in capturing syllable rate, syllable complexity, and stress distribution in large speech corpora, which can reduce the time required for speech segmentation.

6.4 Limitations

The experimental design of speech cycling in our study was not without limitations. First, we have demonstrated in Chapter 3 that the shorter unstressed syllables in the iambic sentences

than in the trochaic sentences led to earlier phase alignment in the former than in the latter. Therefore, differences in phase alignment between the iambic and trochaic sentences do not reflect prosodic constraints but it is due to text materials. It is cumbersome to have similar unstressed syllable structures across the iambic and trochaic sentences because most iambic words in Kuwaiti Arabic have simple unstressed syllable structure, CV, and most trochaic words have complex unstressed syllable structure, CVC. However, a potential way that might aid in exploring prosodic timing differences between the iambic and trochaic sentences in speech cycling is by considering alignment with metronome beeps at the start of the sentences. We have suggested earlier in section 2.7 that a difference between the iambic and trochaic patterns may reside in the alignment with metronome beeps, which may be perceived as Phase 0 in the repetition cycle. Phrase-initial stressed syllables in the trochaic sentences could be easily aligned with metronomes, while in the iambic sentences, phrase-initial stressed syllables are separated from metronomes with phrase-initial unstressed syllables. Given the salience of metronomes in speech cycling, it is possible that speakers may tend to start stressed syllables in the iambic sentences with metronomes (with unstressed syllables uttered earlier than metronomes) or apply strong reduction to unstressed syllables so that stressed syllables are closer to metronomes. Investigating these two possibilities would have been possible had we accounted for the metronomes position when we segmented the data. Thus, future work in speech cycling may be improved by considering the alignment with metronomes, especially if trochaic and iambic materials are to be used.

Second, the statistical analyses of this study may be improved by reducing the predictors and interaction levels. It may be difficult for the reader to follow the results in complex interactions, such as the four-way interaction between dialect, rate, weight, and stress pattern in phase analysis. Perhaps, one factor that may be dropped is the stress pattern because, as discussed above, the different syllable structures between iambic and trochaic words in Kuwaiti Arabic dialects obliterate the comparison in prosodic timing constraints between the two stress patterns. Also, a simpler statistical analysis with fewer predictors and with lower interaction levels may have the advantage of facilitating cross-linguistic comparisons in speech cycling.

6.5 Theoretical implications

Cummins and Port (1998) and Tajima (1999) have accounted for the temporal alignment of stressed vowel onsets at simple phase angles within the phrase repetition cycle with an oscillatory process. The repetition of sentences at a constant metronome period leads to the emergence of harmonic timing effects, in which stressed vowel onsets will tend to lie at simple integer ratios, e.g., 1:2 and 1:3, of the phrase repetition cycle. The harmonic timing effect is a case of an oscillatory process in which two oscillators (the phrase oscillator and the stressed vowel onsets oscillator) are phase-locked at simple phase angles, e.g., 1/2 and 1/3, which act as attractors to prominent events of speech. The coupled oscillators model is also representative of the hierarchical structuring between the phrase repetition cycle and stressed vowel onsets; the harmonic timing effects constrain the temporal alignment of stressed vowel onsets at simple phases, which establishes a nesting relation between higher-level and lower-level speech units. Furthermore, the metrical structure of hierarchically coordinated units is reflected in the phase ratios established between prosodic units. For instance, a 1/2 phase ratio between the final stressed syllable and the phrase repetition cycle in sentences made of three stresses reflects a structure of four beats, with the last beat being a silent one.

Importantly, however, the findings from our speech cycling experiment, as well as from Tajima's (1999) study, suggest that speaking rate may influence the emergence of temporal attractors at certain phase angles. Findings from Tajima (1999) showed that Japanese speakers showed different phase alignment at different metronome rates, which are associated with different speaking rates, and we found that stronger temporal coordination with 1/2 phase in Hadari and Bedouin Kuwaiti Arabic dialects emerged at the fastest speaking rate. Therefore, the model of coupled oscillators may be improved by accounting for rate effects, as suggested by Haken et al. (1985) in accounting for phase patterns in finger movements tasks in Kelso et al.'s (1979) study. As we have reviewed in section 1.5.4.1, finger movements, which were controlled by a pacing metronome, exhibited two phase relations. At the slow metronome rate, there was an anti-synchronous phase relation at phase 1/2 and a synchronous phase at phase 0; however, at the fast rate there was only a single stable phase which is the synchronous phase at phase 0 (see Figure 1.5). Haken et al. modelled the stable phase angles with the superposition of two cosine waves, which established attractors at 1/2 and 0 phase angles. Crucially, the stability of attractors was influenced by a control parameter, which captures metronome rate effects on stable rhythmic

modes. As the value of the control parameter decreases, corresponding to an increase in rate, only a single attractor at phase 0 is stable (see Figure 1.6). Such control parameter appears to be useful in capturing speech rate effects on the temporal coordination patterns in speech cycling since speech rate mediates temporal coordination with certain phase angles in speech cycling.

