9 Assessing Second Language Pronunciation: Distinguishing Features of Rhythm in Learner Speech at Different Proficiency Levels

Evelina Galaczi, Brechtje Post, Aike Li, Fiona Barker and Elaine Schmidt

Introduction

Ongoing globalization in the 21st century, as well as the primary role of English as a lingua franca in educational and professional contexts, has brought to the forefront the fundamental role of second/foreign language (L2) pronunciation, due to the growing need for interlocutors from different first language (L1) backgrounds to engage in meaningful and intelligible communication in English. While pronunciation is clearly an important skill, and insufficient pronunciation ability is detrimental to the intelligibility of speech, it has a surprisingly minor role in theoretical L2 models, in assessment scales of speaking, and in practical assessment and teaching contexts involving examiners, teachers and learners. In terms of theoretical models, pronunciation ability is only minimally dealt with in the widely accepted framework of communicative language ability (Bachman, 1990), where pronunciation (phonology) is neglected and categorized together with graphology, despite the fundamental difference between these two constructs (Isaacs, 2014). The backstage role of pronunciation in theoretical models is also reflected in some well-known assessment scales, where pronunciation tends to be captured inconsistently or to be entirely absent. For example, the ACTFL Proficiency Guidelines for Speaking (ACTFL, 2012) do not include descriptors for pronunciation at every level. Similarly, the Common European Framework of

Reference for Languages (CEFR) omits pronunciation as a criterion from the Spoken language use scale (Council of Europe, 2001: 28–29), attributing this decision to the difficulty in distinguishing across pronunciation ability levels in the same way as other language skills and the difficulty in interpreting descriptors of pronunciation consistently across languages (North & Hughes, 2003: 6). Pronunciation is, instead, captured in the phonological control scale (Council of Europe, 2001: 117), which includes only brief descriptors. It does not consistently describe the construct of interest, and does not distinguish between CEFR Levels C1 and C2 (see Harding, this volume).

The underdeveloped nature of the pronunciation construct is also seen in the challenge presented by pronunciation for teachers and examiners: studies have indicated that examiners find the assessment of pronunciation to be more challenging than that of other skills, and that they tend to be more confident when making global judgements of intelligibility than when making judgements about micro-level segmental and prosodic features (e.g. Brown & Taylor, 2006; Isaacs *et al.*, 2015; Yates *et al.*, 2011). Teachers have also been found to lack training and confidence in their pronunciation expertise (Levis, 2006).

The under-researched nature of the pronunciation construct is reflected in L2 acquisition research as well, where relatively little is known about how segmental and prosodic features develop over time. What is known is that cross-linguistic differences in segmental and prosodic proficiency are apparent due to language transfer (e.g. difficulties with the /l/ and /r/ contrast for Japanese learners), especially at lower levels of proficiency, reflecting properties of the L1 (Major, 2008). While the L1 influence is important, research has also indicated that acquisition does not necessarily proceed in a uniform fashion, and that some features are subject to L1 transfer while others, such as accentual lengthening, show common developmental paths across languages (e.g. Li & Post, 2014). The picture is further complicated by the fact that proficiency emerges with the acquisition of phonology, morphosyntax and information structure and the mapping between them (Post *et al.*, 2010) but phonological and prosodic acquisition can be out of step with acquisition in other areas of language competence at higher levels.

To sum up, there are difficulties with the theoretical conceptualization and practical operationalization of L2 pronunciation. One useful line of enquiry to pursue in making pronunciation a well-understood and integral part of learning, teaching and assessment contexts is to better understand the pronunciation features of learners with different L1s at different proficiency levels. The present study aims to contribute to this need through providing an in-depth micro-analytic investigation of prosodic features observed in learner speech. The prosodic features of interest relate to the rhythm of speech, chosen for investigation here largely due to the under-researched nature of rhythm in L2 speech, which is at odds with its important role in comprehensibility (Isaacs & Trofimovich, 2012). Through its dual interdisciplinary phonetics and assessment perspective, the study aims to establish a profile of the rhythmic properties of learner speech at different proficiency levels, which can in turn contribute towards a more comprehensive definition and operationalization of the construct of L2 pronunciation. A broader aim is to raise awareness about micro-level features of rhythm and prosody which play a role in learner speech, and which teachers and assessors are likely to benefit from.

