SPEECH RHYTHM IN SPONTANEOUS AND CONTROLLED L2 SPEAKING MODES: EXPLORING DIFFERENCES AND DISTANCE MEASURES

KATHERINE FRASER
University of Barcelona
kfrasefr18@alumnes.ub.edu

JOAN C. MORA
University of Barcelona
mora@ub.edu

Abstract
Studies of speech rhythm have often used read speech rather than spontaneous speech in their comparisons. However, read speech has been shown to be perceptually different from spontaneous speech, which may be due to rhythmic differences between the two modes. To examine this, the effect of speaking mode (spontaneous or controlled) was assessed in a group of 82 Spanish-Catalan learners of English relative to a control group of 8 native English speakers. Results found strong rhythmic differences between the two modes, but minimal differences between the learners and native speakers. Additionally, Mahalanobis distance analyses revealed that non-native speakers differed significantly more from the native control group in the spontaneous condition than the controlled condition.

Keywords: speech rhythm, rhythm metrics, English, Spanish, Catalan, Mahalanobis distances

1. Introduction

Speech rhythm is a key suprasegmental component of L2 phonological acquisition, however, most research findings in L2 speech rhythm have come from read speech and have not examined spontaneous speech or compared between spontaneous and controlled speech modes. This study assesses differences in L2 speech rhythm between controlled speech and spontaneous speech for Spanish-Catalan learners of English and how they differ from a control group of English native speakers. Additionally, it examines how pairs of individual rhythm metrics can be used together to measure distances between L2 learners' and native speakers' speech through a novel Mahalanobis distance analysis that may account better for the multivariate nature of speech rhythm.

2. Literature Review

Since Pike (1945) first introduced the terms "stress-timed" and "syllable-timed" to describe the linguistic rhythm classes languages belong to, researchers have attempted to find and measure the phonetic basis underlying these perceptual phenomena (Mairano, 2007). Although the isochrony (of syllables, feet or mora) that these terms came to be based on has not held up to scrutiny (Roach, 1982; Dauer, 1983; Arvaniti, 2012), researchers like Bertinetto (1989) and Dauer (1983), theorized that the presence or absence of specific phonological properties in a language may be what creates the impression of syllable-timing or stress-timing. Such properties centrally included the presence or degree of vowel reduction and the allowance and degree of complexity of consonant clusters (indicative of syllable structure complexity), all of which can in principle be captured by measuring the duration of vocalic and consonantal speech intervals. This suggested that languages, rather than belonging to discrete syllable-timed or stress-timed categories, could be placed along a continuum between these two prototypical rhythm classes. English, as a language with consonantal clusters and significant vowel reduction in unstressed syllables, occupies a place near the stress-timed end of the continuum, while Spanish, with more restrictions on consonant clusters and a lack of vowel reduction is near the opposite, syllable-timed end of the continuum. Languages with a combination of these features, like Polish, remain difficult to categorize (Cantarutti & Szczepek-Reed, 2021).

Based on the refinement of the idea of linguistic rhythm as it relates to language-specific phonological features, rhythm has been operationalized as patterns of durational variability of consonantal (C) and vocalic (V) intervals (Ramus et al., 1999; Low et al., 2000). Languages ranking low on such variability indices would be closer to the syllable-timed end of the continuum, whereas those ranking high due to duration differences between stressed and unstressed vowels and variability in the complexity of consonant clusters would be closer to the stress-timed end of the continuum (Ramus et al., 1999). A variety of rhythm metrics using durational variability (Table 1) were developed in conjunction with this approach. Rate-normalized rhythm metrics were developed from the original raw metrics upon noticing that durational variability is very sensitive to speech rate (Dellwo, 2010).

Table 1: Rhythm metrics commonly used in the research of durational variability in speech

Metric	Description
Delta (Δ)	the standard deviation of the interval (vocalic or V), expressed in milliseconds (Ramus et al., 1999)
Varco	rate-normalized Delta, calculated by dividing Delta by the mean interval duration (i.e. the standard deviation of the V or C interval divided by the mean duration of all V or C intervals) (Ramus et al., 1999; Dellwo, 2010)
%V	%V is the percentage of all speech material that is made up of vocalic intervals (Ramus et al., 1999)
rPVI	the raw Pairwise Variability Index (rPVI) for C or V intervals computes the durational differences between each C or V interval and the one that immediately follows it, then uses all of these separate differences to compute an average for the full speech interval (Low et al., 2000)
nPVI	the normalized Pairwise Variability Index is rate-normalized rPVI, calculated by dividing the rPVI by the mean interval duration (Grabe & Low, 2002)

Rhythm metrics have been used mainly to determine the rhythm class languages belong to (e.g. Ramus & Mehler, 1999; Loukina et al., 2011; Grabe and Low, 2002; Arvaniti, 2012), but more recently have been used to compare the speech rhythm of L2 and L1 speakers of the same language (e.g. Algethami & Hellmuth, 2023; Ordin and Polyanskaya, 2015; van Maastricht et al., 2021). While cross-language comparisons of speech rhythm reflect differences in language-specific phonological and phonotactic properties, within-language comparisons of the speech rhythm of learners and native speakers instead would characterize L2 prosodic development by reflecting the extent to which learners have acquired the rhythmic patterns of the L2 (Levis, 2018). When speaking their L2, as learners attempt to accommodate the L2 rhythmic characteristics, they will exhibit measurable differences from both their L1 rhythm and the rhythm of native speakers of their L2 (White & Mattys, 2007).

For example, Ordin and Polyanskaya (2015) examined how French and German learners of English at different levels of proficiency developed their speech rhythm, with German being rhythmically similar to English (stress-timed) and French being rhythmically dissimilar (syllable-timed). They found that

learners from both language backgrounds exhibited progression from a more syllable-timed rhythmic pattern towards more stress-timed rhythm as proficiency increased. However, only the L1-German group exhibited native-like rhythmic patterns at advanced levels of proficiency. These findings shed light on the importance of the learner's L1 background in acquiring L2 rhythmic patterns. Additionally, it showed that the German learners progressed through a syllable-timed stage early in their acquisition of English, despite the fact that their L1 is considered, like English, to be stress-timed. This led the authors to question whether L2 rhythm acquisition progresses through predicable patterns, as it does with children when acquiring their L1 rhythm (Ordin & Polyanskaya, 2014; Polyanskaya & Ordin, 2015). Still, Li and Post (2014) demonstrated that L1-Mandarin and L1-German learners acquired vocalic variability in L2-English similarly, although the two groups differed in the ratio of vocalic material (%V) in their speech, providing evidence of transfer of their differential L1 rhythmic properties.