Note, however, that Cummins and Port (1998) have pointed out that the change in phase alignment with different speaking rates may not reflect hierarchical temporal coordination with the phrase repetition cycle; rather, it may reflect preference towards alignment of vowel onsets at a preferred durational window. For instance, if at slow rates the final stressed syllable was at 2/3 phase angle and at faster rates phase alignment changes to 1/2, it may be due to preference to align stressed vowel onsets' intervals at longer durational window in the slow rate (since 2/3 is a later target in the cycle) and at a shorter durational window in the fast rate. However, as we have argued in section 2.11, the effect of speaking rate and hierarchical temporal coordination need not be mutually exclusive. The different temporal coordination patterns around 1/2 phase of stressed heavy and stressed light syllables between Hadari and Bedouin dialects only emerged at the fast speaking rate, which reflects different hierarchical organizations based on different degrees of temporal stress contrast between the two dialects.

While accounting for the hierarchical nesting between prosodic units with a system of oscillators that are phase-locked at harmonic frequencies is specific to speech cycling tasks, there are some aspects of the coupled oscillators model that seem to be useful in accounting for timing interaction in natural dialogue. For instance, Wilson and Wilson (2005) suggested that in smooth turn-taking, listeners entrain to the syllable oscillation rate of the speaker and time the onset of their turns with an *anti-phase* relation to the speaker's syllable oscillation rate to avoid overlap. In section 1.5.4, we have referred to Włodarczak et al.'s (2012a,b) work, which showed that in overlapped speech, interlocutors initiate their overlaps at syllables' vowel onsets, demonstrating an *in-phase* relation between the speech of interlocutors. Of course, the nature of temporal coordination in speech cycling and natural dialogue is different. The simple division of the phrase repetition cycle into simple integer ratio, i.e., harmonic timing effects, are clearly not expected in natural dialogue since these harmonic timing effects emerge due to specific task demands of speech repetition with regular metronomes. However, the influences of speech rate and the degree of temporal stress contrast on temporal coordination, as our speech cycling findings show, may provide insights

into the nature of temporal coordination in natural dialogue. We have demonstrated in our speech cycling findings that speech rate, especially the fast rate and the relative stress contrast may afford temporal coordination with simple phase angles. The assertion of Wilson and Wilson (2005) that smooth turn-taking is achieved through entrainment to the interlocutor's speech rate may support our findings on the affordance of speech rate to temporal coordination. Furthermore, Włodarczak et al. (2012a) have demonstrated that overlapped speech is modulated by speech rate; the likelihood of initiating an overlap increases at faster speaking rates of the overlappee speech. As for the potential effect of temporal stress contrast, it is plausible to assume a scenario in which an overlaper may favour initiating overlaps at more prominent vowel onsets of stressed syllables relative to unstressed syllables. This tendency may be facilitated by entrainment to the rate of alternation between stressed and unstressed syllables throughout the speech signal, so that the dialogue partner may predict the occurrence of stressed vowel onsets to initiate an overlap. Being modulated with acoustic cues such as speech rate and temporal stress contrast, temporal coordination in natural speech can be understood as an *affordance* of the acoustics of speech to the listener's entrainment to the speech of their dialogue partner (Cummins, 2009, p. 17). Temporal phase relations, in-phase and anti-phase, can be considered hallmarks for entrainment; they reflect the listener's entrainment to speech acoustic cues throughout the signal to predict the occurrence of salient points in the speech stream, e.g., stressed vowel onsets. It is noteworthy that the notion of affordance in temporal coordination demonstrates a departure from hierarchical constituent size influences that certain hierarchical timing models may imply (e.g., O'Dell & Nieminen, 2009). In particular, constituent size effects imply syllabic shortening (see section 1.5.1) to mark the structure of speech; however, temporal coordination may be greatly facilitated through local lengthening effects rather than shortening. As listeners in natural dialogue may tend to predict the occurrences of stressed vowel onsets, greater stress-based lengthening may facilitate this tendency as it contributes to greater salience of stressed vowel beats.

Considering the plausibility of temporal coordination in natural speech, we will attempt in the next section to infer predictions on potential temporal coordination patterns between speakers of Arabic dialects, especially Hadari and Bedouin Kuwaiti dialects, in natural dialogue. We will provide more details on how temporal stress contrast and speech rate may afford temporal coordination in natural dialogue.

6.6 Empirical implications and future research

We have discussed briefly in the previous section the study of Wilson and Wilson (2005) on temporal coordination in smooth turn-taking and the studies of Włodarczak et al. (2012a,b) on temporal coordination in overlapped speech. Discussion of these studies suggested that temporal coordination in natural dialogue emerges as an affordance of temporal acoustic cues, such as speech rate and temporal stress contrast, to listeners' entrainment to the speech structure. In this section, we seek to elaborate more on how timing cues may afford temporal coordination. Then, we will attempt to predict potential patterns of temporal coordination between Hadari and Bedouin Arabic dialects based on prosodic timing differences that we found in our speech cycling experiment.

We referred in section (1.5.4) to Włodarczak et al.'s (2012a,b) studies on overlapped speech. They found that English, German and French speakers timed their overlaps towards the end of vowel-to-vowel (VTV) boundaries. There were some language-specific patterns. Within VTV units, English and German speakers timed their overlaps close to the onset consonant(s) of the following syllable, while French speakers timed their overlaps closer to the vowel onsets of the following syllable. These patterns may suggest the relevance of the p-centre (Marcus, 1981) in overlapped speech. As reviewed in section 1.5.4.2, the p-centre occurs near vowel onsets, but its exact location can be influenced by syllable structure; with longer onset consonants, the p-centre may occur earlier than the vowel onsets, within syllabic onset consonant(s). As English and German have more complex onset consonant clusters, it may not be surprising that overlaps were at syllabic consonant onsets. On the other hand, simpler onset consonant clusters in French may have led to overlap initiation closer to vowel onsets. Importantly, the tendency to overlap at VTV unit boundaries reflects temporal coordination between interlocutors, with an *in-phase* relation between interlocutors.