Role of Rhythm in English Speech

Rhythm has traditionally been seen as a key distinguishing feature between languages, with stress-timed and syllable-timed languages regarded as distinct on the basis of differences in rhythmic properties (Abercrombie, 1967; Pike, 1945). In stress-timed languages the durations between each stressed syllable tend to be approximately equal, whereas in syllable-timed languages the durations of syllables tend to be approximately equal. In stress-timed languages, stressed syllables are significantly longer than unstressed syllables, and unstressed syllables which occur between consecutive stressed syllables are compressed (e.g. through vowel reduction) to fit into the time interval. For example, in the phrases ‘i LIKE to TRAvel’ and ‘i LIKE very much to TRAvel’, the unstressed words/syllables (in lower case letters) are reduced and shorter than their stressed counterparts and, importantly, the intervals taken by the unstressed syllables are of approximately equal duration. In syllable-timed languages, in contrast, syllables have more equal duration and prominence, with little or no vowel reduction. Table 9.1 provides a summary of key properties that have been associated with the stress-timed versus syllable-timed rhythm classification.

Although the rhythmic distinctions between stress-timed and syllable-timed languages have been empirically supported (e.g. Ramus *et al.*, 2003), many instrumental studies have failed to find constant and systematic evidence to support the dichotomous approach to categorizing languages (e.g. Bolinger, 1965; Roach, 1982). In addition, it is clear that rhythm manifests itself along a number of phonetic dimensions, including duration, pitch and loudness. As a result, rather than categorizing languages in terms of two (or three) distinct rhythm classes, crosslinguistic differences in rhythm are now accounted for in terms of a continuous model of rhythm in which a combination of language-specific properties (besides more general factors such as speaking rate) result in different rhythmic patterning gradients along a continuum. Depending on these properties, an individual language will fall at a particular point along the continuum (e.g. Dauer, 1983; Prieto *et al.*, 2012). Bearing this in mind, we adopt the terms ‘stress-timed’ and ‘syllable-timed’ in this chapter for ease of reference, and only operationalize speech rhythm in terms of duration in the first instance.

Table 9.1 Stress-timed and syllable-timed languages: Rhythmic properties

Rhythmic property	*Stress-timed languages (e.g. Dutch, English, German)*	*Syllable-timed languages (e.g. French, Mandarin, Spanish)*
Vowel reduction Vowels in unstressed position tend to be shorter and to converge towards the central/neutral vowel schwa /ə/.	Evident reduction of unstressed vowels.	Vowel reduction is not evident.
Syllable structure complexity Number of consonants allowed in a syllable.	Complex consonant clusters, therefore high amount of consonants in speech, e.g. C, CC and CCC.	Open syllables (CV) are far more common than complex syllable structures, therefore a lower amount of consonants in speech.
Durational marking of accentuation The lengthening of accented syllables compared to unaccented syllables.	Large durational difference between accented syllables and unaccented ones.	Little durational difference between accented and unaccented syllables.
Final lengthening The lengthening of phrase-final and utterance-final syllables compared to non-final syllables.	Final syllables are lengthened compared to non-final syllables.	Little durational distinction between final and non-final syllables.

Note: C = consonant; V = vowel.

Apart from such language-specific properties, rhythm is affected by non-linguistic factors such as speech rate (e.g. Prieto *et al.*, 2012). Figure 9.1 provides an illustration of how the consonants and vowels that are produced in the speech stream form intervals of different types with variable durations, which in turn create the rhythm of English speech. The example chosen is the fragment '… if you have a team leader, strictly speaking …' (see Figure 9.2 for the speech pressure wave and spectrogram). As Figure 9.1 shows, syllable boundaries do not necessarily coincide with the edges of words (e.g. in 'if you' the /f/ forms a syllable with the following word /juː/), nor with the intervals that contain the sequences of vocalic or consonantal material (e.g. in 'strictly', the second consonantal interval straddles a syllable boundary). The figure also illustrates that the intervals are delimited by major phrase boundaries (marked by a silent pause # in this example). The effect of syllable structure can be seen when we compare intervals that contain complex

