Methods for investigation of L2 speech rhythm: Insights from the production of English speech rhythm by L2 Arabic learners

Abstract

Rhythm metrics can detect L2 development of target-like speech rhythm but the mapping of metrics to underpinning phonological features is indirect. Results from rhythm metrics are thus hard to interpret, and even harder to translate into practical priorities for learners. We investigate speech rhythm in L1 Arabic/L2 English which differ in properties which contribute to the percept of rhythm: unstressed vowel reduction and phonemic vowel length. Our production data are examined using local measures of stressed and unstressed vowels and evaluation of consonant cluster realization, alongside standard rhythm metrics; this combination facilitates disambiguation of competing interpretations of metric results. The findings confirm the importance of using multiple rhythm metrics to study L2 speech rhythm and demonstrate how local measures can guide interpretation of their results. The results for L2 Saudi Arabic learners support prioritization of practice in reduction of unstressed vowels for those seeking to develop target-like SSBE speech rhythm.

Keywords: rhythm, metrics, Arabic, English, unstressed vowels, vowel reduction.

Methods for investigation of L2 speech rhythm: Insights from the production of English speech rhythm by L2 Saudi learners

Background

Rhythm in Speech

Rhythm in speech has long been debated among linguists: whether languages vary in their rhythmic properties, and, if so, how that variation can be captured and measured. Early attempts suggested a strong tendency in English for stressed syllables to occur at regular or equal intervals (i.e., isochrony) (Jones, 1962). The concept was then used by Lloyd James (1940, cited in Pike, 1945) to describe English and French as having 'Morse-code rhythm' versus 'Machine-gun rhythm', respectively. Pike (1945) then coined the terms 'stress-timed' and 'syllable-timed' to describe rhythm in languages. Notably, he mentioned that all languages appeared to display both kinds of rhythm but differ in that they may favor one more than the other. Subsequent studies did not find concrete evidence of isochrony in the speech signal (e.g., Dauer, 1983; Roach, 1982; Wenk and Wioland, 1982). For this reason, it was suggested that rhythm is instead a perceptual phenomenon (e.g., Allen, 1975; Donovan and Darwin, 1979; Lehiste, 1977). However, the question of how rhythmic variation between languages could be measured was not answered.

Roach (1982) and Dasher and Bolinger (1982) suggested that auditory classification of languages into 'stress-timed' and 'syllable-timed' might be attributable to differences that those languages exhibit in degree of complexity of syllable structure, and in existence of vowel length distinctions and/or reduction of unstressed syllables. It was suggested that 'stress-timed' languages, such as English, German and Dutch, tend to have more complex syllable structure and are more likely to exhibit vowel reduction than 'syllable-timed'

languages, such as French, Chinese and Italian. This suggestion was elaborated further by Dauer (1983), who proposed that speech rhythm is not a phonetic feature or a phonological primitive, but rather a manifestation of multiple phonological features, namely, stress, vowel reduction and syllable structure. Dauer maintained that all languages are more or less stress-based and cannot be divided into two dichotomous rhythmic types.

An acoustic implementation of this phonological stance on speech rhythm was first put forth by Ramus et al. (1999), who support the phonological basis for classifying languages rhythmically, but propose that the resulting phonetic timing differences are independently measurable. Previous experiments showed that neonates are able to discriminate between two languages conventionally classified into two different rhythmic types relying merely on rhythmic cues (Nazzi et al., 1998). Ramus et al. (1999) argued that infants cannot be using complex language-specific phonological concepts to segment speech, but rather the succession of vowels of variable durations separated by unanalyzed speech segments; a similar view was expressed in Mehler et al. (1996). Thus, Ramus et al. (1999) combined the phonological explanation of rhythmic typology (Dauer, 1983) with the simpler task of segmenting speech into vowels and consonants (Mehler et al., 1996; Nazzi et al., 1998) to propose a new acoustic quantification of rhythmic typology.

Ramus et al. (1999) proposed three acoustic metrics of rhythm: %V, percentage of the total duration of vocalic intervals; ΔV, standard deviation of the duration of vocalic intervals; and ΔC, standard deviation of the duration of consonantal intervals. The authors hypothesized that ‘syllable-timed’ languages would display lower ΔV and ΔC values than ‘stress-timed’ languages because ‘stress-timed’ languages tend to show more durational variation between consonantal intervals (due to complexity of consonantal clusters) and

between stressed and unstressed vowels (due to shortening of unstressed vowels). %V was hypothesized to be higher in 'syllable-timed' languages than in 'stress-timed' languages for the same reasons as for ΔV. A combination of %V and ΔC was found to be successful in distinguishing languages belonging to different rhythm classes.

Low et al. (2000) proposed a rate-normalized acoustic measure to capture the durational differences between successive vocalic segments: the normalized pairwise variability index (nPVI-V). The measure was shown to be more effective than the metrics suggested by Ramus et al. (1999) in capturing the rhythmic difference between British English and Singapore English. Grabe and Low (2002) also devised a similar, non-normalized, variability index for successive consonantal intervals (rPVI-C). The rPVI-C did not achieve the same success, though was helpful in explaining results for languages that do not fit neatly into either of the two traditional rhythmic classes.

Because of the negative effect of speech rate on both ΔC and ΔV, Dellwo (2006) and White and Mattys (2007a) developed variation coefficients for intervocalic and vocalic durations, respectively. Dellwo (2006) proposed the VarcoC measure as an alternative to ΔC and White and Mattys (2007a) proposed the VarcoV measure as an alternative to ΔV. Both proved more successful in distinguishing the languages examined.

Arabic and English are both traditionally described as 'stress-timed' languages (Abercrombie, 1976; Miller, 1984). Unlike Arabic, English speech rhythm has been studied extensively using rhythm metrics. English is widely considered an archetypical 'stress-timed' language, thus assumed to show relatively higher durational variability between vocalic segments and between consonantal segments (e.g., Grabe and Low, 2002; Ramus et al., 1999; White and Mattys, 2007a). However, some dialectical variation has been

observed between English dialects (e.g., White et al., 2012). Few studies have examined rhythmic variation in Arabic using metrics (e.g., Hamdi et al., 2004; Ghazali et al., 2002), but it was generally found that Western Arabic dialects (e.g., Moroccan) are more 'stress-timed' than Eastern dialects (e.g., Jordanian). Due to the absence of any prior study on Saudi Arabic speech rhythm, we include samples of L1 Saudi Arabic in the present study to facilitate interpretation of the results in terms of potential L1 transfer.

Many studies have examined the success, stability and reliability of rhythm metrics (e.g., Arvaniti, 2012; Knight, 2011; White & Mattys, 2007a; Wiget et al., 2010). The potential of the metrics to classify languages into traditional rhythm classes is generally agreed to be weak, but their capacity to distinguish languages and dialects is acknowledged. The rhythm metrics provide a useful quantitative tool to study acquisition and production of speech rhythm by L2 learners, especially in the absence of any other reliable rhythm measures, and given the importance of speech rhythm to L2 speech (e.g., Li and Post, 2014; Ordin and Polyanskaya, 2014; Behrman et al., 2019).

L2 Speech Rhythm

Since the introduction of the acoustic rhythm metrics, numerous studies have used them to examine production and acquisition of L2 speech (Behrman et al., 2019; Benet et al., 2012; Carter, 2005; Grenon and White, 2008; Gut, 2003; Li and Post, 2014; Mok and Dellwo, 2008; Polyanskaya and Ordin, 2014; Schaeffler, 2001; Stockmal, Markus and Bond, 2005; White and Mattys, 2007a; White and Mattys, 2007b; White and Mok, 2019).

Schaeffler (2001) evaluated the usefulness of rhythm metrics to study L2 speech rhythm. He used PVI, ΔV, %V, and ΔC to examine production of German rhythm by L2 Spanish speakers. Apart from ΔC, the metrics were successful in distinguishing the L1 and

L2 speakers. Carter (2005) examined the rhythm of Hispanic English using PVI-V, and showed that Hispanic English was rhythmically intermediate between L1 English and Spanish. Stockmal et al. (2005) used ΔV, %V, nPVI-V, ΔC and rPVI-C to study acquisition of Latvian rhythm by L2 Russian learners, divided into two groups based on self-reported proficiency. Both L2 groups yielded similar %V, ΔV, and nPVI-V scores to L1 Latvian speakers. The low proficiency L2 speakers showed higher scores of ΔC and rPVI-C than all other speaker groups, which was ascribed to their slow and careful delivery in L2 speech.