Overlaps in Włodarczak et al.'s data were initiated at later positions of the interlocutor's utterances, suggesting that listeners may use timing variation throughout the signal to predict the point of overlap, i.e., the p-centre. One potential source of timing variation that may afford predicting the p-centre is temporal stress contrast. As we have suggested in the previous section, listeners in overlapped speech may tend to initiate overlaps at stressed syllable beats. As such, entrainment to the stress rate, i.e., the rate of alternation between

stressed and unstressed syllables, may facilitate predictions of the occurrence of stressed syllable beats to initiate an overlap at these points.

Włodarczak (2014) provided a more detailed analysis regarding the potential effect of temporal stress contrast on temporal coordination in overlapped speech (although there was no indication whether overlapped VTVs were stressed or not). Włodarczak investigated the normalized variability index (nPVI) (Low et al., 2000) of three VTV units that preceded the overlapped VTV. Overlaps at phrase-final positions were excluded from the analysis, as these overlaps may be due to prosodic levels higher than lexical stress, i.e., phrase-final boundaries. Scores of nPVI were divided into three categories: high, mid, and low. The findings were that lower scores of nPVI (mid and low) increased the likelihood of overlap initiation more than the high score of nPVI. As nPVI measures the temporal stress contrast of sequential vocalic intervals, these findings suggest that lower temporal stress contrast throughout the signal is associated with a greater likelihood of overlap initiation. This is contrary to our suggestion that greater temporal stress contrast may facilitate temporal coordination. However, another interpretation regarding the effect of temporal stress contrast is possible. There might be an interaction between speaking rate and temporal stress contrast in affording temporal coordination. The lower scores of the nPVI metric may be due to faster speaking rate since durational intervals tend to be shorter at faster rates, hence the lower nPVI scores. Therefore, listeners may tend to entrain to the rate of stress occurrences at faster speaking rate, which in turn facilitates predicting salient syllable beats to initiate overlaps.

As the above analysis was made for overlaps in the middle of the overlapped utterances, Włodarczak (2014) noted a tendency for overlaps at phrase-final positions to be associated with longer VTVs, that range between 220 ms and 500 ms. Włodarczak pointed out that temporal coordination at phrase-final positions is of a different nature than temporal coordination at phrase-medial positions, as the former reflects entrainment at prosodic levels higher than lexical stress, particularly, at prosodic boundaries. However, an important question that arises here is what kind of timing variation listeners use to predict the end of the phrase.

Entrainment to stress occurrences rate, suggested for overlaps phrase-medially, may not explain coordination with phrase boundaries. Overlaps at phrase boundaries may not be

driven by predicting semi-regular occurrences of salient beats; rather, there is an additional factor, which is the *extra* lengthening of VTVs at the phrase-final position.

A potential explanation of what listeners use to predict the end of the phrase may be inferred from speech segmentation literature. In particular, White et al. (2015) suggested that listeners segment the linear speech stream into words and phrases based on the experience of forgoing speech rate. That is, listeners generate expectations about segment duration with sufficient experience with utterance's speech rate. Timing deviation from expected unit duration, for example, due to phrase-final lengthening, can be interpreted by listeners as a linguistically meaningful cue to phrase boundaries. Furthermore, White et al. suggested that listeners may interpret deviation from expected unit duration through delays in the p-centre point relative to the listeners' expectation. Reflecting on overlapped phrase-final VTVs, lengthened vowels may lead to delays in the p-centre of the following syllable (in accordance with vocalic rhymes influences reported in the p-centre literature), which in turn will be interpreted as a prosodic boundary. Thus, potentially, overlaps that occurred later in VTVs, within the onset consonant(s) of the following syllable may reflect the effect of delayed p-centre on temporal coordination with phrase boundaries.

In summary, two types of timing variation may afford temporal coordination in natural dialogue, especially overlapped speech. The first is the rate of stress occurrences, and the second is segmental lengthening at prosodic phrase boundaries.

As temporal coordination is mediated by certain different temporal cues in the speech signal, we may predict that temporal coordination in natural dialogue between Hadari and Bedouin speakers may depend on the available durational cues that signal structure in each dialect. As the Hadari dialect has stronger temporal stress contrast, as found in Chapter 3, a Bedouin dialogue partner may entrain to the rate of stress occurrences to predict salient syllable beats to initiate an overlap. On the other hand, less temporal stress contrast in the Bedouin dialect may lead a Hadari dialogue partner to entrain to a stronger cue to structure, that is, lengthened segments at prosodic boundaries. These predictions may be supported by the findings that dialectal discrimination depends on the strongest available temporal cues that signal structure (White et al., 2012). These predictions may be tested in future research with natural speech corpus of Hadari and Bedouin Kuwaiti Arabic speakers, which may enhance our understanding of timing interaction in natural dialogue.