While the majority of studies on speech rhythm have examined controlled speech (typically read speech) and some have examined spontaneous speech, very few have compared speech rhythm between controlled and spontaneous speech and so far only in the context of an L1. Despite difficulties in controlling for phonetic content (and therefore V and C interval durations) across and within speakers in spontaneous speech, the characterization of prosodic development in L2 acquisition and the role of L2 speech rhythm in communication cannot be fully accomplished if limited exclusively to controlled speech. Furthermore, it is currently unknown to what extent learners are able to inhibit their L1 speech rhythm and use a more target-like rhythm in controlled or in spontaneous speech. The dynamic nature of spontaneous speech (compared to controlled speech), however, might increase the variability of C and V interval durations, which might favour the implementation of L2 stress-timed rhythmic patterns. It is also currently unclear whether rhythm metrics can capture differences between controlled and spontaneous speech that L1 listeners can perceptually identify independently of variations in speech rate (Dellwo et al., 2015b). For example, Kim and Jang (2009) found read speech to be more syllable-timed, with lower V interval variability (lower Varco-V and nPVI-V) and higher %V than spontaneous speech. By contrast, Dellwo et al. (2015a) and Leemann et al. (2014) did not find read speech to differ significantly from spontaneous speech in rhythm (%V, ΔC and nPVI-V) for native speakers of Zurich German.

To the best of our knowledge, no research to date has investigated whether rhythm metrics differ significantly in controlled versus spontaneous speech production in L2 learners or the extent to which learners can be rhythmically more target-like in controlled or spontaneous speech. Although in general we are expecting less consistency across speaking styles in speech rhythm in L2 than L1 speech, and consequently potentially larger rhythmic differences between

controlled and spontaneous speech in L2 learners than in L1 speakers, the validity of current rhythm metrics to capture such differences needs further investigation.

Related to the underexamined question of within-language rhythmic differences between non-native and native speech and between controlled and spontaneous speech, is the question of whether such speech rhythm differences are consistent within speakers. Arvaniti (2012) ran correlational analyses between several rhythm metrics across pooled individual data, but correlations between controlled and spontaneous L2 speech have not been explored yet. Although differences may exist between controlled and spontaneous speech as a function of whether speech is native or non-native, as predicted by the perceptual differences between these speech types (Dellwo et al., 2015a), it is yet to be seen whether such differences are predictable on an individual level. Correlating rhythm scores between speech types for both native speakers and learners could elucidate these relationships.

Given the multidimensional nature of the phonological properties distinguishing between languages (e.g. vowel reduction, syllable structure complexity, stressed to unstressed syllable duration ratios), pairs of V and C rhythm metrics (rather than single measures, e.g. %V or Varco-C) are used to determine the position of languages in a two-dimensional rhythm space (e.g., %V–Varco-C; Grabe & Low, 2002; White &Mattys, 2007). This allows visually determining the extent to which languages differ from one another or cluster together multidimensionally according to the rhythm class they belong to or the specific pair of metrics used. It also offers the possibility, not usually implemented, of computing spatial distance scores to assess such distances. For example, Arvaniti (2009) computed cross-language Euclidean distances using various combinations of metrics between English and German, Italian, Korean, Spanish and Greek and found German to be rhythmically much closer to English (smaller Euclidean distance scores, 3.8) on a space defined by %V and ΔC than to the other languages (15.4, 12.6, 11.2, 13.1, respectively). Such Euclidean distance measures have also been used in within-language comparisons assessing differences in speech rhythm between controlled and spontaneous speech (Arvaniti, 2012), which have found English to present much larger distance scores between read sentences and spontaneous speech (14.2) than Spanish (4.2), despite the fact that in both languages spontaneous speech appears to be more vocalic (i.e. higher %V) and present higher variability of C intervals (ΔC) than read speech.

Due to the limitations of rhythm metrics and Euclidean distance scores in unambiguously capturing differences between languages across controlled and spontaneous speech materials (Arvaniti, 2012), it is challenging to use rhythm metrics and distance scores to characterize the potentially systematic differences between L2 English and L1 English in controlled and spontaneous speech materials, which is what the current study sets out to do. We aim to partially overcome these limitations by selecting a set of rhythm metrics that maximally

distinguish learners' L1 (Spanish) from their L2 (English) rhythmically and by computing a rhythm distance measure (Mahalanobis distances) that, unlike Euclidean distances, takes into account the variability associated with the speech samples measured when determining locations and distances between controlled and spontaneous L2 and L1 English speech.

The current study addressed the following research questions (RQ):

RQ1. Does the English speech rhythm of Spanish-Catalan learners of English exhibit differences between controlled and spontaneous speech?

RQ2. For individual rhythm metrics, does the English speech rhythm of Spanish-Catalan learners of English differ from that of a native English speaker control group in controlled speech? Is this reflected in spontaneous speech?

RQ3. Are the Mahalanobis Distances between the English speech rhythm of Spanish-Catalan learners of English and the native English speaker control group different for controlled and spontaneous speech?

RQ4. On an individual level, is the English speech rhythm of Spanish-Catalan learners of English consistent across speaking modes?

3. Methods

3.1 Participants

Eighty-two Spanish-Catalan bilingual learners of English (70 female, 12 male) participated in this study for course credit (see Table 2 for demographics). Their level of English proficiency ranged from upper-intermediate to advanced (CEFR level B2-C2) according to the outcome of an elicited imitation task (EIT) and self-estimated proficiency. Eight native speakers of British English participated as a control group providing L1-English data. They were English teachers in the region of Barcelona with a neutral British accent and were paid for their participation.

Table 2: Spanish-Catalan participants' demographics and L2 proficiency

Measure	***M***	***SD***
Age at testing (years)	20.23	2.94
Age of onset of English learning (years)	5.13	1.78
*Vocabulary size (0-10,000 words)**	6741	1232
*Self-estimated proficiency***	7.12	0.97
Tested proficiency (EIT; 0-120)	96.60	13.40

*Tested using a Yes/No receptive vocabulary size test (Meara & Miralpeix, 2016)
**Reading, listening, speaking, writing: 9-point Likert scale from 1=very poor to 9=native-like

3.2 Production tasks

3.2.1 Spontaneous speech tasks

Spontaneous speech samples were elicited from learners and native English speakers through an adaptation of the Dinner Table Task (Ur, 1981), whereas an elicited imitation task was used to obtain controlled speech samples. In the Dinner Table task participants were asked to come up with a seating arrangement for 6 characters attending a dinner party (based on their personalities, hobbies, and professions) that would maximize pleasant conversations. In the first part of the task, characters were already seated around the table and the participants had to justify why the arrangement would not work. In the second part, they were given character cards and asked to come up with a seating arrangement of their own. Participants performed a simple and a complex version of the task on two separate days, differing in the number of characters at each table (2 versus 3) and whether their personality traits were coherent (for example "open-minded" and "humble") or incoherent ("kind" and "greedy for money"). A total of 4-10 minutes of speech were recorded for each participant. In this way, four recordings were obtained per participant in this speaking task (Part 1 and Part 2 for the simple and the complex versions of the task).