White and Mattys (2007a) tested the effectiveness of the metrics (ΔV, %V, ΔC, nPVI-V, rPVI-C, VarcoV and VarcoC) for detecting influence of L1 on production of L2 rhythm. They examined L1 and L2 speech samples of English, Dutch and Spanish, and found VarcoV, %V, and nPVI-V to be the most discriminative measures. The authors also suggested that a combination of VarcoV and %V provides particular insight for study of L2 rhythm. For instance, for English, the L2 Dutch speakers had more similar scores for VarcoV and %V to those of the L1 English speakers than the L2 Spanish speakers. Grenon and White (2008) examined production of Japanese and English speech rhythm by L2 English and L2 Japanese speakers, respectively, using %V, VarcoV and rPVI_C. The metrics captured some aspects of L2 rhythm, but the authors cautioned that the results require careful interpretation, taking the phonological systems of the languages under study into account. Mok and Dellwo (2008) used various rhythm metrics to examine English speech rhythm produced by L2 Chinese speakers. As for White and Mattys (2007a) and Grenon and White (2008), VarcoC and %V were found to be most effective in classifying Chinese-accented English.

More recently, Ordin and Polyanskaya (2014) used rhythm metrics to examine acquisition of English speech rhythm in longitudinal data from two Punjabi and two Italian L2 learners (both 'syllable-timed' languages). The metrics were successful in capturing development of more English-like rhythmic properties over time. Li and Post (2014) studied acquisition of English speech rhythm by L2 German and Mandarin speakers, and the vocalic metrics captured the developmental path of the L2 speakers. Finally, Behrman et al (2019) studied English speech rhythm of L2 Spanish speakers and found that %V, VarcoV, and VarcoC distinguished L1 and L2 speech only in conversational speech, not careful speech.

These previous studies show that the vocalic rhythm metrics can be used successfully to study acquisition and production of L2 speech rhythm. The consonantal rhythmic metrics have not achieved the same level of success but may still shed light on the complexity of consonant clusters in produced by L2 learners. Because of the sensitivity of the rhythm metrics to speaking styles and data collection methods, their scores should be taken as indicative rather than absolute measures of rhythm (Grenon and White, 2008). Given this restriction, it is possible that, in addition to these global rhythm measures, local measures of individual vowels and consonants may contribute to a better account of L2 speech rhythm.

The current study examines the acquisition of L2 English speech rhythm by L2 Saudi learners, divided into two groups based on their length of residence in the UK. The study contributes to the growing literature on L2 speech rhythm in three ways. First, it examines English speech rhythm of an L2 population not examined before (Arabic learners of English), in languages which differ along key phonological parameters relevant to the

global percept of speech rhythm. Second, it examines the effect of length of residence, as a rough index of L2 experience, on the production of English speech rhythm. Third, it goes beyond use of rhythm metrics alone to also examine vowel reduction and consonant clusters realization, as the assumed phonological building blocks of speech rhythm.

Method

Speakers

Three groups of speakers were recruited. The first group were six L1 speakers of Standard Southern British English (SSBE) aged 20-40, drawn from students and staff at the University of [xxx]. The second group were 12 L2 speakers of English, aged 19-32, who were L1 speakers of Najdi Saudi Arabic (NSA), a colloquial variety of Arabic spoken in the central region of Saudi Arabia. They were recruited among international Saudi students in the UK, with length of residence in the UK ranging from one to five years. The third group were six L1 speakers of NSA, aged 19-32, selected randomly from the same population of L2 NSA speakers of English (the second group); their length of residence in the UK ranged from one to three years.

The 12 L2 Saudi speakers were sub-divided into two groups, labelled ‘more experienced’ (ME) and ‘less experienced’ (LE), based on the number of years spent in the UK. The ME L2 speakers were six university students in the UK, aged 27-32, who had spent from two and half to five years in the UK. The LE L2 speakers, aged 19-32, were six English language students, who had spent one year in English language schools in [xxx], UK. Length of residence (LoR) was used in many previous studies as an index of L2 experience, even though it provides only a rough measurement of L2 experience, since longer residency does not always entail greater language experience (Piske et al., 2001).

Nevertheless, LoR arguably provides a more objective measure of L2 experience than L2 speakers' self-reported language use. For convenience and consistency with prior studies, the current study classified the LE and ME groups based solely on their LoR in the UK.

SSBE is taken as the language model for the L2 speakers because they all learned English in formal situations in English language schools in the UK accredited by the British Council, in which the variety of English taught is assumed to be closer to standard English than to regional UK varieties. In addition, it was not possible to select a single regional language model for the L2 speakers, as they were residents of different cities in the UK.

The sample size of speakers in the present study was restricted to six speakers per group due to the time needed for manual speech segmentation. All were male speakers to control for gender effects. None reported any hearing or speech impairments. All were paid an honorarium for participation. The participants form four speaker groups: SSBE, NSA, ME L2, and LE L2.

Materials and Procedure

L2 speech elicited by direct pronunciation assessment tasks, such as sentence reading, may encourage L2 speakers to monitor their speech production. This could lead to underestimation of the extent of L1 transfer or phonetic variation observed in speech produced under more natural conditions. However, read speech tasks provide control over lexical content and phonetic/phonological features to be examined, and also maximize comparability of speech samples across speakers. One way to avoid self-monitored speech while keeping the advantages of controlled speech is to place a moderate cognitive load on participants. In this way, they are more preoccupied with composing the message than with monitoring their pronunciation accuracy. The current study used an elicitation method

(adopted from Author XXX) which offers control over the content and lexical items in the utterances of the L2 speakers, but deflects them from monitoring their L2 speech production. However, it should be emphasized that the speech elicited is still read speech, and may not fully represent the speech of L2 speakers in natural settings.

The L1 and L2 English speakers were asked to paraphrase ten English sentences. They were first asked to write a paraphrase in response to a written prompt word, then after writing each paraphrase, they were asked to read it aloud twice into a microphone at normal speech pace. An example is given in (1).

(1) Example Stimulus: One of the developed countries in the world is Japan.

Prompt Word: Japan……

Paraphrase response: Japan is one of the developed countries in the world.

The second rendition was analysed only when hesitation or disfluency affected the first. Although the L2 speakers all had sufficient proficiency in English to engage in university studies, they were invited to stop the test at any time to ask what a certain word meant or how it should be pronounced. The paraphrase task was designed to be difficult enough to engage the L2 participants and deflect their attention from focusing on their pronunciation while reading. The test was also time-constrained, with 15 minutes per participant. Although the paraphrase task is not strictly needed for L1 English speakers, it was considered preferable to elicit all the English speech samples under the same conditions. Another advantage of the paraphrase task is that writing the paraphrase

sentences out first familiarizes speakers with the sentence to be read, and should reduce the occurrence of pauses and hesitations which affect the rhythmic flow of utterances.

Elicitation of NSA speech samples was designed in the context of Arabic diglossia . Reading and writing in Arabic colloquial varieties such as NSA is unnatural to L1 Arabic speakers; reading/writing are associated with Standard Arabic, which is not used in daily conversation and is phonologically distinct from colloquial varieties. Therefore, when constructing Arabic sentences to be read by the L1 NSA speakers, one must consider the possibility that they may lean towards reading the sentences in Standard Arabic.

Ten NSA sentences were constructed by the first author who is an L1 speaker of Saudi Arabic. Prior to recording, the sentences were checked verbally with three of the NSA speakers (from among the participants sample), who confirmed that the sentences sounded natural and acceptable in NSA. To avoid the sentences being read in Standard Arabic, the first author read the full set of sentences aloud in Saudi colloquial Arabic to each speaker at the start of each session, to provide an example of the speech register to be used. Reading all the sentences at once avoids biasing participants' responses towards imitation of the model rendition of the sentences; by the end of reading the last sentence they would have forgotten the acoustic detail of how the first one was produced. The NSA speakers then read each sentence twice at a normal speech rate in their own dialect.

Most of the speakers were recorded in a sound-treated phonetics laboratory at the University of [xxx]. Four NSA speakers were not resident in [xxx], so were recorded in a quiet furnished room in each participant's home, using a Marantz PMD660 digital recorder with Shure SM10A-CN headset condenser microphone. All recordings were digitised at 16 bit with 44.1 kHz sampling frequency, then transferred into computer memory for analysis.