Future research on temporal coordination between dialects may also benefit from experimentally controlled tasks that may allow for the examination of the effect of specific timing cues in affording temporal coordination. For example, Rathcke et al. (2021) used a constrained tapping task in which metronomes were displayed throughout the test sentences, and English participants were asked to tap to prominent beats within the test sentences. The findings were that stressed syllables attracted more taps than unstressed syllables, with strong tendencies to align taps with vowel onsets. Interestingly, there were beat anticipation effects; at earlier phrase positions, participants' taps were earlier than stressed vowel onsets, while at later positions of the phrase, taps were more closely aligned with stressed vowel onsets. The asymmetry between earlier and later taps in the test phrases suggests that participants' use timing variation in the signal to predict salient points of synchronization. Such tapping tasks may be improved in future research, for example, by manipulating stress occurrences rate in order to calibrate specific timing effects that afford synchronization. Another promising controlled experimental task is the stop-signal paradigm suggested by Tilsen (2008). In this task, English speakers were presented with a signal at random points of the test sentences to stop speaking as fast as they could. The general findings were that lags between speech cessation and stop signals were larger when the stop signal occurred before a stressed syllable than before an unstressed syllable. This suggests stronger anticipation of stressed syllables in the speech signal than unstressed syllables. The advantage of the stop-signal task compared with tapping tasks, is that the former allows for direct examination of entrainment to speech structure in speech production. Different manipulation of the speech materials in the stop-signal task, e.g., stress rate, and final lengthening, could be made to examine the effect of specific temporal cues on temporal coordination and the anticipation of stressed syllable beats.

6.7 Summary

In speech cycling, we found variable stressed vowel onsets alignment between Hadari and Bedouin Kuwaiti dialects. In the external phase, Bedouin aligned heavy and light syllables similarly, around 0.5 phase, in the trochaic pattern in the fast rate, while Hadari aligned light syllables earlier than heavy syllables. In the internal phase, both dialects aligned heavy syllables closer to 0.5 phase than light syllables; however, Hadari aligned light syllables earlier in the phrase and closer to 0.5 than Bedouin. We interpreted the closer alignment of heavy syllables in both dialects, in both phase measures, as reflecting a top-down effect; due

to the phonological length of heavy syllables, they exhibit stronger stress contrast with unstressed syllables, which may attract closer alignment with harmonic phases. As for the earlier alignment of light syllables in Hadari, we asserted that it could be due to greater unstressed syllables reduction. Analysis of syllable duration revealed greater unstressed syllable reduction in Hadari than in Bedouin, which may lead to earlier phase alignment of light syllables alignment in Hadari (see Figure 6.1). In the analysis of the amplitude envelope, the power distribution metric, Centroid, reflected the dominance of stressed syllables time scale. We also interpreted the scores in the SBPr and Ratio21 metrics as reflecting variation in stressed syllable duration since the reference to stressed syllables timing better captures the syllabic durational patterns investigated in Experiment 1 (b). Thus, we did not find evidence to support a potential role of stress feet timing in temporal coordination in speech cycling.

The demonstrated effects of speech rate and syllable weight in speech cycling support the view that temporal coordination between different rhythmic time scales emerges as an affordance of temporal acoustic cues to speech structure (Cummins, 2009). We discussed patterns of temporal coordination between interlocutors in natural dialogue (e.g., Włodarczak, 2014) and suggested specific timing cues (speech rate, temporal stress contrast, and final lengthening) that afford the timing interaction between interlocutors. Following the potential effects of timing cues on the affordance of temporal coordination, we suggested a working hypothesis for exploring timing interaction in natural dialogue between Hadari and Bedouin speakers in future research. Specifically, in situations of natural dialogue between Hadari and Bedouin speakers, a listener may use the strongest available timing cues to speech structure in order to entrain to the speech of the dialogue partner.

Bibliography

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: University Press.

Abu-Haider, F. (2006). Bedouinization. In K. Versteegh, M. Eid, M. Woidich, A. Zaborski, & E. Alaa (Eds.), *The Encyclopaedia of Arabic Language and Linguistics* (pp. 269-274). Laiden, Brill.

Ahn, M. (2002*). Phonetic and functional bases of syllable weight for stress assignment* [Doctoral dissertation, University of Illinois].

Allen, G. D. (1972a). The location of rhythmic stress beats in English: an experimental study I. *Language and Speech*, *15*, 72-100.

Allen, G. D. (1972b) "The location of rhythmic stress beats in English: an experimental study II. *Language and Speech*, *15*, 179-195.

Allen, G. D. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, *5*, 75-86.

Almalki, H. (2020). *The Production and Perception of Prosodic Prominence in Urban Najdi Arabic* [Doctoral dissertation, George Mason University, Virginia].

Almbark, R., Bouchhioua, N., & Hellmuth, S. (2014). Acquiring the phonetics and phonology of English word stress: comparing learners from different L1 backgrounds. In *Proceedings of the International Symposium on the Acquisition of Second Language Speech: Concordia Working Papers in Applied Linguistics 2014, 5*, 19-35.

Alzaidi, M. (2014). *Information Structure and Intonation in Hijazi Arabic* [Doctoral dissertation, University of Essex].

Arantes, P. (2018). Slicer Praat script. Retrieved from: https://github.com/parantes/slicer.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46–63.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351-373.