Speech chunks containing no pauses, hesitations, repetitions or sound elongations were manually extracted from each of the four speech samples from each participant. Chunks from within each speech sample were concatenated into a single audio file of approximately 100 words using Praat (Boersma & Weenink, 2022). While for the L1 British English Control group, full sentences and complete thoughts were able to be extracted due to the speakers' fluency, L2 speech (in particular, that produced by less proficient learners) was hard to divide into full sentences of the type described in Leemann et al. (2014). Chunks of speech between pauses were sometimes very short, consisting of a single noun phrase or verb phrase. Following Arvaniti (2012), utterances were separated mainly based on pause placement. Omitting silent and filled pauses along with segmental lengthenings and repetitions allowed the spontaneous speech samples to be more comparable with the controlled speech samples collected from elicited imitation.

3.2.2 Controlled speech: elicited imitation task

An elicited imitation task (Wu et al., 2022) consisting of 30 sentences of varying length and complexity was originally used to measure oral proficiency. Six of the eight native speakers completed the same task as a baseline. A recording of each sentence was played over headphones, followed by a 2.5 second pause, after which participants were signaled to repeat the sentence they had heard. This procedure was selected in place of using read materials in order to avoid speakers entering a "reading mode" (Ordin & Polyanskaya, 2015). In an effort to

avoid read speech, some studies have used a combination of memorization and photo cues to elicit speech, while others have used transcriptions of a speaker's own previously uttered spontaneous speech in place of more traditional text or sentences (Ordin & Polyanskaya, 2015; Dellwo et al., 2015a). Additionally, clear cognitive differences between the act of spontaneously producing speech and the act of reading, repeating or reciting make this a particularly challenging methodological hurdle. In the present study, the 2.5 second delay between hearing the sentence and repeating it was intended to impede the phonological loop from repeating the sentence exactly as it was heard directly from rote memory (it had to be previously processed for repetition) and imitating its rhythmic characteristics. The speaker in the recording is a native speaker of American English and spoke in a conversational style at a mean speech rate of 4.58 (*SD*=0.58) syllables per second.

Of the 30 sentences, 12 were selected (which the majority of participants had accurately repeated), extracted from the recording of each speaker, and used as samples of controlled speech. The selected sentences ranged from 7-19 syllables in length (*M*=11.8, *SD*=3.8; see Appendix A).

Of the resulting 984 sentences (12 x 82 participants), 5 sentences (0.5%) were not included in the analyses due to major repetition inaccuracies. A further 320 sentences (33%) contained very minor deviations from the original recording (for example "had" in place of "have") but were deemed acceptable for analysis.

3.3 Procedures

Participants were recorded in a soundproof booth using an external Shure SM58 vocal dynamic microphone and a Marantz PMD-661 MKII solid-state digital recorder with a sampling frequency of 44.1KHz. Participants completed the Elicited Imitation Task on the same day as either the simple or complex parts of the Dinner Table Task, which were conducted on separate, counterbalanced days. In addition to an informed consent form, participants also completed a language background questionnaire and a vocabulary size test.

3.4 Speech rhythm distance measures

One issue with using the rhythm metrics outlined in Table 1 is that each measure provides a one-dimensional picture of a single phonological property underlying speech rhythm. Instead, the use of a distance metric computed on a space defined by two rhythm metrics (e.g., %V and Varco-C) allows us to characterize more robustly the extent to which two languages, or two sets of speech samples (e.g. native vs. non-native speech, controlled vs. spontaneous speech), differ from one another along a rhythm continuum between syllable-timed and stress-timed ends, by including two phonological dimensions known to set syllable-timed and stress-timed systems apart (e.g., vowel reduction and syllable structure

complexity). While previous L2 speech research has used several distance metrics to characterize acoustic distances between distributions of vowel tokens in a two-dimensional F1-F2 vowel space, such as Pillai scores (e.g., Amengual, 2016; Amengual & Chamorro, 2015), Euclidean distances (Flege et al., 1997; Lengeris, 2016) and Mahalanobis distances (Kartushina et al., 2016; Melnik-Leroy et al., 2022), only Euclidean distances have been used in speech rhythm research to gauge distances between languages or speaking styles (Arvaniti, 2012). Mahalanobis distances are a measure of the distance between a point and a distribution in two-dimensional space for multivariate data. They have an advantage over Euclidean distances in that they take into account the shape of the distribution by using a covariance matrix to calculate a distance between a point and the center of a distribution based on standard deviations (Brereton, 2015). Mahalanobis distances have been previously used to characterize distances between distributions of tokens of contrastive vowels on a two-dimensional vowel space defined by first-formant (height) and second-formant (fronting) frequencies (e.g. Kartushina et al., 2016; Melnik-Leroy et al., 2022).

In previous research, pairs of metrics have been used to visualize distances in speech rhythm between native and non-native speech. White and Mattys (2007) employed the pairs %V–Varco-V and rPVI-C–nPVI-V to differentiate between English and Spanish as first and second languages, while Valls Ferrer (2011) found that of the pair %V–Varco-C was the most effective at distinguishing between L1 Spanish, L2 English at two testing times, and native English controls. To our knowledge, Mahalanobis distances have not been used yet to compare between different pairs of metrics to examine differences in speech rhythm between L2-English learners and native speakers in controlled and spontaneous speech. However, they have been used in other linguistic studies comparing vowels in language and dialect acquisition (Riverin-Coutlée et al., 2022). Mora (2021) found that Mahalanobis distance measures were more sensitive in capturing changes over time resulting from phonetic training in the production of vowel contrasts than Euclidean distances or Pillai scores. For rhythm, Mahalanobis distances based on pairs of metrics may help to distinguish more accurately between native speakers and learners of the same language as well as to map changes over time as learners make progress in acquiring the rhythmic characteristics of their L2.

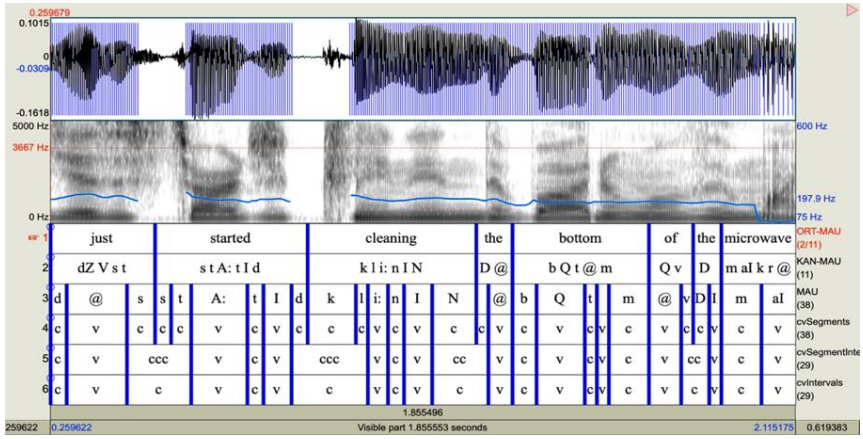
4. Analysis

4.1 Speech segmentation procedures

To obtain the rhythm metrics, spontaneous speech sound files were transcribed into standard orthography, with filled pauses manually annotated. Next, sound-transcript pairs were submitted to the WebMAUS basic automatic aligner (Schiel, 2015) to obtain phonetic transcriptions at the segment level (tier 3) that were used

as input to a Praat script (*CVtierCreator*) that generated C and V intervals (tiers 4-6 in figure 1). After manually adjusting boundaries for segmentation errors, another Praat script (durationAnalyzer_04) was used to obtain raw and rate normalized rhythm metrics (https://www.cl.uzh.ch/de/people/team/phonetics/vdellw/software.html).