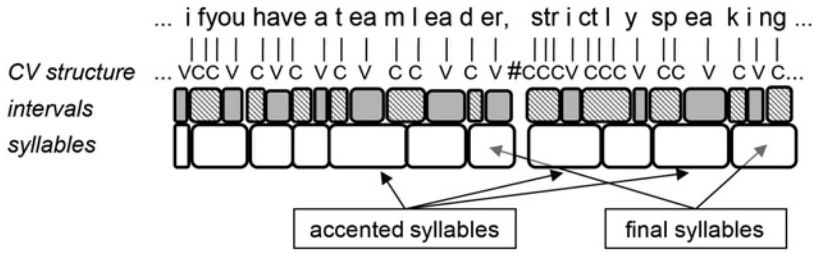

Figure 9.1 Stress timing: An example of syllable durations and vocalic and consonantal intervals

Notes: (i) C = consonant; V = vowel; grey boxes = vocalic intervals, striped boxes = consonantal intervals; # = silent pause. (ii) Vocalic interval = section of speech between vowel onset and offset; consonantal interval = section between vowel offset and onset. (iii) Durations are shown approximately to scale.

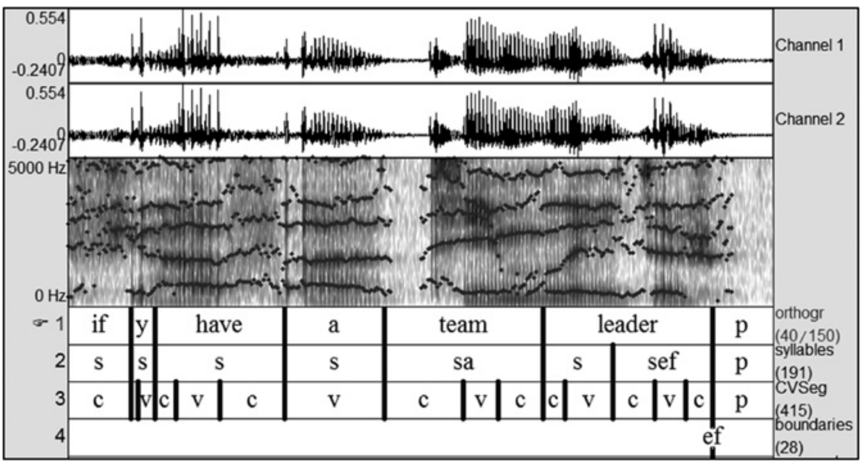

Figure 9.2 Example of the segmental and prosodic labelling

consonant clusters (e.g. /str/ in 'strictly') with those that contain singleton consonants (e.g. /t/ in 'team'). As a result, the duration of the consonantal intervals is quite variable, and will be much more variable than in a language that only allows syllables with singleton consonants. Also, accented syllables with full vowels tend to be considerably longer than unaccented syllables which often contain reduced vowels (e.g. /ɪ/ in 'strict', 'if' and '-ly'). The effect of boundary lengthening can be seen in the longer duration of syllables that precede a major phrase boundary (compare the duration of the schwa vowel in 'a' and '-der'). Although the durational marking of accented syllables and phrase boundaries is very common across the languages of the world, the amount of lengthening varies considerably, and Figure 9.1 exemplifies that that effect is particularly strong in English.

Rhythm Metrics

A number of metrics have been developed to quantify rhythm in (learner) speech (for a review, see White & Mattys, 2007).

%V (Proportion of vocalic material in speech)

Ramus *et al.* (1999) proposed several utterance-level measures of rhythm by dividing speech into vocalic and consonantal parts and computing the proportion of the vocalic interval duration in speech, expressed as a percentage from the total utterance duration: %V (i.e. comparing the amount of material in the grey against the striped boxes in Figure 9.1). In a stress-timed language such as English, %V would typically be lower due to vowel reduction and to the presence of consonant clusters, compared to a syllable-timed language such as Spanish.