Asu, E. L. & Nolan, F. (2006). Estonian and English rhythm: A two dimensional quantification based on syllables and feet. In *Proceedings of Speech Prosody 2006* (pp. 249-252). Dresden.

Barbosa, P.A. 2003. Beat extractor Praat script. Retrieved from: https://uk.groups.yahoo.com/neo/groups/praat-users/conversations/topics/921.

Barbosa, P.A., Arantes, P., Meireles, A.R., & Vieira, J.M. (2005). Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors. *In Proceedings of interspeech 2005,* 1441–1444. Lisbon.

Barkat, M., Hamdi, R., & Pellegrino, F. (2004). *De la caractérisation linguistique à l'identification automatique des dialectes arabes*. *MIDL 2004*. Paris.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? In *Proceedings of the 15th International Congress of Phonetic Sciences,* 2693–2696. Barcelona.

Beckman, M. (1982). Segment duration and the 'mora' in Japanese. *Phonetica*, *39*(2-3), 113-135.

Beckman, M. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457–463). Tokyo: OHM Publishing Co.

Beckman, M., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology* (pp. 152-178). Cambridge: Cambridge University Press.

Bell, A., & Fowler, A. (1984). Perception of the rhythm of English and of non-speech analogues. Paper presented to the 108th meeting of the Acoustical Society of America, Minneapolis, October 1984.

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, *37*(3), 237–250.

Biadsy, F., & Hirschberg, J. (2009). Using Prosody and Phonotactics in Arabic Dialect Identification. *In Proceedings of Interspeech 2009, 208-211*. Brighton, UK.

Bloomfield, P. (2000). *Fourier analysis of time series: an introduction*. John Wiley & Sons.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer. Version 6.0.43. http://www.praat.org/.

Bolinger, D. (1965). *Forms of English: Accent, Morpheme, Order*. Cambridge, Massachusetts: Harvard University Press.

Bouchhioua, N. (2008). Duration as a cue to stress and accent in Tunisian Arabic, Native English and L2 English. *In Proceedings of Speech Prosody 2008* (pp. 535-538). Campinas, Brazil.

Bouzon, C., & Hirst, D. (2004). Isochrony and prosodic structure in British English. In *Proceedings of Speech Prosody 2004* (pp. 223-226). Nara, Japan.

Bowman, D. C., & Lees, J. M. (2013). The Hilbert–Huang transform: A high resolution spectral method for nonlinear and nonstationary time series. *Seismological Research Letters*, *84*(6), 1074-1080.

Bruggeman, A. (2018). *Lexical and postlexical prominence in Tashlhiyt Berber and Moroccan Arabic* [Doctoral dissertation, University of Cologne].

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149-180.

Chahal, D. (2001). *Modelling the intonation of Lebanese Arabic using the autosegmental-metrical framework: a comparison with English* [Doctoral dissertation, University of Melbourne].

Chen, G. (2017). Spectral balance Praat script. Retrieved from: https://github.com/chengafni/praat/blob/master/plugin_SpectralEmphasis.zip

Classè, A. (1939). *The rhythm of English prose*. B. Blackwell.

Crystal, J. (1990). *Oil and politics in the Gulf: Rulers and merchants in Kuwait and Qatar*. Cambridge University Press.

Cumming, R. E. (2011a). Perceptually informed quantification of speech rhythm in pairwise variability indices. *Phonetica*, *68*(4), 256–277.

Cumming, R. E. (2011b). The language-specific interdependence of tonal and durational cues in perceived rhythmicality. *Phonetica*, *68*(1-2), 1-25.

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*, 121–126. Aix-en-Provence.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics, 31*(2), 139-148.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16-28.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145-171.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical

access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113-121.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics*, *25*(3), 287–312.

Darwin, C. J., & Donavan, A. (1980). Perceptual studies of speech rhythm: Isochrony and intonation. In J. C. Simon (Ed.), *Spoken Language Generation and Understanding* (pp. 77-85). Dordrecht, Holland: D. Riedel Publishing Company.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*, 51–62.

De Jong, K. (1994). The correlation of p-center adjustments with articulatory and acoustic events. *Perception & Psychophysics*, *56*(4), 447-460.

De Jong, K., & Zawaydeh, B. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics, 27*(1), 3-22.

De Jong, K., & Zawaydeh, B. (2002). Comparing stress, lexical focus, and segmental focus: Patterns of variation in Arabic vowel duration. *Journal of Phonetics, 30*(1), 53-75.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics, 4*, 183-198.

Dell, F. & ElMedlaoui, M. (2002). *Syllables in Tashlhiyt Berber and in Moroccan Arabic*. Dordrecht: Kluwer.

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence* [Doctoral dissertation, Bonn University].

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and

speech rate. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 471–474). Barcelona.
Eriksson, A. (1991). *Aspects of Swedish speech rhythm* [Doctoral dissertation, University of Göteborg].

Farwaneh, S. (1995). *Directionality Effects in Arabic Dialect Syllable Structure* [Doctoral dissertation, University of Utah].

Faure, G., Hirst, D. J., & Chafcouloff, M. (1980). Rhythm in English: Isochronism, pitch, and perceived stress. In Linda R. Waugh, & C. H. van Schoonevald (Eds.), *The melody of language* (pp. 71-79). Baltimore: University Park Press.