The construction of the English and Arabic target sentences (provided in the online supplementary materials) was not random. An attempt was made to make the sentences representative of the phonological and metrical features relevant to rhythm for both English and NSA. This was achieved by including all permissible syllable structures and all possible degrees of vowel reduction (secondary stress, function words, and schwas) within the full set of sentences for each language, and by avoiding consecutive syllables that carry primary stress. The latter step does not reflect natural speech, where two stressed syllables may follow each other, with the clash potentially resolved by assigning more prominence to one of them (Nespor and Vogel, 1989). However, because one of the objectives of the current research is to examine how L2 speakers temporally differentiate stressed and unstressed vowels, consecutive stressed syllables were avoided. In the English sentences, multisyllabic words expected to contain schwa were also included to examine how the L2 speakers produced target syllables containing schwa and unstressed function words, in terms of degree of vowel reduction.

The total number of syllables in the English and the NSA target sentences was designed to be equal (155 syllables in each), to reduce the influence of speech rate on the rhythm metrics (Ramus et al., 1999); although some metrics are normalized for speech rate they may not eliminate the effect of speech rate completely (White and Mattys, 2007a). Based on citation forms of the words in the sentences, the average number of syllables per sentence in each language was 15.5, ranging from 13-21 for NSA, and 9-17 for English. Mean sentence duration was 2.02 seconds for NSA and 2.5 for English. The difference in mean sentence duration between languages may be because NSA syllable structure is

simpler than that of English (Ingham, 1994), with greater preponderance of CV syllables in NSA than in English.

Segmentation

Segmenting speech into phonemic categories is a complex process (Laver, 1994); clear segmentation protocols are necessary for consistency within a study and for comparison of results between studies. Following generally accepted criteria (Peterson and Lehiste, 1960; Turk et al., 2006), all utterances were manually segmented and labelled by the first author into vocalic and intervocalic (i.e. consonantal) intervals, based on auditory impression and visual inspection of waveforms and spectrograms in Praat (Boersma and Weenink, 1992-2018). Vocalic intervals are defined as the stretch of speech from the onset of a vowel to its offset, and consonantal intervals are defined as the stretch of speech from the offset of a vowel to the onset of the next vowel, regardless of the number of intervening consonants (Grabe and Low, 2002). The boundary between vocalic and intervocalic intervals was placed at the zero-crossing point on the waveform. Vowel-consonant boundaries were mainly delimited by the end of the pitch period preceding a break in the structure of the second vowel formant (F2) accompanied by a significant drop in the waveform amplitude; consonant-vowel boundaries were delimited by the start of a pitch period consistent with the beginning of the second vowel formant (White and Mattys, 2007a; Wiget et al., 2010)

Additionally, stop consonants were identified by the end of a pitch period characterized by a significant drop in amplitude of the waveform and a break in the second formant. For consistency, aspiration following release of stop consonants was included in the intervocalic intervals. Fricative and nasal consonants were identified by the start of

visible friction, and by abrupt spectral changes characterized by reduction of amplitude and spectrographic energy, respectively. Glottalized intervals, as often observed between two successive vowels, were identified by changes in the pitch period such as reduction, doubling and lengthening (Dilley et al., 1996), and were labelled as consonantal intervals. Two successive vowels were labelled as one vocalic segment when there was no glottalization or pause separating them. This contrasts with White and Mattys (2007a) who omitted glottalized intervals and summed the vocalic intervals that preceded and followed them. However, because of the possibility that speakers may use these glottalized intervals to maintain the overall rhythmic flow of the utterance, and because the current research also examines the durational differences between stressed and unstressed vowels which could be affected by summing vowels together, glottalized intervals were not excluded from measurement and were treated as consonantal intervals. In addition, it was observed that the L2 speakers used more frequent and longer glottalized intervals to separate successive vowels than the SSBE and NSA speakers, which might be relevant to the overall percept of non-target-like rhythm. The approach used for identifying glides and liquids followed Low and Grabe (2002), who based their judgements on acoustic, rather than phonological/phonemic, criteria. Where there were no clearly noticeable changes in the formant structure or amplitude of the signal, glides/liquids were treated as part of the vocalic interval. This strategy was also applied to segmentation of the Arabic pharyngeal /ʕ/, which has been shown to have vowel-like formant structure (Laufer, 1996). This was also deemed the best way of dealing with semivowels, since the rhythm metrics are fundamentally acoustic-based. For the same reason, any devoiced vowels or syllabic consonants were treated as (part of) intervocalic intervals.

Some prior studies attempted to exclude approximants from the English stimuli they used due to segmentation difficulty, but this was at the expense of other important methodological concerns, such as number of sentences used, inclusion of varied syllable structures, and avoidance of consecutive primary stressed syllables. Most of these studies used the same set of target sentences, which are five English sentences taken from a corpus of sentences from Nazzi et al, (1998). Moreover, all five sentences include tokens of the grapheme <r>, as in the word 'more', which is pronounced in many rhotic accents of English as an approximant. Many L2 speakers of English may also produce <r> as rhotic, due an effect of orthography (Wells, 2005) and/or due to exposure to rhotic varieties such as American English. Therefore, it was decided not to attempt to exclude approximants from the data in the current study.

The first consonant in all utterances was excluded from the measurements due to the sometimes extreme difficulty in demarcating its beginning. This holds particularly for stop consonants, but for consistency the exclusion was applied to all consonants. Due to possible final-syllable lengthening effects (Delattre, 1966; Klatt, 1976; de Jong and Zawaydeh, 1999), final syllables were excluded from the measurements. Some previous studies (e.g., Low and Grabe, 2002; White and Mattys, 2007a) did include final syllables in their measurements. White and Mattys (2007a) argued that final-syllable lengthening may be language specific and might possibly contribute to the overall perception of rhythm. However, it was sometimes difficult to segment the final syllable, as in many cases the spectral energy decreases significantly, making it extremely hard to mark the boundaries of the phonemes. It is also often difficult to delimit the end of utterance-final consonants (Deterding, 2001). Besides, the present research also examines durational differences

between stressed and unstressed vowels, and inclusion of utterance-final vowels may affect the results due to possible lengthening. Intervals of perceptible silent pauses within utterances were excluded from calculations. In the few cases where these silent pauses were preceded by a stop consonant, both the stop consonant closure and the pause silences were excluded, due to the difficulty of distinguishing the pause from the consonant closure (White and Mattys, 2007a).

Analysis

After segmenting all the utterances, scores for %V, ΔC, rPVI-C, VarcoV, and nPVI-V (see Table 1) were calculated for each sentence produced by each speaker in the four groups. A measure of speaking rate (SR) is also included in the analysis because speech rhythm has been shown to be affected by speech rate (e.g., Dellwo, 2008; Meireles and Barbosa, 2008). Following some previous studies that have investigated speech rate in L2 speech (Munro, 1995; Towell et al., 1996; Trofimovich and Baker, 2006), SR was measured by dividing the number of syllables in an utterance by the total duration of that utterance. The number of syllables for each utterance was calculated based on the number of labelled vocalic intervals in the utterance (i.e., number of syllables in an utterance is equated to number of vowels produced).

Table 1: *Summary of the acoustic rhythmic measures*

Metric	Measurement	Related Work
%V	Percentage of the total duration of vocalic intervals.	Ramus et al. (1999)

ΔV	Standard deviation of the durations of vocalic intervals.	Ramus et al. (1999)
ΔC	Standard deviation of the durations of consonantal intervals.	Ramus et al. (1999)
nPVI-V	Mean of the durational differences between successive vocalic intervals divided by their sum, and multiplied by 100.	Low et al. (2000)
rPVI-C	Mean of the durational differences between successive consonantal intervals.	Grabe & Low (2002)
VarcoV	Standard deviation of the durations of vocalic intervals divided by the mean duration of vocalic intervals, and multiplied by 100.	White & Mattys (2007)
VarcoC	Standard deviation of the durations of consonantal intervals divided by the mean duration of consonantal intervals, and multiplied by 100.	Dellwo (2006)

The vocalic intervals segmented in each utterance from each speaker were labelled as either stressed or unstressed. Primary stressed vowels were labelled as stressed, and all other vowels were labelled as unstressed. Categorising vowels only as stressed or unstressed ones is not the only way of dividing vowels in terms of the degree of stress they bear, since vowels, in English at least, can have more than two degrees of stress e.g. primary, secondary, and weak (Fear et al., 1995; Roach, 2009). The current study, however, took a more general view of stress, dividing vowels into stressed and unstressed only, following Ladefoged (1975) who argues for two levels of phonetic stress in English. A few vocalic intervals contained two consecutive vowels (a stressed vowel preceded by an

unstressed one, as in ‘the outcome’ [ðiaʊt.kʌm]), and in these cases the interval was labelled as a stressed vowel.