Figure 1: Example of segmentation into consonantal and vocalic intervals



4.2 Rhythm metrics

Our choice of rhythm metrics (%V, Varco-V, nPVI-V, Varco-C) and pairs of measures when computing Mahalanobis distances (%V–Varco-V, %V–nPVI-V, %V–Varco-C) took into account differences in articulation rate between native and non-native speakers (Polyanskaya & Ordin, 2019; Munro & Derwing, 2001) by using rate-normalized measures and previous findings by Prieto et al. (2012), who found vocalic measures to maximally distinguish English from Spanish and Catalan, while they did not differ significantly on consonantal measures. Thus, the %V–Varco-V pair, for example, would provide a measure of distance in rhythm space between controlled and spontaneous L2 English speech that would take into account both the amount of vocalic material in speech and the extent to which vocalic segments vary in duration (likely reflecting learners' ability to apply in controlled and spontaneous speech the vowel reduction processes typical of English). White and Mattys (2007) also found %V and Varco-V to discriminate well between learners' L2 speech and their L1 (English, Dutch, Spanish, and French).

Mahalanobis distances between learners and the native speaker control group were calculated using an R script (Borràs, 2022). Distances were calculated unidirectionally between each individual learner's speech sample and the centroid of the native speaker distribution (based on 4 speech samples per speaker in spontaneous speech and 12 sentence speech samples in controlled speech).

All rhythm measures were screened for values above 3 standard deviations from the group's (learners, native speakers) mean, which were excluded from analysis (0.8% excluded for %V, 1.4% for Varco-V, 0.9% for nPVI-V, 0.9% for Varco-C, 2.9% for %V–Varco-V, 2.9% for %V–nPVI-V and 2.3% for %V–Varco-C). Given the huge inter- and intra-speaker variability in speech rate typical of connected speech, extreme values were excluded to ensure that a small number of them did not cause distances between groups to be artificially inflated. Following screening, Mahalanobis distances were positively skewed. A square root transformation was applied, after which visual inspection of histograms and Q-Q plots and Shapiro-Wilk statistics (W>.98) indicated acceptable normality.

5. Results

5.1 Speaking mode and speaker group effects on speech rhythm metrics

In general, for both learners and native speakers, Varco-V and Varco-C were higher in spontaneous than controlled speech indicating greater variability of interval durations. nPVI-V was lower in spontaneous speech for both learners and native speakers indicating lower variability of adjacent vocalic intervals. %V was slightly higher for learners (but not for native speakers) in spontaneous speech (see Table 3).

Table 3: Descriptive statistics for rhythm metrics

Group	**Mode**	**%V**		**Varco-V**		**nPVI-V**		**Varco-C**	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Learner	Controlled	43.5	5.6	61.1	16.5	67.2	17.0	58.1	13.3
	Spontaneous	46.6	3.6	69.4	9.3	63.1	6.3	67.9	7.4
Native	Controlled	43.9	6.4	64.2	14.3	67.1	17.3	57.9	13.1
	Spontaneous	43.3	1.8	71.4	8.1	62.6	3.8	65.0	4.3

We tested these effects by submitting each of the rhythm metrics (%V, Varco-V, nPVI-V, Varco-C) to a linear mixed-effects model with *Mode* (spontaneous, controlled) and *Group* (learners, natives) and their interaction as fixed factors and *Speaker* as a random intercept (in SPSS 26). Tests of fixed effects (see Table 4) showed that spontaneous speech involved higher variability of interval durations for Varco-V, nPVI-V and Varco-C (consistent with a more stress-timed rhythm), whereas for %V the main effect of *Mode* was driven by the significantly higher %V score of learners in spontaneous speech relative to controlled speech (t(1389)=9.32, SE=.32, p<.001), as evidenced by the significant *Mode* x *Group* interaction (native speakers' %V was similar in spontaneous and controlled speech: t(1389)=-.49, SE=1.11, p=.625). None of the main effects of group (or the other *Mode* x *Group* interactions) reached significance, suggesting that L2 English speech did not differ significantly from L1 English speech on these metrics overall.

Table 4: Tests of fixed effects

Source	*df*	%V		Varco-V		nPVI-V		Varco-C	
		F	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Mode	1, 1389	4.59	.032*	21.3	<.001*	7.97	.005*	38.5	<.001*
Group	1, 1389	3.78	.052	2.79	.095	.000	.983	1.52	.218
Mode* Group	1, 1389	9.48	.002*	.068	.794	.117	.732	1.39	.239

.*indicates significance at the .05 level

5.2 Mahalanobis Distances

Mahalanobis distances for the pairs of rhythm metrics %V–Varco-V, %V–nPVI-V and %V–Varco-C (Figures 2, 3, 4) were much larger for spontaneous speech (S) than controlled speech (C), suggesting that the rhythm of L2-learners' English speech did not differ so much from that of native speakers in controlled speech as it did in spontaneous speech (RQ3). It is worth noting that it is mainly %V (rather than segmental duration variability) that contributes to the size of the Mahalanobis distance in spontaneous speech, whereas the small Mahalanobis distances in controlled speech are mainly due to the slightly higher variability of segmental durations of native speakers compared to L2 learners.

Table 5: Descriptive statistics for Mahalanobis distances

Group	Mode	%V–Varco-V			%V–nPVI-V			%V–Varco-C		
		M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
Learner	Controlled	2.73	1.92	2.78	2.14	1.41	2.15	2.09	1.26	2.16
	Spontaneous	10.8	7.67	10.0	11.2	8.53	10.0	13.1	8.25	14.2