Fraisse, P. (1956). *Les structures rythmiques*. Paris.

Fraisse, P. (1978). Time and rhythm perception. In C. Edward, C. Carterette & M. P. Friedman (Eds.), *Handbook of Perception* (VIII, pp. 203—254). New York: Academic Press.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.*), The Psychology of Music* (pp. 149-180). New York: Academic Press.

Frota, S., & Vigário, M. (2001). On the Correlates of Rhythmic Distinctions: The European/Brazilian Portuguese Case". *Probus* 13(2), 247-273.

Gibbon, D. (2003). Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of the 15th international congress of phonetic sciences*, 2489–2492. Barcelona.

Gibbon, D. (2006). Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In *Methods in Empirical Prosody Research* (Originally published 2006 ed., Vol. 3, pp. 181-210). Berlin, Boston: DE GRUYTER.

Grabe, E., & Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven, & Natasha Warner (Eds.),

*Papers in laboratory phonology (*VII, pp. 515-546). Berlin and New York: Mouton de Gruyter.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*(3-4), 465-485.

Gussenhoven, C., & Rietveld, A. C. M. (1992). Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics*, *20*(3), 283–303.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model for phase transitions in human hand movements. *Biological Cybernetics 51*, 347-356.

Hamdi, R., Barkat-Defradas, M., Ferragne, E., & Pellegrino, F. (2004). Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches. *In Proceedings of Interspeech 2004*. Jeju Island, Korea.

Hamdi, R., Ghazali, S., & Barkat-Defradas, M. (2005). Syllable structure in spoken Arabic: a comparative investigation. *In Proceedings of Interspeech 2005* (pp. 2245-2248). Lisbon.

Hayes, B. (1985). *A metrical theory of stress rules*. New York, NY: Garland.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics, 31*(1), 39-62.

Heselwood, B. (2004). The 'tight approximant'variant of the Arabic 'ayn. *Journal of the International Phonetic Association*, *37*(1), 1-32.

Hirst, D. (2009). The rhythm of text and the rhythm of utterances: from metrics to models. In *Proceedings of Interspeech 2009* (pp. 1519-1522). Brighton.

Hoequist, C. (1983). The perceptual centre and rhythm categories. *Language and Speech, 26*(4), 367-376.

Holes, C. (2006). Kuwaiti Arabic. In K. Versteegh, M. Eid, M. Woidich, A. Zaborski, & E. Alaa (Eds.), *The Encyclopaedia of Arabic Language and Linguistics* (2, pp. 608-620). Leiden: Brill.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. *In Proceedings of the 10th International Congress of Phonetic Sciences,* pp. 429–433. Utrecht.

Howell, P. (1988a). Prediction of P-center location from the distribution of energy in the amplitude envelope. I. *Perception & Psychophysics, 43*(1), 90-93.

Howell, P. (1988b). Prediction of P-center location from the distribution of energy in the amplitude envelope. II. *Perception & Psychophysics, 43*(1), 99.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences 454*, 903-995.

Ingham, B. (1994). *Najdi Arabic central Arabian*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Jassem, W., Hill, D. R., & Witten, I. H. (1984). Isochrony in English speech: Its statistical validity and linguistic relevance. In D. Gibbon, & H. Richter (Eds.), *Intonation, accent and rhythm: studies in discourse phonology* (pp. 203- 225). Berlin: Walter de Gruyter.

Jones, D. (1942). Chronemes and tonemes. *Acta Linguistica 3*, 1-10.

Katz, J., Chemla, E., & Pallier, C. (2015). An attentional effect of musical metrical structure. *PLOS ONE, 10*(11), E0140895.

Kelly, Niamh E. (2021). Phrase-final lengthening across segments in Lebanese Arabic. *In 179th Meeting of the Acoustical Society of America 2021, 42*(1).

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, Mass. MIT Press.

Kelso, J. A. S., Southard, D., & Goodman, D. (1979). On the nature of human interlimb coordination. *Science, 203*, 1029-1031.

Kim, D., & Oh, H. (2009). EMD: A package for empirical mode decomposition and Hilbert spectrum. *The R Journal*, *1*, 40-46.

Kim, D., Kim, K. O., & Oh, H. S. (2012). Extending the scope of empirical mode decomposition by smoothing. *EURASIP Journal on Advances in Signal Processing*, *2012*(1), 1-17.

Kim, H., & Cole, J. (2005). The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase. *In Proceedings of Interspeech 2005* (pp. 2365-2368). Lisbon.

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

Krivokapić, J. (2013). Rhythm and convergence between speakers of American and Indian English. *Laboratory Phonology 4*(1). 39–65.

Lea, W. A. (1974). *Prosodic Aids to Speech Recognition. IV. A General Strategy for Prosodically-Guided Speech Understanding*. Technical report PX10791, Sperry Univac, DSD, St Paul, Minnesota.

Leenberg, E. H. (1967). *Biological foundations of language*. John Wiley and Sons.

Lehiste, I. (1972). The role of temporal factors in the establishment of linguistic units and boundaries. In W. U. Dressler, & F. V. Mares (Eds.), *Phonologica* (pp. 115-122). Munich-Salzburg: Verlag.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*, 253–263.

Liberman, M. (1975). *The intonational system of English* [Doctoral dissertation, Cambridge, MA].