For the identification of stressed and unstressed vowels, we first checked stress placement in English in dictionaries and reference books (Cambridge Dictionary Online, 2011; Couper-Kuhlen, 1986; Pike, 1945). All function words were considered unstressed unless they were stressed by the speaker to express contrast (Couper-Kuhlen, 1986; Pike, 1945; Roach, 2009). All monosyllabic content words were labelled as stressed. Stress assignment in polysyllabic words was checked in the Cambridge Dictionary (2011). This was followed by auditory and visual inspection of the waveforms and spectrograms of all the vowels in Praat (Boersma and Weenink, 1992-2018); the expectation was that stressed vowels would have longer duration, greater intensity and higher pitch than unstressed vowels (e.g., Fear et al., 1995; Fry, 1955; Roach, 2009). This procedure was difficult to follow for the utterances of the L2 speakers of English. In many cases, they appeared to stress function words to the same degree as monosyllabic content words, and misplaced stress in polysyllabic words. For function words, the decision was made to consider all function words as unstressed, since one of the main aims of the current study is to find out whether the L2 speakers make a durational difference between (what are expected to be) stressed and unstressed vowels. In the case of polysyllabic words, stress was assigned to vowels based on auditory judgement combined with visual inspection of the vowels’ waveforms and spectrograms.

A parallel procedure was followed to segment and label the NSA vowels. Function words were labelled as unstressed and monosyllabic words were labelled as stressed. Stress placement in polysyllabic words is fully predictable by phonological rules in NSA. Stress

falls on the final syllable if the syllable has the shape CVCC or CV:C; otherwise, it falls on the penultimate if it is CVC or CV: ; otherwise, it falls on the antepenultimate (Al-Mozainy, 1982; Ingham, 1994). Resyllabification sometimes occurs across word boundaries, where the final consonant of a word is resyllabified with the initial syllable of the next word (Kenstowicz, 1986). As this might affect the weight of the syllable, and thus, possibly, the placement of stress, resyllabification was taken into consideration when identifying stress in polysyllabic words.

Having segmented and labelled all the vowels, we calculated the duration of all stressed and unstressed vowels, and measured their first and second formants (F1 and F2) at the midpoint. Because of the spectral transitions in diphthongs, their formant measurements were not included in the analysis. Formant tracking errors were checked and corrected manually. All duration measurements were then normalized for speech rate by first dividing the duration of each sentence by its number of syllables to obtain an average syllable duration for each sentence, then dividing the duration of each vowel in each sentence by the obtained average syllable duration for that sentence (following Kavanagh, 2012). The normalized vowel durations were then divided by 100 to give a more readable number than the large numbers obtained. Mean durations of normalized stressed and unstressed vowels were calculated for each sentence produced by each speaker. Finally, a ratio of the mean duration of stressed vowels to the mean duration of unstressed ones was calculated for each sentence.

A scatterplot of F1 and F2 for all stressed and unstressed vowels, corresponding to the acoustic vowel space, was drawn for each speaker group to see whether they differed in the way they centralize unstressed vowels relative to the stressed ones. The current study

focusses on temporal aspects of rhythm, so no attempt was made to further quantify vowel quality reduction or centralization of unstressed vowels.

All consonant clusters (i.e., CC, CCC, CCCC) in the speech of the L2 speakers were examined auditorily, and the analysis was supported by visual inspection of the waveform and spectrogram in Praat (Boersma and Weenink, 1992-2018) to examine whether the speakers produced them in a target-like way. All consonant clusters were judged to be either correct or incorrect. No attempt was made to categorise alternate realisations of consonant clusters, but the production of a cluster was deemed incorrect if a vowel was inserted to break it up (e.g., /nst/ in 'against' realise as /nɪst/), or if one of the consonants was deleted (e.g., /ksts/ in 'texts' realized as /kst/ or /kɪst/). The L2 speakers' productions of consonant clusters were comparise to canonical citation forms. In other words, their production was not compared to the SSBE speakers' production in the current study, but rather to the citation or dictionary forms of how the clusters are canonically produced by L1 English speakers. Although the position of a consonant cluster might have an effect on how the L2 speakers produced them, context was not considered in the analysis. An overall percentage of target-like production of consonant clusters was calculated for each speaker.

Results

Rhythm metrics

Table 2 provides the mean scores and standard errors (between parentheses) for all the rhythm metrics for each speaker group. For each rhythm metric, a mixed-effects model, with the rhythm metric as a dependent variable, Speaker Group as a fixed factor, and

random intercepts for speakers and utterances, was run to examine whether the speaker groups differed significantly from each other in terms of the metric scores.

Table 2: *Mean scores and standard errors (between parentheses) for rhythm metrics for each speaker group (definitions of the metrics are in Table 1).*

Metric	NSA	SSBE	ME L2	LE L2
%V	39 (0.6)	32 (0.6)	40 (0.6)	38 (0.5)
VarcoV	56 (1.2)	63 (1.7)	46 (1.4)	45 (1.2)
nPVI-V	60 (1.4)	77 (2.3)	48 (1.7)	49 (1.5)
ΔC	45 (1.6)	61 (1.8)	57 (1.7)	63 (2.3)
rPVI-C	54 (2.0)	67 (2.2)	65 (1.9)	68 (2.4)
Speech rate	6.5 (0.1)	5.6 (0.1)	5.1 (0.1)	4.9 (0.1)

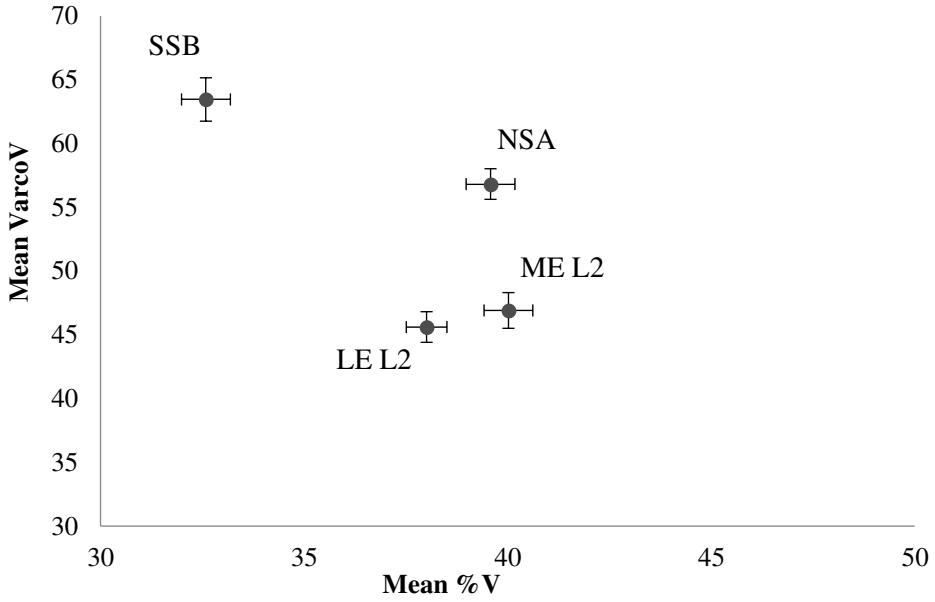
For %V, the results showed a significant main effect of Speaker Group, F (3,236) = 15.07, $p < .01$. A post-hoc test showed that only SSBE was significantly different from all other speaker groups ($p < .01$). This means that the utterances produced by the SSBE speakers had a significantly lower percentage of total vowel duration, relative to overall consonant duration, than did the utterances produced by the NSA and the L2 speaker groups. It is not clear from the metric score alone whether the lower scores for %V on the part of L1 English speakers was because they shortened unstressed vowels to a greater degree than did the other groups, or because their utterances had longer consonantal intervals than the utterances produced by the other groups. LoR did not have an effect on

the L2 speakers' scores for %V, as the two L2 speaker groups were found to have similar scores.