Figure 2: %V–Varco-V
N=Native, L=Learner, C=Controlled, S = Spontaneous

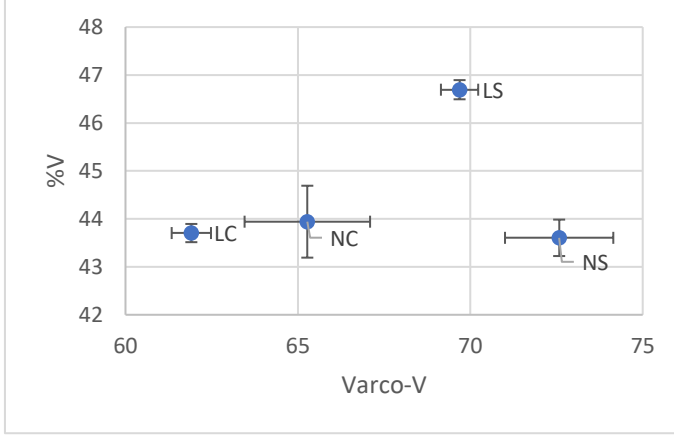


Figure 3: %V–nPVI-V
N=Native, L=Learner, C=Controlled, S = Spontaneous

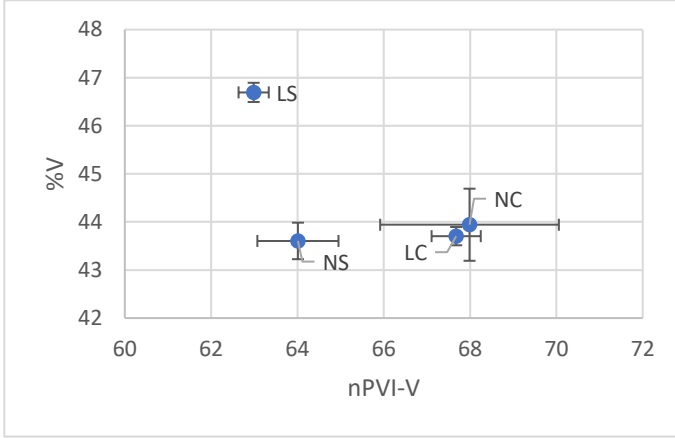
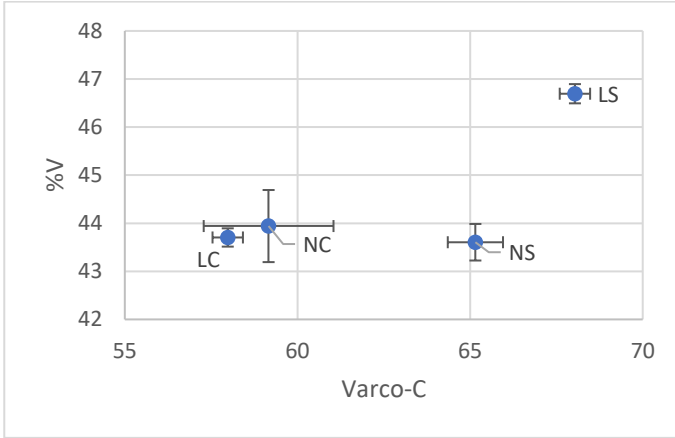


Figure 4: %V–Varco-C
N=Native, L=Learner, C=Controlled, S = Spontaneous



Mahalanobis distances for the pairs of rhythm metrics %V–Varco-V, %V–nPVI-V and %V–Varco-C were submitted to a linear mixed-effects model with *Mode* as the fixed factor and *Speaker* as a random intercept. Tests of fixed effects (see Table 6) indicated that Mahalanobis distances were significantly larger in spontaneous than controlled speech for all pairs of rhythm metrics, suggesting that L2 learners' English speech rhythm was less target-like in spontaneous than controlled speech.

Table 6: Tests of fixed effects

Source	*df*	%V–Varco-V		%V–nPVI-V		%V–Varco-C	
		F	*p*	*F*	*p*	*F*	*p*
Mode	1, 1260	38.4	<.001*	100.8	<.001*	77.1	<.001*

* indicates significance at the .05 level

5.3 Inter-learner consistency in speech rhythm across speaking modes

In order to ascertain whether learners' English speech was consistently syllable-timed (i.e. non-target like) or stress-timed (i.e. target-like) across speaking modes (controlled versus spontaneous speech) we computed *Pearson*-r correlation coefficients on each subject's averaged score for both the individual rhythm metrics and the Mahalanobis distance metric pairings. Only 6 native speakers completed both the controlled and spontaneous tasks, so associations relating to native speaker data are not conclusive and should be corroborated in further research. Table 7 shows that correlations between L2 learners' speech rhythm metrics in controlled and spontaneous speech are relatively weak, suggesting that speaking mode substantially affects learners' speech rhythm. The strongest significant correlation was found for %V, suggesting that %V is a relatively stable rhythm metric across speaking modes in Spanish learners of English, that is, learners' whose speech rhythm is more syllable-timed in controlled speech (higher %V) is also more syllable-timed in spontaneous speech. No significant associations emerged for native speakers. However, Mahalanobis distances in controlled and spontaneous speech are unrelated to one another, suggesting that (as shown in Figures 2-4) the size of the difference between L2 learners' and native speakers' speech rhythm in controlled and spontaneous speech differs substantially at an individual level.

Table 7: Learner correlations between controlled and spontaneous speech

	Metrics(s)	*r*	*p*
Controlled vs. Spontaneous	%V	.377**	<.001
	Varco-V	.195	.080
	nPVI-V	.272^{+}	.014
	Varco-C	.226^{+}	.043
	%V–Varco-V	.110	.329
	%V–nPVI-V	-.031	.781
	%V–Varco-C	.025	.827

An asterisk (*) indicates a correlation that is significant following adjustment for multiple comparisons using Benjamini and Hochberg's False Discovery Rate (FDR) procedure at the 0.05 level, resulting in a threshold *p* value of 0.007 (Benjamini & Hochberg, 1995). A cross ($^{+}$) indicates *p* values that become non-significant following adjustment.

6. Discussion

For individual rhythm metrics, controlled and spontaneous speech exhibited differences in rhythm for both natives and learners consistent with Arvaniti's (2012) findings. Both groups exhibited significantly larger vocalic and consonantal duration variability (Varco-V and Varco-C) in spontaneous than controlled speech. This is typical of speech that is more stress-timed. However, nPVI-V was lower in spontaneous than controlled speech, indicating a more syllable-timed rhythm. Despite such inconsistencies dependent on the rhythm metric used (also attested in Arvaniti, 2012), effects of speaking mode could be consistently observed for both learners and native speakers, suggesting that it may be important to characterize learners' acquisition of L2 speech rhythm by obtaining rhythm metrics from both controlled (read) and spontaneous speech materials.

Although the rhythm metrics we used had been shown to maximally distinguish Spanish and Catalan from English (Prieto et al., 2012), only %V in spontaneous speech clearly discriminated between learners' and native speakers' English. This could be related to the fact that %V was the only non-rate-normalized metric used in this study. While %V is a ratio measure (proportion of vocalic content relative to overall speech content) and has not been shown in L1 speech to vary substantially based on speech rate (Dellwo, 2006), it is possible that vowels are more sensitive to speech rate than consonants. Additionally, in the spontaneous condition %V might have been affected by slight elongations as a means of buying time for lexical search and grammatical encoding.