Lloyd James, A. (1929). *Historical introduction to French Phonetics*. London: ULP.

Lloyd James, A. (1940). *Speech Signals in Telephony*. London: Pitman.

Loukina, A., Kochanski, G., Rosner, B., & Keane, E. (2011). Rhythm Measures and Dimensions of Durational Variation in Speech. *The Journal of the Acoustical Society of America, 129*(5), 3258-3270.

Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, *43*(4), 377-401.

Manrique, A. M. B., & Signorini, A. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics*, *11*(2), 117-128.

Map of Kuwait, Middle East. Retrieved from: https://www.nationsonline.org/oneworld/map/kuwait-map.htm

Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics, 30*(3), 247-256.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan, & K. Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition* (pp. 101–116).

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review*, *83*, 405–408.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica, 38*(1–3), 84–105.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris Publications.

Nolan, F., & Jeon, H-S. (2014). Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B*, *369*(1658), 20130396.

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth International Congress of Phonetic Sciences 2* (pp. 1075–1078).

O'Dell, M. L., & Nieminen, T. (2009). Coupled oscillator model for speech timing: Overview and examples. *Nordic Prosody: Proceedings of the 10th Conference* (pp. 179-190). Helsinki.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America, 54*(5), 1235–1247.

Pike, K. L. (1946). *Intonation of American English*. Ann Arbor: University of Michigan.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics, 8*(3), 293–304.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-centre phenomenon. *Journal of Phonetics*, *17*, 175–192.

Port, R. F. (1981) Linguistic timing factors in combination, *Journal of the Acoustical Society of America*, *61*(1), 262-274.

Prieto, P., Vanrell, 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.

Quenè, H. (2005). Metronome Praat script. Retrieved from: https://www.hugoquene.nl/tools/index.html

Quenè, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics, 35*(3), 353-362.

R Development Core Team. (2019). R: A language and environment for statistical computing, R foundation for statistical computing. Vienna, Austria. http://www.Rproject.org/

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In *Proceedings of speech prosody 2002* (pp. 115–120). Aix-en-Provence.

Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 337–342). Barcelona.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265-292.

Rathcke, T., Lin, C., Falk, S., & Bella, S. (2021). Tapping into linguistic rhythm. *Laboratory Phonology, 12*(1), 1-32.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic Controversies* (pp. 73–79). London: Edward Arnold.

Rosario-Martinez, H., Fox, J., R Core Team. (2015). Package 'phia'. *CRAN Repos*. *https://cran.r-project.org/web/packages/phia/index.html*

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367-373.

Rosenhouse, J. (2006). Bedouin Arabic. In K. Versteegh, M. Eid, M. Woidich, A. Zaborski, & E. Alaa (Eds.), *The Encyclopaedia of Arabic Language and Linguistics* (pp. 259-269). Laiden, Brill.

Sagher, M.A. (2004). *The impact of economic activities on the social and political structures of Kuwait* [Doctoral dissertation, Durham University].

Saltzman, E. L., Nam, H., Krivokapi´c, J., & Goldstein, L. (2008). A task-dynamic toolkit for modelling the effects of prosodic structure on articulation. In *Proceedings of Speech Prosody 2008* (pp. 175-184). Campinas, Brazil.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics*, *13*(2), 155-162.

Scott, S. K. (1993). *Perceptual centres in speech: an acoustic analysis* [Doctoral dissertation, University College London].

Selkirk, E.O. (1986). On derived domains in sentence phonology. *Phonology Yearbook 3*, 371-405.

Shen, Y., & G.G. Peterson. (1962). Isochronism in English. *Studies in Linguistics, Occasional Papers 9*. University of Buffalo.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*. https://CRAN.Rproject.org/package=afex

Sleimen-Malkoun, R., Temprado, J., & Hong, S. (2014). Aging induced loss of complexity and dedifferentiation: Consequences for coordination dynamics within and between brain, muscular and behavioral levels. *Frontiers in Aging Neuroscience, 6*, 140.

Sluijter, A., & 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.

Steele, J. (1779). *Prosodia Rationalise: Or, an Essay Towards Establishing the Melody and Measure of Speech to Be Expressed and Perpetuated by Peculiar Symbols*. London: J. Nichols.

Šturm, P., & Volin, J. (2016). P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics, 55*, 38-52.

Sueur, J., Aubin, T., Simonis, C. (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*, 213-226.

Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems* (2nd ed., Cambridge handbooks for language teachers). Cambridge; New York: Cambridge University Press.

Tajima, K. (1999) *Speech rhythm in English and Japanese: Experiments in Speech Cycling* [Doctoral dissertation, Indiana University].

Taqi, H. (2010). *Two ethnicities, three generations: phonological variation and change in Kuwait* [Doctoral dissertation, Newcastle university].

Tilsen, S. (2006). Rhythmic coordination in repetition disfluency: a harmonic timing effect. *UC-Berkeley Phonology Lab Annual Report* (pp. 73-114). UC Berkeley.

Tilsen, S. (2008). Preliminary results of a stop-signal experiment. *UC Berkeley Phonology lab annual report* (pp. 686-712). UC Berkeley.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America, 134*(1), 628-639.

Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America, 124*(2), EL34-EL39.

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.