The results for VarcoV showed a main effect of Speaker Group, F (3,236) = 25.66, $p < .01$. Unlike %V, post-hoc tests showed no significant difference between NSA and SSBE ($p < .7$). The L2 speaker groups differed significantly from both NSA ($p < .02$ for the difference between ME L2 and NSA, and $p < .01$ for the difference between LE L2 and NSA) and SSBE group ($p < .01$). This means that the utterances produced by the NSA and the SSBE speaker groups exhibited significantly higher durational variability between vocalic intervals than did the utterances produced by the L2 speakers. The L2 groups had similar scores for VarcoV, which indicates no significant effect of LoR on their VarcoV scores.

White and Mattys (2007a) suggested that %V and VarcoV are complementary and thus provide insights into the influence of L1 on L2. Figure 1.3 below plots the average scores for %V and VarcoV for all speaker groups.

Figure 1: *Scatterplot of mean scores for %V and VarcoV for all groups.*



The graph shows the separation between the SSBE group, on the one hand, and all other speaker groups, on the other. For VarcoV, the NSA group appear to be intermediate between the SSBE and the L2 speaker groups. The graph also clearly illustrates the similarity of the results for the two L2 speaker groups.

For nPVI-V, the results showed a significant effect of Speaker Group, $F(3,236) = 39.01$, $p < .01$. Post-hoc tests for pairwise comparisons revealed a significant difference only between the SSBE group, on the one hand, and all other speaker groups on the other ($p < .01$). The difference between the NSA and L2 speaker groups only approached significance ($p < .08$ for the difference between ME L2 and NSA, and $p < .07$ for the difference between LE L2 and NSA speaker groups).

The utterances produced by the SSBE speakers displayed significantly greater durational variability between successive vowels than did the utterances produced by the NSA and the L2 speaker groups. A possible reason for this finding is that the SSBE

speakers shortened unstressed vowels to a greater degree than the other groups. The results for nPVI-V showed a similar trend to those for %V, as only the SSBE group was found to differ significantly from the other groups. Unlike the scores for VarcoV, nPVI-V scores showed a significant difference between NSA and SSBE. LoR had no effect on the L2 speakers' scores for nPVI-V, as there was no significant difference between the two L2 groups.

The results of the consonantal rhythm metrics, ΔC and rPVI-C, showed significant differences only between the SSBE and L2 speaker groups, on the one hand, and the NSA group on the other (ΔC: F (3,236) = 5.98, p < .01; rPVI-C: F (3,236) = 2.86, p = .05). LoR did not affect the scores calculated for the L2 speakers, as there was no significant difference between the two L2 speaker groups in either measure. This suggests that the utterances produced by the L2 and SSBE speakers showed similar degrees of durational variability between consonantal intervals. However, previous studies have cast doubt on the reliability of the consonantal-based rhythm metrics, as their scores were shown to be easily affected by speech rate (e.g., Barry et al., 2003; Dellwo and Wagner, 2003; White and Mattys, 2007a).

The results for speech rate showed a main effect of Speaker Group , F (3,236) = 16.46, p < .01. Post-hoc tests showed that only NSA was significantly different from all other speaker groups (p < .01). The NSA speaker group spoke at a faster speaking rate than the SSBE and L2 speaker groups. This might be because NSA has simpler syllable structure than SSBE, as noted before. Previous studies have shown that languages with simple syllabic structures are usually spoken at a faster speaking rate (syllable/second) than languages with more complex syllabic structures (e.g., Dellwo, 2010). Although the L2

speakers spoke at a lower speaking rate than the SSBE speakers, the differences between the L2 and the SSBE speakers were not statistically significant ($p < .08$ for the difference between ME L2 and SSBE groups and $p < .4$ for the difference between LE L2 and SSBE groups). The difference between the L2 speaker groups was not significant, which indicates that LoR had no effect on the L2 speakers' production in terms of speaking rate.

Unstressed vowels

The mean durations of stressed and unstressed vowels, and the durational ratios of stressed to unstressed vowels (SUR) were calculated for all the utterances produced by all the speakers in the four speaker groups, and are reported in Table 3.

Table 3: *Means and standard deviations (between parentheses) of the durations of stressed and unstressed vowels and scores for SUR (durational ratio of stressed to unstressed vowel durations) for all speaker groups.*

Speaker Group	Stressed vowel	Unstressed Vowels	SUR
NSA	5.64 (1.1)	2.67 (0.4)	2.14 (0.4)
SSBE	4.56 (0.8)	2.12 (0.4)	2.22 (0.5)
ME L2	4.84 (0.6)	3.19 (0.5)	1.56 (0.3)
LE L2	4.32 (0.6)	3.02 (0.3)	1.44 (0.2)

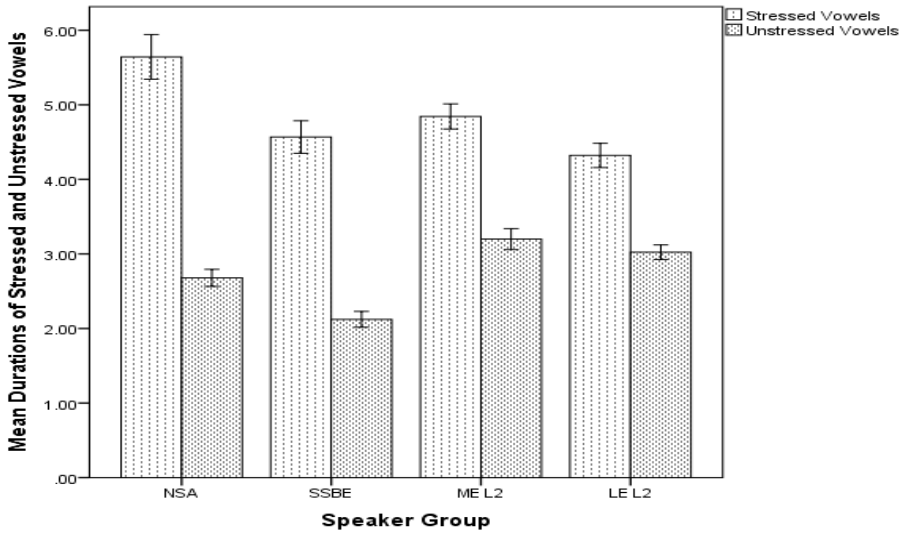
A mixed-effects model, with SUR as dependent variable, Speaker Group as a fixed factor, and random intercepts for utterances and speakers, was run to find out whether the

four speaker groups differed significantly from each other in terms of SUR scores. The model showed a main effect of Speaker Group for SUR, F (3,236) = 32.41, $p < .01$. Post-hoc tests for pair-wise comparisons showed significant differences only between the L1 speaker groups (NSA and SSBE) on the one hand and the L2 groups on the other (all differences were significant at $p < .01$). No significant difference was found between the NSA and the SSBE speaker groups, or between the L2 speaker groups.

The L2 speakers did not make as much durational difference between stressed and unstressed vowels as the NSA and SSBE speakers. As in the results for VarcoV, the NSA speakers had similar SUR scores to the SSBE speakers, which indicates that the speech of both groups had similar durational differences between stressed and unstressed vowels. LoR showed no effect on the results for the L2 speaker groups, as their SUR scores were not significantly different.

It is not clear from the SUR ratio measure whether the similar durational difference between stressed and unstressed vowels in the utterances produced NSA and SSBE are because both groups reduce unstressed vowels. Figure 2 below illustrates durations of each type of vowel for all speaker groups.

Figure 2 *Mean durations of stressed and unstressed vowels for all speaker groups.*



A pair of mixed-effects models with durations of stressed vowels and unstressed vowels as separate dependent variables were run, with Speaker Group as a fixed factor and random intercepts for utterances and speakers, to find out whether the four groups differed in terms of the durations of stressed and unstressed vowels. Both models showed a significant main effect of Speaker Group, for mean durations of stressed vowels F (3,236) = 4.54, $p < .01$ and for mean durations of unstressed vowels $F(3,236) = 39.80$, $p < .01$. Post-hoc tests were run for pair-wise comparisons between the four groups. Although the NSA and the SSBE speaker groups show similar SUR scores, they differ to in the mean durations of stressed and unstressed vowels independently, to a significant extent ($p = .05$ for stressed vowels and $p < .01$ for unstressed vowels), with NSA vowels of both types longer than their SSBE counterparts. The L2 speaker groups differed significantly from the NSA and the SSBE speaker groups in terms of the mean durations of unstressed vowels ($p < .01$ for the differences between the L2 groups and SSBE, $p < .01$ for the differences between ME L2 and NSA, and $p < .05$ for the difference between LE L2 and NSA), with

learners producing longer unstressed vowels than NSA and SSBE. There were no significant differences between the L2 speaker groups for mean durations of stressed and unstressed vowels.