Our findings (at least for controlled speech) do not provide strong support for the hypothesis that speakers of a syllable-timed L1 will exhibit measurable rhythmic transfer when acquiring the prosodic patterns of a stress-timed L2 (Lai et al., 2013; Ordin & Polyanskaya, 2015). This could be related to the advanced proficiency level of the learner group in our study, or to the use of an elicited imitation task, rather than a read-aloud task, to elicit controlled speech samples. In addition, our group comparison must be taken with caution, given the large sample size difference between learners and the native speaker control group, and the large inter-learner differences in speech rhythm found for all measures, also common in all speakers (Dellwo et al., 2015b).

As in previous research we also found various rhythm metrics to capture differences between spontaneous and controlled speech differently in learners and native speakers. This is not uncommon, as different measures tend to capture separate properties of rhythm (Loukina et al., 2011) and two languages (and by extension the native and non-native varieties of English examined in the current study) can simultaneously be similar and dissimilar as a function of the metric used (Arvaniti, 2009; Nolan & Asu, 2009). The fact that rhythm metrics in spontaneous and controlled speech did not correlate very strongly with one another, and Mahalanobis distances in spontaneous and controlled speech were

unrelated to one another, provide support for the multidimensional nature of speech rhythm, the complexity in capturing speech rhythm differences in spontaneous and controlled speech and in native and non-native speech through vocalic and consonantal interval duration data and, particularly in spontaneous speech, the multiplicity of factors (e.g., phonetic correlates of speaking fluency) that might affect measures of speech rhythm. Such limitations can partly be overcome by using Mahalanobis distances based on selected pairs of metrics to capture degree of nativelikeness of L2 learners' speech rhythm. Based on these distances we provided evidence that learners are more target-like relative to the native speakers in controlled speech than in spontaneous speech.

6.1 Pedagogical Implications

This research has important implications for language learning and teaching. Kohler (2009) asserted that speech rhythm has a "guiding function for the listener", while both Levis (2018) and Quené and Delft (2010) point out that deviations from expected rhythmic patterns can cause challenges for native listeners, highlighting the importance of rhythm in L2 speech intelligibility and effective communication. Rhythm metrics have already been applied to speech recognition software for computer-assisted pronunciation training programs to improve comparisons between learner productions and target speech for "listen and repeat" pronunciation exercises (Bogach et al., 2021). A better understanding of spontaneous rhythm through applying combined metrics could lead to further improvement in these technologies for measuring improvements in spontaneous speech.

Furthering our understanding of rhythm in general is also fundamental to improving the teaching of speech rhythm in classroom settings and designing task-based approaches (Levis, 2018). Finally, it also highlights the related importance of preserving natural rhythm in audio materials for learners, as these materials often use "read aloud" speech samples or stilted, unnaturally slowed speech as listening materials, despite the fact that learners need access to high quality input in order to acquire features of the L2 (Gass & Mackey, 2014).

In general, a major challenge with speech rhythm research comes with connecting measurable differences in rhythm metrics with perceptual data from native listeners. For example, in English %V is relatively easy to conceptualize, as native listeners are looking for vowel reduction cues in unstressed syllables, which happens less in syllable-timed speech. However, it is unclear to what extent listeners are attuned to measures of durational variability, both globally and locally, in the speech of a non-native speaker. A follow-up study could look at which metrics translate well into differences in speech processing and comprehensibility among native speakers. This could be done through native judges listening to true/false questions recorded by non-native speakers and measuring how long it takes them to respond, as an assessment of processing difficulty.

While rhythm has been recognized as an important aspect of pronunciation teaching for many years (e.g. Pike, 1945) and contributes to intelligibility and comprehensibility (Levis, 2018; Quené and van Delft, 2010), learners often do not receive direct training in speech rhythm as part of pronunciation learning (Henderson et al., 2012). For Spanish and Catalan learners of English, this direct training could involve drawing learners' attention to the lengthening of stressed vowel sounds in English, leading to improved recognition and production timing distinctions between stressed and unstressed vowels (Levis, 2018; Low, 2015). Correspondingly, learners can apply explicit attention to reducing unstressed vowels of English, with particular attention to the schwa sound (Munro & Derwing, 2001). Additionally, watching training videos involving a speaker using beat gestures that integrated prosodic prominence into their responses to speaking prompts was shown to lead to decreased ratings of accentedness in learners' spontaneous responses to similar prompts (Gluhareva & Prieto, 2017). With authentic audiovisual materials widely available to students and teachers (Levis, 2018) it is easier than ever to develop high quality materials for the perceptual side of rhythm pronunciation training. However, more studies on the effectiveness of specific pronunciation training methods with a lens on rhythm would lead to more research-informed teaching practices and improved rhythm pronunciation outcomes for language learners.

7. Conclusions

This study has provided preliminary evidence that L2 speakers exhibit different speech rhythm in controlled and spontaneous speech (RQ1). However, our data did not support previous research in that rhythm metrics did not provide a reliable distinction between the English speech of learners and native speakers (RQ2). Mahalanobis distances based on pairs of rhythm metrics specifically selected to maximally distinguish between pairs of languages (e.g. L1-English vs. L2-English) provide a promising way of characterizing prosodic development in L2 learners (RQ3). Finally, on an individual level, learners were not shown to be consistent in their speech rhythm between speech styles (RQ4).

Because of these differences, we should be cautious in generalizing findings from research on the rhythm of read speech to speech rhythm in general. Additionally, further research in the area of quantifying speech rhythm is called for, as improvements in measurement and comparison with native speaker controls for educational purposes could prove to be incredibly helpful for L2 rhythm acquisition. Accurate speech rhythm has been shown to be an important factor in intelligibility and effective communication for L2 speakers, motivating the development of research-informed pedagogical approaches targeted at developing learners' speech rhythm. However, this begins with accurate measurement and quantification of speech rhythm, as well as identifying which components of speech contribute most heavily to a speaker's rhythm.

References

Algethami, Ghazi and Sam Hellmuth. 2023. Methods for investigation of L2 speech rhythm: Insights from the production of English speech rhythm by L2 Arabic learners. *Second Language Research, 0*(ahead of print), 1-26. https://doi.org/10.1177/02676583231152638

Amengual, Mark. 2016. Cross-linguistic influence in the bilingual mental lexicon: Evidence of cognate effects in the phonetic production and processing of a vowel contrast. *Frontiers in Psychology, 7*(617), 1-17. https://doi.org/10.3389/fpsyg.2016.00617

Amengual, Mark, & Pilar Chamorro. 2016. The effects of language dominance in the perception and production of the Galician mid vowel contrasts. *Phonetica, 72*(4), 207-236. https://doi.org/10.1159/000439406

Arvaniti, Amalia. 2009. Rhythm, Timing and the Timing of Rhythm. *Phonetica, 66*(1-2), 46–63. https://doi-org.sire.ub.edu/10.1159/000208930

Arvaniti, Amalia. 2012. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373. https://doi.org/10.1016/j.wocn.2012.02.003

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological), 57(*1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bertinetto, Pier Marco. 1989. Reflections on the dichotomy "stress" vs "syllable timing". *Revue de Phonétique Appliqué, 91-92-93*, 99-129.