Tuller, B., & Kelso, J. A. S. (1989). Environmentally-specified patterns of movement coordination in normal and split-brain subjects. *Experimental brain research, 75*, 306-316.

Turk, A., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*(4), 397-440.

Turk, A., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445-472.

Turk, A., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics, 27*(2), 171-206.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic Segment Durations in Prosodic Research: A Practical Guide. In S. Sudhoff, D. Lenertovà, R. Mayer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schleiber (Eds.), *Methods in Empirical Prosody Research* (3, pp. 1-28). Berlin, Boston: DE GRUYTER.

Van Heuven, V. J. (2018). Acoustic correlates and perceptual cues of word and sentence stress. In R. Goedemans, J. Heinz, & H. van der Hulst (Eds.), *The Study of Word Stress and Accent: Theories, Methods and Data (*pp. *15-59)*. Cambridge: Cambridge University Press.

Van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication, 11*(6), 513-546.

Van Santen, J. P. H., & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *The Journal of the Acoustical Society of America, 107*(2), 1012-1026.

Villing, R., Timoney, J., & Ward, T. E. (2004). Automatic blind syllable segmentation for continuous speech. *In Irish Signals and Systems Conference 2004*.

Vogel, I., Athanasopoulou, A., & Pincus, N. (2017). Acoustic properties of prominence and foot structure in Jordanian Arabic. In H. Ouali (Ed.), *Perspectives on Arabic Linguistics* (XXIX, pp. 55-88). John Benjamins.

Wagner, P., Ćwiek, A., & Samlowski, B. (2019). Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies. *Journal of Phonetics, 76*, 100911.

Wagner, P., Malisz, Z., Inden, B. & Wachsmuth, I. (2013). Interaction phonology – A temporal co-ordination component enabling representational alignment within a model of communication. In I. Wachsmuth, J. de Ruiter, P. Jaecks, & S. Kopp (Eds.), *Alignment in Communication: Towards a New Theory of Communication* (pp. 109-132). Amsterdam: John Benjamins.

Warner, N., & Arai, T. (2001). Japanese mora-timing: a review. *Phonetica, 58*(1–2), 1–25.

Watson, J. C. E. (2011a). Arabic dialects (general article). In S. Weninger, G. Khan, M. Streck, & J. Watson (Eds.), *The Semitic Languages: An International Handbook* (pp. 851-896). Boston: De Gruyter Mouton.

Watson, J. C. E. (2011b). Dialects of the Arabian Peninsula. In S. Weninger, G. Khan, M. Streck, & J. Watson (Eds.), *The Semitic Languages: An International Handbook* (pp. 897-908). Boston: De Gruyter Mouton.

Watson, J. C. E. (2011c). Word stress in Arabic. In M. Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (5, pp. 2990-3018). Oxford: Blackwell.

White, L. (2002). *English speech timing: a domain and locus approach* [Doctoral dissertation, University of Edinburgh].

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication, 63-64*(Sep), 38-54.

White, L., & Mattys, S. L. (2007a). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501-522.

White, L., & Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and Prosodic Issues in Romance Phonology* (pp. 237–257). Amsterdam: John Benjamins.

White, L., & Turk, A. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*, 459-471.

White, L., Mattys, S., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665-679.

White, L., Mattys, S., Stefansdottir, L., & Jones, V. (2015). Beating the bounds: Localized timing cues to word segmentation. *The Journal of the Acoustical Society of America, 138*(2), 1214-1220.

White, L., Payne, E., & Mattys, S.L. (2009). Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In M. Vigario, S. Frota & M.J. Freitas (Eds.), *Phonetics and Phonology: Interactions and Interrelations* (pp. 137-158). Amsterdam: John Benjamins.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. (2010). How stable are acoustic metrics of contrastive speech rhythm? *The Journal of the Acoustical Society of America*, *127*(3), 1559-1569.

Wilson, M., & Wilson, T. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957-968.

Windmann, A. (2016). *Optimization-based modeling of suprasegmental speech timing* [Doctoral dissertation, Bielefeld University].

Włodarczak, M., Simko, J., & Wagner, P. (2012a). Temporal entrainment in overlapped speech: Cross-linguistic study. *In Proceedings of Interspeech 2012*. Portland, OR.

Włodarczak, M., Simko, J., & Wagner, P. (2012b). Syllable-boundary effect: Temporal entrainment in overlapped speech. *Proceedings of Speech Prosody 2012*, pp. 611–614. Shanghai, China.

Włodarczak, M. (2014). *Temporal coordination in overlapping speech* [Doctoral dissertation, Bielefeld University].

Woodrow, H. (1951). Time perception. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1224- 1236. New York: Wiley.

Yamnishi, J., Kawato, M., & Suzuki, R. (1980). Two coupled oscillators as a model for the coordinated finger tapping by both hands. *Biological Cybernetics, 37*, 219-225.

Yeou, M., Embarki, M., & Al-Maqtari, M. (2007). Contrastive focus and F0 patterns in three Arabic dialects. *Nouveaux cahiers de linguistique française 28*, 317–326.

Zawaydeh, B.A., Tajima, K., & Kitahara, M. (2002). Discovering Arabic rhythm through a speech cycling task. In D. Parkinson, & E. Benmamoun (Eds.), *Perspectives in Arabic Linguistics* (XIII-XIV, pp. 39-58). California: John Benjamins.