Although the NSA and the SSBE speakers had similar durational ratios of stressed to unstressed vowels (SUR), the NSA speakers did not shorten unstressed vowels to the same degree as the SSBE speakers. This suggests that the similarity between the SSBE and NSA scores for SUR is not because the NSA speakers shortened unstressed vowels to the same degree as the SSBE speakers, but instead most likely due to the fact that vowel length is phonemically contrastive in NSA, where all long vowels have short counterparts which are about half their lengths (Alghamdi, 1998). English also has long vowels but the durational difference between short and long vowels in Arabic is larger than in English (Mitleb, 1981). In addition, the Arabic quantity sensitive stress algorithm means that the overwhelming majority of long vowels will attract stress.

A lower score for SUR then, can be caused not only by shortening of unstressed vowels, but by the phonemic length contrast between short and long vowels. Looking at the metric results above, it seems that %V and nPVI-V (which set SSBE apart from the NSA and L2 speaker groups) can account slightly better for unstressed vowel shortening; in contrast, VarcoV (which sets SSBE and NSA apart from the two L2 speaker groups) can account better for more general temporal differences between stressed and unstressed vowels, and is robust to the fact that not all languages shorten unstressed vowels.

As unstressed vowel durational shortening is also associated with reduction in vowel quality (see Section 1.1 above) (e.g., Flemming, 2004; Lindblom, 1963), plotting the

formants of stressed and unstressed vowels for all speaker groups should further illustrate the difference between the SSBE group and all other speaker groups regarding unstressed vowel shortening. Figure 3, 4, 5 and 6 show scatterplots of F1 and F2 of stressed and unstressed vowels for all speaker groups.

Figure 3

Scatterplot of F1 and F2 of stressed and unstressed vowels for the NSA speaker group.

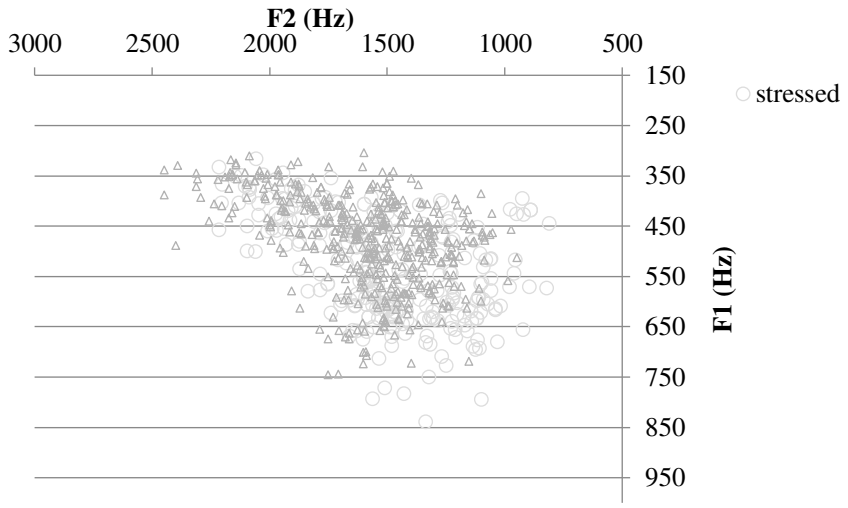


Figure 4

Scatterplot of F1 and F2 of stressed and unstressed vowels for SSBE speaker group.

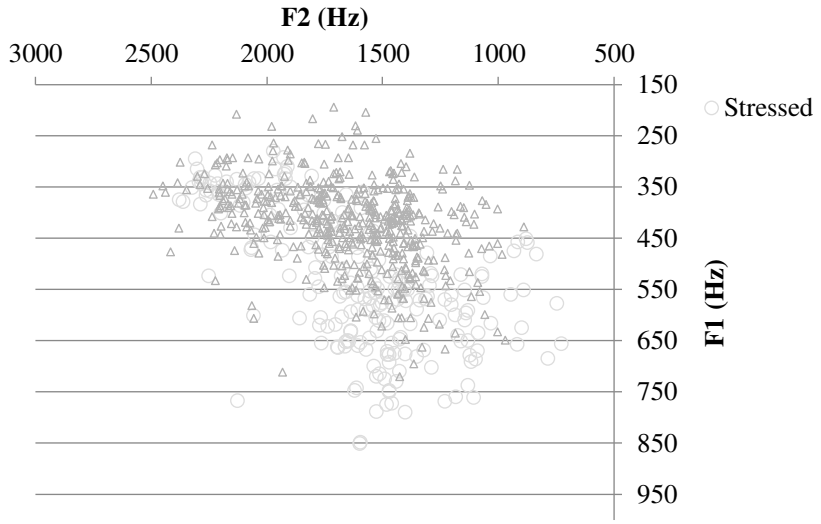


Figure 5

Scatterplot of F1 and F2 of stressed and unstressed vowels for ME L2 speaker group.

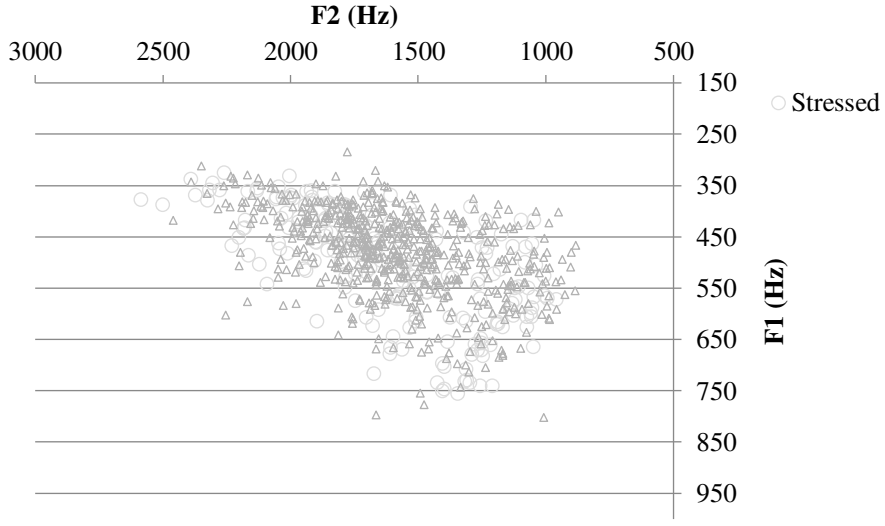
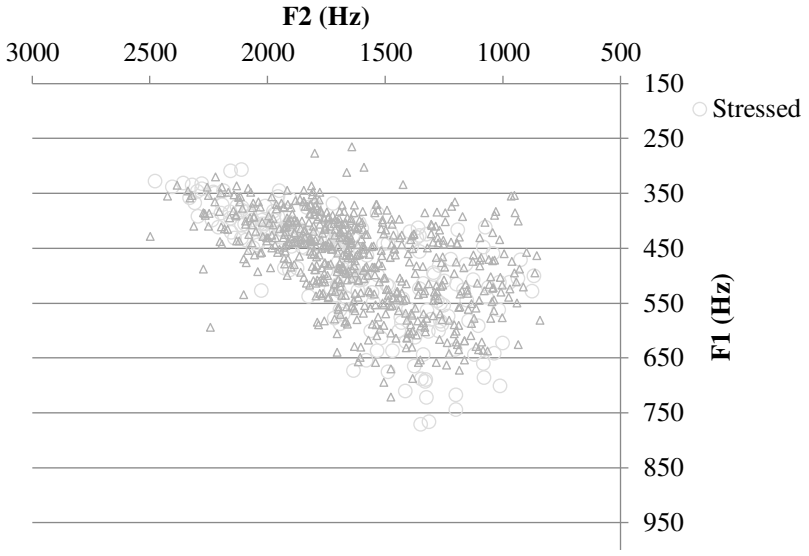


Figure 6

Scatterplot of F1 and F2 of stressed and unstressed vowels for LE L2 speaker group.