Boersma, Paul and David Weenink 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.07, retrieved 28 January 2022 from http://www.praat.org/

Bogach, Natalia, Elena Boitsova, Sergey Chernonog, Anton Lamtev, Maria Lesnichaya, Iurii Lezhenin, Andrey Novopashenny et al. 2021. Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching. *Electronics (Switzerland)*, *10*(3), 1–22.

Borràs, Joan. 2022. Voweldist.R [R script]. Retrieved from: https://github.com/ebrenc/rstats/blob/main/voweldist.R

Brereton, Richard G. 2015. The Mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics*, *29*(3), 143–145. https://doi.org/10.1002/cem.2692

Cantarutti, Marina N. and Beatrice Szczepek-Reed. 2021. Stress and Rhythm. In *The Cambridge Handbook of Phonetics*, 159–184. Cambridge University Press. https://doi.org/10.1017/9781108644198.007

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

Dellwo, Volker. 2006. Rhythm and Speech Rate: A Variation Coefficient for deltaC. In P. Karnowski and I. Szigeti (eds.), *Language and Language-Processing.* Frankfurt am Main.

Dellwo, Volker. 2010. *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence*. Ph.D. dissertation, Universität Bonn, p1-185.

Dellwo, Volker, Adrian Leemann, and Marie-José Kolly. 2012. Speaker idiosyncratic features in the speech signal. In *Proceedings of interspeech 2012*, 1584–1587. Portland, USA. https://doi.org/10.21437/Interspeech.2012-342

Dellwo, Volker, Adrian Leemann, and Marie-José Kolly. 2015a. The recognition of read and spontaneous speech in local vernacular: The case of Zurich German. *Journal of Phonetics*, *48*, 13–28. https://doi.org/10.1016/j.wocn.2014.10.011

Dellwo, Volker, Adrian Leemann, and Marie-José Kolly. 2015b. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

Flege, James Emil, Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437-470. https://doi.org/10.1006/jpho.1997.0052.

Gass, Susan M. and Alison Mackey. 2007. Input, interaction and output: An overview. In K. Bardovi-Harlig & Z. Dornyei (eds.), *AILA Review*, 3-17. Benjamins. https://doi.org/10.1075/aila.19.03gas

Gluhareva, Daria and Pilar Prieto. 2017. Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research, 21*(5), 609-631.

Grabe, Esther and Ee Ling Low. 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. *Papers in laboratory phonology, 7*(1982), 515-546. https://doi.org/10.1515/9783110197105.2.515

Henderson, Alice, Dan Frost, Elina Tergujeff, Alexander Kautszch, Deirdre Murphy, Anastazija Kirkova-Naskova, Ewa Waniek-Klimczak, David Levey, Una Cunningham, and Lesley Curnick. 2012. English pronunciation teaching in Europe survey: Selected results. *Research in language, 10*(1), 5-27. https://doi.org/10.2478/v10015-011-0047-4

Kartushina, Natalia, Alexis Hervais-Adelman, Ulrich Hans Frauenfelder, and Narly Golestani. 2016. Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics, 57*, 21-39. https://doi.org/10.1016/j.wocn.2016.05.001

Kim, Sul-Ki and Tae-Yeoub Jang. 2009. Rhythmic differences between spontaneous and read speech of English. *Phonetics and Speech Sciences, 1*(30), 49-55.

Kohler, Klaus J. 2009. Rhythm in speech and language: A new research paradigm. *Phonetica*, *66*(1–2), 29–45. https://doi.org/10.1159/isbn.978-3-8055-9117-1

Lai, Catherine, Evanini, Keelan, & Zechner, Klaus. 2013. Applying Rhythm Metrics to Non-native Spontaneous Speech. In P. Badin, T. Hueber, G. Bailly, D. Demolin, & F. Raby (eds.), *Proceedings of the ISCA workshop on speech and language technology in education (SLaTE)*, 159–163.

Leemann, Adrian, Marie-José Kolly, and Volker Dellwo. 2014. Speaker-individuality in the time-domain: Implications for forensic voice comparison. *Forensic Science International, 238*, 59–67. https://doi.org/10.1016/j.forsciint.2014.02.019

Lengeris, Angelos. 2016. Comparison of perception-production vowel spaces for speakers of Standard Modern Greek and two regional dialects. *Journal of the Acoustical Society of America. 140*(4), EL314-EL319. https://doi.org/10.1121/1.4964397

Levis, John M. 2018. *Intelligibility, Oral Communication, and the Teaching of Pronunciation* (first ed.). Cambridge University Press. https://doi.org/10.1017/9781108241564

Li, Aike and Brechtje Post. 2014. L2 acquisition of prosodic properties of speech rhythm. *Studies in Second Language Acquisition*, *36*(2), 223–255. https://doi.org/10.1017/S0272263113000752

Loukina, Anastassia, Greg Kochanski, Burton Rosner, Elinor Keane, and Chilin Shih. 2011. Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, *129*(5), 3258–3270. https://doi.org/10.1121/1.3559709

Low, Ee-Ling. 2015. The Rhythmic Patterning of English(es): Implications for Pronunciation Teaching. In M. Reed and J.M. Levis (eds.), *The handbook of English pronunciation*, 125-138. Wiley Blackwell. https://doi.org/10.1002/9781118346952.ch7

Low, Ee Ling, Esther Grabe, and Francis Nolan. 2000. Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, *43*(4), 377–401. https://doi.org/10.1177/00238309000430040301

Mairano, Paolo. 2007. *Rhythm typology: acoustic and perceptive studies.* [Doctoral Dissertation. University of Torino]. HAL. https://theses.hal.science/tel-00654261

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783096473

Melnik-Leroy, Gerda Ana, Rory Turnbull, and Sharon Peperkamp. 2022. On the relationship between perception and production of L2 sounds: Evidence from Anglophones' processing of the French /u/–/y/ contrast. *Second Language Research*, *38*(3), 581-605. https://doi.org/10.1177/0267658320988061

Mora, Joan C. 2021. Assessing L2 vowel production gains after high-variability phonetic training: acoustic measurements vs. perceptual judgements. *Proc. 3rd International Symposium on Applied Phonetics (ISAPh 2021)*, 9-18. https://doi.org/10.21437/isaph.2021-2

Munro, Murray J. and Tracey M. Derwing. 2001. Modeling perceptions of the accentedness and comprehensibility of l2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/S0272263101004016

Nolan, Francis and Eva Liina Asu. 2009. The pairwise variability index and coexisting rhythms in language. *Phonetica*, *66*(1–2), 64–77. https://doi.org/10.1159/000208931

Ordin, Mikhail and Leona Polyanskaya. 2014. Development of timing patterns in first and second languages. *System*, *42*, 244-257. https://doi.org/10.1016/j.system.2013.12.004

Ordin, Mikhail and Leona Polyanskaya. 2015. Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *Journal of the Acoustical Society of America*, *138*(2), 533–544. https://doi.org/10.1121/1.4923359

Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Polyanskaya, Leona and Mikhail Ordin. 2015. Acquisition of speech rhythm in first language. *Journal of the Acoustical Society of America*, *138*(3), 199-204. https://doi.org/10.1121/1.4929616

Polyanskaya, Leona and Mikhail Ordin. 2019. The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language. *Applied Psycholinguistics*, *40*(3), 795–819. https://doi.org/10.1017/S0142716419000067

Polyanskaya, Leona, Mikhail Ordin, and Maria Grazia Busa. 2017. Relative Salience of Speech Rhythm and Speech Rate on Perceived Foreign Accent in a Second Language. *Language and Speech*, *60*(3), 333–355. https://doi.org/10.1177/0023830916648720

Prieto, Pilar, Maria del Mar Vanrell, Lluïsa Astruc, Elinor Payne, and Brechtje Post. 2012. Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, *54*(6), 681–702. https://doi.org/10.1016/j.specom.2011.12.001

Quené, Hugo and L. E. Van Delft. 2010. Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*(11–12), 911–918. https://doi.org/10.1016/j.specom.2010.03.005

Ramus, Franck and Jacques Mehler. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, *105*(1), 512–521. https://doi.org/10.1121/1.424522

Ramus, Franck, Marina Nespor, and Jacques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*(1), 265–292. https://doi.org/10.1016/s0010-0277(00)00101-3

Riverin-Coutlée, Josiane, Johanna-Pascale Roy, and Michele Gubian. 2022. Using Mahalanobis Distances to Investigate Second Dialect Acquisition: A Study on Quebec French. *Language and Speech*, *66*(2), 291–321 https://doi.org/10.1177/00238309221097978

Roach, Peter. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal. (ed.), *Linguistic Controversies: Essays in Linguistic Theory and Practice*, 73-79. London: Edward Arnold.

Schiel, Florian. 2015. A statistical model for predicting pronunciation. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, Rachel, M. MacMahon, J. Stuart-Smith, & J. Scobbie (eds.), *Proceedings of the 18th International Congress of Phonetic Sciences 2015, Glasgow, UK (ICPhS 18)*.

Ur, Penny. 1981. *Discussions that work: Task-centered fluency practice*. Cambridge University Press.

Valls-Ferrer, Margalida. 2011. *The development of oral fluency and rhythm during a stay abroad period*. [Doctoral dissertation, Universitat Pompeu Fabra].

Van Maastricht, Lieke, Tim Zee, Emiel Krahmer, and Marc Swerts. 2021. The interplay of prosodic cues in the L2: How intonation, rhythm, and speech rate in speech by Spanish learners of Dutch contribute to L1 Dutch perceptions of accentedness and comprehensibility. *Speech Communication*, *133*, 81–90. https://doi.org/10.1016/j.specom.2020.04.003

White, Laurence and Sven L. Mattys. 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

Wu, Shu-Ling, Yee Pin Tio, and Lourdes Ortega. 2022. Elicited imitation as a measure of L2 proficiency: new insights from a comparison of two L2 English parallel forms. *Studies in Second Language Acquisition*, *44*(1), 302–302.

Acknowledgments

We are grateful to Miren Adrian, Cristina Aliaga-Garcia, Gonzalo Bermejo Miranda, Athenea Botey, Josh Frank, Natalia Fullana, Valeria Galimberti, Ingrid Mora-Plaza, Mireia Ortega and Gisela Sosa for their contributions in data collection and data processing and analysis. We would like to thank Volker Dellwo for his advice on the use the Praat plug-ins for obtaining the rhythm metrics and Joan Borràs for his help in implementing the rhythm distance measures. We are grateful to two RiL anonymous reviewers for their very helpful and insightful comments and suggestions on a previous draft of this manuscript and the audience at ACCENTS 2022 for their inspiring questions and comments. This study was supported by grant PID2019-107814GB-I00 from the Spanish Ministry of Science, Innovation and Universities. The participants in the study gave their written informed consent to participate in the study. The study protocol adhered to the good practices of data collection, anonymization, processing, and storage of the Institutional Review Board of the University of Barcelona (IRB0003099).

Appendix A: EIT Sentences

Num	Sentence	Syl
1	I have to buy a bus pass.	7
2	The red book is on the table.	8
3	The parks in this town are old.	7
4	He takes a shower every morning.	9
5	Where did you think you were going tonight?	10
6	It is possible that it will rain tomorrow.	12
7	The houses are very nice but too expensive.	12
8	That restaurant is supposed to have very good food.	13
9	You really enjoy listening to country music, don't you?	14
10	She just started cleaning the bottom of the microwave.	14
11	The most fun I've ever had was when we went to the opera.	16
12	There are a lot of people who don’t eat anything at all in the morning.	19

Appendix B: Generalized linear mixed-models parameter estimates for individual rhythm metrics

Table B1: %V

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	43.334	1.040	41.685	<.001	41.294	45.373
Mode	.543	1.111	.488	.625	-1.637	2.722
Group	3.312	1.088	3.044	.002	1.178	5.446
Mode x Group	-3.563	1.157	-3.079	.002	-5.833	-1.293

Table B2: Varco-V

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	71.841	2.668	26.926	<.001	66.507	77.075
Mode	-7.292	3.205	-2.275	.023	-13.580	-1.004
Group	-2.385	2.796	-.853	.394	-7.870	3.101
Mode x Group	-.875	3.348	-.261	.794	-7.442	5.692

Table B3: nPVI-V

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	62.541	2.783	22.474	<.001	57.082	68.000
Mode	5.408	3.276	1.651	0.099	-1.017	11.834
Group	0.625	2.912	0.214	0.830	-5.088	6.337
Mode x Group	-1.170	3.417	-0.342	0.732	-7.872	5.532

Table B4: Varco-C

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	64.42	2.2236	28.971	<.001	60.058	68.782
Mode	-6.846	2.6098	-2.623	0.009	-11.966	-1.727
Group	3.445	2.3301	1.478	0.14	-1.126	8.016
Mode x Group	-3.21	2.7232	-1.179	0.239	-8.553	2.132

Appendix C: Generalized linear mixed-models parameter estimates for Mahalanobis distances

Table C1: Mahalanobis distances %V–Varco-V

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	2.952	0.205	14.427	<.001	2.551	3.354
Mode	-1.451	0.234	-6.194	<.001	-1.911	-0.991

Table C2: Mahalanobis distances %V–nPVI-V

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	3.032	0.149	20.319	<.001	2.739	3.325
Mode	-1.713	0.171	-10.041	<.001	-2.047	-1.378

Table C3: Mahalanobis distances %V–Varco-C

Source	*β*	*SE*	*t*	*p*	*95% CI*	
					Lower	Upper
Intercept	3.220	0.193	16.719	<.001	2.842	3.598
Mode	-1.925	0.219	-8.778	<.001	-2.355	-1.495