Figures 3-6 show that the SSBE speaker group made a clearer spectral distinction between stressed and unstressed vowels than all other speaker groups. The unstressed vowels, relative to the stressed ones, produced by the SSBE speakers are more clustered in the F1-F2 formant space than the unstressed vowels produced by the NSA and L2 speakers. In contrast, the NSA and the L2 speaker groups show overlapping distributions of F1/F2 values for stressed and unstressed vowels. These patterns provide further support for interpretation of %V and nPVI-V scores as sensitive to degree of unstressed vowel reduction.

Consonant clusters

Both L2 speaker groups showed high percentages of target-like production of consonant clusters (ME L2: $M = 87.61\%$, $SD = 7.3$; LE L2: $M = 80.47\%$, $SD = 9.1$). The ME L2 group had a higher raw percentage of target-like productions than the LE L2 group; nonetheless, an independent samples t-test showed no significant difference between the

two speaker groups, $t(10) = 1.49$, $p > 0.01$. To examine whether the results for the L2 groups differ significantly from a hypothesized L1 English group, with a mean of 100% correct production, a one-sample t-test was run for each L2 speaker group. Both ME L2 and LE L2 groups differed significantly from the hypothesized population, $t(5) = 4.11$, $p < 0.01$ and $t(5) = 5.20$, $p < 0.01$, respectively. The percentage of target-like production did vary considerably, however, according to the type of consonant cluster. As might be expected, the percentage decreased as complexity of consonant cluster increased.

Table 4: *Mean and standard deviations (between parentheses) for percentage of target-like productions of English consonant cluster types by the L2 speaker groups.*

Speaker Group	CC Cluster ($N = 23$)	CCC Cluster ($N = 10$)	CCCC Cluster ($N = 2$)
ME L2	97.10 (2.2)	80 (20.9)	16.67 (25.8)
LE L2	92.75 (5.2)	66.67 (20.6)	8.33 (8.3)

The ME L2 speaker group had higher mean percentages of target-like productions of consonant clusters for all cluster types than the LE L2 speaker group. However, independent sample t-tests showed no significant difference between the two groups on all types of consonant clusters. The high percentage of target-like production consonant clusters by the L2 speakers may explain why the consonantal rhythm metrics did not show any significant differences between the L1 and L2 English speakers.

Discussion

The current study used a range of rhythm metrics to examine the production of English speech rhythm by two groups of L2 Saudi learners: 'more experienced' (ME) and 'less experienced' (LE), based on their length of residence in the UK. It also examined the speech rhythm of Najdi Saudi Arabic (NSA) and Sothern Standard British English (SSBE) for comparison and to help in explaining the results. Similar to most previous L2 studies that have used the rhythm metrics (e.g., Behrman et al., 2019; Ordin & Polyanskaya, 2014; White and Mattys, 2007b), all three vowel-based rhythm metrics used in the current study (%V, VarcoV and nPVI-V) showed significant differences between the L1 and L2 English speakers. Given that the vowel-based rhythm metrics were originally developed to capture the durational variability between vocalic segments arising from shortening of unstressed vowels, the initial conclusion would be that the L2 speakers did not shorten unstressed vowels to the same degree as the SSBE speakers. However, the rhythm metric results for NSA point to a more nuanced picture. While the NSA group show significantly lower nPVI-V scores than the SSBE group, the two groups had similar VarcoV scores. This result gives further support to the previous studies that have recommended use of more than one measure for studying rhythm in speech (e.g., Wiget et al., 2010).

To make more sense of the data, we analyzed the duration of stressed and unstressed vowels for all the speaker groups. The durational ratio of stressed to unstressed vowels (SUR) showed similar results to VarcoV. However, a closer look at the durations of stressed and unstressed vowels independently showed that the durational variability between vocalic segments in the case of NSA must derive from another source of temporal variability, and not because of any shortening or reduction of unstressed vowels; we

ascribe this result to is the particular phonetic exponence in Arabic of the phonemic difference between short and long vowels. This finding was supported by the analysis of unstressed vowel quality reduction, as only the SSBE group was shown to clearly centralize unstressed vowels.

The results of the consonantal rhythm metrics showed similar results for both the L1 and L2 English speakers. This may either be due to the instability of the consonantal metrics, as shown in some previous studies (White & Mattys, 2007a; Wiget et al., 2010), or the success of the L2 learners in producing similar durational variability of consonant segments. The latter interpretation is supported by analysis of consonant cluster production by the L2 learners, the majority of which was target-like. The NSA speech exhibited less durational variability of consonantal intervals than the speech of the L1 and L2 English speakers. This result is consistent with the fact that NSA has a simpler consonantal structure than English.

The utterances produced by the L2 speakers showed similar speech rate to those produced by the SSBE speakers. This contrasts with the results of most previous studies, where L2 English speakers have been found to speak at a lower speech rate than L1 English speakers (e.g., Munro & Derwing, 2001). We ascribe this positive outcome to the fact that the speech elicited from the speakers was read, and the speakers were familiar with the utterances from the paraphrasing task before being asked to read them. In contrast, the NSA utterances were spoken at a faster rate than the L1 and L2 English utterances which we also ascribe to the simpler syllable structure of NSA.

Length of residence, as a rough index of L2 experiences, did not affect any of the results, as both L2 groups showed similar scores in all measures. This might be due to the

relatively short difference in LoR between the two groups. Trofimovich and Baker (2006) found a positive effect of LoR on the production of English stress timing by L2 Korean speakers, which was measured as a durational ratio of unstressed to stressed syllables. However, their learners' amount of experience varied more than in the current study. While their L2 "moderately experienced" speaker group had spent two to three and half years in the United States, their L2 "experienced" speaker group had spent from seven to 15 years. Another possible reason for not finding a positive effect of LoR on the production of English rhythm by the L2 speakers is because prosody, of which rhythm is a component, is an extremely difficult aspect to learn, given that languages vary not only in what prosodic structures they exhibit but also in how these structures are implemented (Mennen & de Leeuw, 2014).

Conclusion

The current study showed that the L2 speakers, regardless of their length of residence in the UK, had lower durational variability of vocalic intervals than L1 English speakers. This can be seen at least partially as a case of L1 transfer, as the NSA speakers were also found not to reduce unstressed vowels to the same degree as SSBE speakers. The results add to the arguments against categorical classification of languages into rhythmic classes, since both English and Arabic are classified as 'stressed timed' languages, yet show considerable differences. As expected, and perhaps because of their susceptibility to speech rate, the results of the consonantal rhythm metrics were initially difficult to explain. The L2 speakers were found to have similar durational variability of consonantal intervals

to the SSBE speakers, but this is consistent with their largely target-like production of consonant clusters.

Overall, this study provides fresh support for the recommendation to make use of multiple rhythm metrics in investigation of L2 speech rhythm, since the different metrics are here shown to be sensitive to different phonological parameters that contribute to rhythmic variation. We demonstrated how use of simple local measures, such as durations of stressed and unstressed vowels and evaluation of consonant cluster realisation, can aid in disambiguating between competing interpretations of rhythm metric scores. Finally, the present study suggests that the L2 Saudi learners, and their teachers, who wish to foster development of speech rhythm that is closer to that of SSBE should prioritize pronunciation training focused on unstressed vowel reduction.

References

Abercrombie D. (1976) *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Adams C. (1979) *English speech rhythm and the foreign learner*. Berlin: De Gruyter Mouton.

Al-Mozainy H. (1982) *Vowel alternations in a Bedouin Hijazi Arabic dialect: Abstractness and stress*. PhD Thesis, University of Texas, USA.

Allen G. (1975) Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetic* 3: 75-86.

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

Behrman A, Ferguson S, Akhund A, and Moeyaert M. (2019) The effect of clear speech on temporal metrics of rhythm in Spanish-accented speakers of English. *Language and speech* 62: 5-29.

Boersma P, Weenink, D. (1992-2018) *Praat: doing phonetics by computer* (Version 5.1.30) [computer program]. Retrieved from <http://www.praat.org>.

Carter P. (2005) Quantifying rhythmic differences between Spanish, English, and Hispanic English. In: Gess R and Rubin E (eds) *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages*. Amsterdam: John Benjamins, pp.63-75.

Couper-Kuhlen E. (1986) *An introduction to English prosody*. London: Hodder Arnold.
Dasher R and Bolinger D. (1982) On pre-accentual lengthening. *Journal of the International Phonetic Association* 12:58-69.
Dauer R. (1983) Stress-timing and syllable-timing reanalyzed. *Journal of Phonetic,* 11:51–62.
de Jong K and Zawaydeh B. (1999) Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetic* 27:3-22.
Delattre P. (1966) A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics in Language Teaching* 4:183-198.
Dellwo V. (2006) Rhythm and speech rate: A variation coefficient for deltaC. In Karnowski P and Szigeti I (eds) *Language and language processing: Proceedings of the 38th linguistic colloquium*. London: Peter Lang, pp.231–224.
Dellwo V. (2008) The role of speech rate in perceiving speech rhythm. In: *Speech Prosody 2008* (ed P Barbosa, S Madureira and C Reis), Campinas, Brazil, 6-9 May 2008, pp. 375-378.
Deterding D. (2001) The measurement of rhythm: a comparison of Singapore and British English. *Journal of Phonetics* 29:217-230.
Dilley L, Shattuck-Hufnagel S and Ostendorf M. (1996) Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetic,* 24:423-444.
Donovan, A., and Darwin, C. (1979) The perceived rhythm of speech. In: *Proceedings of The Ninth International Congress of Phonetic Sciences*, Copenhagen, Denmark, 6-11 August 1979, pp. 268-274.
Fear B, Cutler A, and Butterfield S. (1995) The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America,* 97:1893-1904.
Fry D (1955) Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27:765-768.
Ghazali, S., Hamdi, R., and Melissa, B. (2002) Speech rhythm variation in Arabic dialects. In: *Speech Prosody 2002*, Aix-en-Provence, France, 11-13 April 2002, pp. 331-334.
Grabe E and Low EL. (2002) Durational Variability in Speech and the Rhythm Class Hypothesis. In: Gussenhoven C and Warner N (eds.) *Papers in Laboratory Phonology 7.* Berlin: Mouton de Gruyter, pp. 515-546.
Grenon I and White L. (2008) Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In Chan H, Jacob H and Kapia E (eds), *BUCLD 32: Proceedings of the 32nd annual Boston University Conference on Language Development*. Somerville: Cascadilla Press, pp. 155-166.
Gut, U. (2003) Non-native speech rhythm in German. In: *Proccedings of the15th International Congress of Phonetic Sciences*, Barcelona, Spain, 3-9 August 2003, pp. 2437-2440.
Hamdi R, Barkat-Defradas M, Ferragne E and Pellegrino F. (2004) Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches. In: *Interspeech-2004*, Jeju Island, Korea, 4-8 October 2004, pp. 1613-1616.
Ingham B. (1994) *Najdi Arabic: Central Arabian*. Merstham: John Benjamins.
Jones D. (1962) *An outline of English phonetics*. Cambridge: Cambridge University Press.
Kavanagh C. (2012) *New consonantal acoustic parameters for forensic speaker comparison* PhD Thesis, University of York, UK.

Kenstowicz M. (1986) Notes on syllable structure in three Arabic dialects. *Revue québécoise de linguistique,* 16:101-127.
Klatt D. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America,* 59: 1208-1221.
Knight R. (2011) Assessing the temporal reliability of rhythm metrics. *Journal of the International Phonetic Association,* 41:271-281.
Ladefoged P. (1975) *A course in phonetics*. New York: Harcourt Brace Jovanovich.
Laufer A. (1996) The common [ʕ] is an approximant and not a fricative. *Journal of the International Phonetic Association,* 26:113118.
Laver J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.
Lehiste I. (1977) Isochrony reconsidered. *Journal of Phonetic,* 5:253-263.
Li A and Post B. (2014) L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German learners of English. *Studies in Second Language Acquisition,* 36:223-255.
Low E, Grabe E, and Nolan F. (2000) Quantitative characterizations of speech rhythm: syllablet-timing in Singapore English. *Language and Speech,* 43:377-401.
Mehler J, Dupoux E, Nazzi T and Dehaene-Lambertz G. (1996) Coping with linguistic diversity: the infant's viewpoint. In: Morgan J and Demuth K (eds) *Signal to syntax: bootstrapping from speech to grammar in early acquisition*. Mahwah: Lawrence Erlbaum Associates, pp. 365-388.
Meireles, A., and Barbosa, B. (2008) Speech rate effects on speech rhythm. In *Speech Prosody 2008* (ed P Barbosa, S Madureira and C Reis), Campinas, Brazil, 6-9 May 2008, pp. 327-330.
Miller M. (1984) On the perception of rhythm. *Journal of Phonetics* 12:75-83.
Mok P and Volker D. (2008) Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Speech Prosody 2008* (ed P Barbosa, S Madureira and C Reis), Campinas, Brazil, 6-9 May 2008, pp. 423-426.
Munro M. (1995) Nonsegmental factors in foreign accent: Ratings of filtered speech, *Studies in Second Language Acquisition* 17:17-34.
Munro M and Derwing T. (2001) Modelling perceptions of the comprehensibility and accentedness of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition,* 23:451-468.
Nazzi T, Bertoncini J and Mehler J. (1998) Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance,* 24:756-766.
Nespor M and Vogel I. (1989) On clashes and lapses. *Phonology,* 6:69-116.
Ordin M and Polyanskaya L. (2014) Development of timing patterns in first and second languages. *System* 42:244-257.
Ordin, M and Polyanskaya L. (2015). Perception of speech rhythm in second language: the case of rhythmically similar L1 and L2. *Frontiers in psychology,* 6:1-15.
Peterson G and Lehiste I (1960) Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32:696-703.
Pike K. (1945) *The Intonation of American English*. Michigan: University of Michigan Press.

Piske T, MacKay I and Flege J. (2001) Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics* 29:191-215.

Polyanskaya L and Ordin M. (2019) The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language. *Applied Psycholinguistics* 40:795-819.

Ramus F, Nespor M and Mehler J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition* 37:265-292.

Roach P. (1982) On the distinction between 'stress-timed' and 'syllable-timed' languages. In Crystal D (eds) *Linguistic Controversies.* London: Edward Arnold, pp.73-79.

Roach P. (2009) *English phonetics and phonology: A practical course.* Cambridge: Cambridge University Press.

Schaeffler F. (2001) Measuring Rhythmic Deviation in Second Language Speech. In: *EUROSPEECH2001,* Aalborg, Denmark, 3-7 September 2001.

Stockmal V, Markus D and Bond D. (2005) Measures of Native and Non-Native Rhythm in a Quantity Language. *Language and Speech* 48:55-63.

Taylor D (1981) Non-native speakers and the rhythm of English. International Review of *Applied Linguistics in Language Teaching,* XIX:219–226.

Towell R, Hawkins R and Bazergui, N. (1996) The development of fluency in advanced learners of French. *Applied Linguistics* 17:84-119.

Trofimovich, P and Baker W. (2006) Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *Studies in Second Language Acquisition,* 28:1-30.

Turk A, Nakai S and Sugahara M. (2006) Acoustic segment durations in prosodic research: a practical guide. In: Sudhoff S, Lenertová D, Meyer R, Pappert S, Augurzky P, Mleinek I, Richter N, and Schließer J (eds) *Methods in empirical prosody research (Language, Context and Cognition).* Berlin: De Gruyter, pp.1-28.

Wells J. (2005) Goals in teaching English pronunciation. In: Dziubalska-Koaczyk K and Przedlacka J (eds) *English pronunciation models: a changing scene*. Bern: Peter Lang, pp.1-11.

Wenk B and Wioland F. (1982) Is French really syllable-timed? *Journal of Phonetic* 10:193-216.

White L and Mattys S. (2007a) Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35:501-522.

White L and Mattys S. (2007b) Rhythmic typology and variation in first and second languages. In: Prieto P, Mascaroī J and Solé, M (eds.) *Segmental and prosodic issues in romance phonology*. Amsterdam: John Benjamins, pp.237–257.

White L, Mattys S and Wiget L. (2012) Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language* 66:665-679.

Wiget L, White L, Schuppler B, Grenon I, Rauch O and Mattys S. (2010) How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America* 127:1559-1569.