In English, L2 learners may incorrectly insert vowels to break up consonant clusters, and as such would increase the proportion of vocalic material in their speech. For example, Japanese learners of English who are struggling with this pronunciation feature may pronounce the word 'Christmas', which has several consonant clusters, as /kurisumasu/. Learners may, in addition, have trouble with vowel reduction in unstressed positions, and make vowel intervals that contain unaccented syllables too long.

Varco-V and Varco-C (variability in vocalic or consonantal duration)

To measure the variability in vocalic or consonantal interval duration (i.e. quantifying the amount of variability within either the grey or the striped intervals in Figure 9.1), Dellwo (2006) developed Varco-C, which calculates the standard deviation of consonantal interval duration (normalized for speech rate), as well as its vocalic counterpart Varco-V. Varco measures the variability globally as an average across the whole speech of a speaker, in contrast to other measures (discussed below) which have a more local focus. Typically, this variability is expected to be larger in stress-timed languages, such as English, due to the higher variety in syllable structures associated with a stress-timed language (e.g. some syllables have complex consonant clusters and some simple structures), and greater accentual and final lengthening. In syllable-timed languages, in contrast, a large proportion of syllables have a simple CV structure and successive syllables are more similar in length, leading to lower variability values (Low, 2006).

In English, L2 speech may show lower variability in vocalic and consonantal durations, partly due to L1 transfer from a syllable-timed language. This will, in part, be due to processing and articulation difficulties associated with differences between the syllable structures in the L1 and L2, but also with differences in accentual and boundary marking. In addition to possible transfer effects, accentuation has been shown to be correlated with proficiency level (Kang, 2013; Kang *et al.*, 2010) and with comprehensibility and accent ratings (Isaacs & Trofimovich, 2012) independent of the L1. Low-proficiency learners have been reported to overuse accentuation, since they accent lexical items regardless of their function or information load (Kang, 2010). This is likely to have knock-on effects on speech rhythm.

PVI (Pairwise Variability Index)

Other measures capture rhythm more locally by focusing on the degree of durational difference between neighbouring intervals. One such measure is the PVI (Pairwise Variability Index) metric which calculates the mean of the durational differences between successive temporal intervals in an intonation phrase (a stretch of speech with its own intonation contour). The raw PVI (rPVI-C for consonantal intervals and rPVI-V for vocalic intervals) contrasts with its normalized counterpart which controls for the effect of speech rate (nPVI-C and nPVI-V) (Grabe & Low, 2002; Low *et al.*, 2000). A stress-timed language would be expected to have high PVI values and a syllable-timed language low PVI values, as confirmed by Low *et al.* (2000) and Low (1998) who compared British English (stress-timed) and Singapore English (syllable-timed) and found that the British English speakers exhibited a significantly higher variability in duration between successive vowels than the Singapore English speakers. In English L2 speech, learners who are producing more syllable-timed speech would be expected to show lower variability in durations between neighbouring intervals.

Prosody, Rhythm and Second Language English Learners

Prosody, which comprises rhythm alongside intonation, tone, accentuation and boundary marking, has been empirically shown to play a fundamental role in comprehensibility (e.g. Anderson-Hsieh *et al.*, 1992; Kang *et al.*, 2010; Munro & Derwing, 1995; Pickering, 2001). In other words, a speaker with otherwise good articulation may be difficult to understand because of weak prosody. Stress – the most widely studied prosodic feature in L2 speech – has consistently been found to relate to L2 comprehensibility, both at word- and sentence-stress level (Field, 2005; Hahn, 2004; Kang *et al.*, 2010). Discussions in the area of English as a lingua franca have also indicated that communication breakdowns could be due to misplaced sentence stress (Jenkins, 2002).

Rhythm has been investigated less extensively than stress as a prosodic feature of L2 speech. Low (2006) suggests that the rhythm of a stress-timed language would be more difficult to acquire than the rhythm of a syllable-timed language, due to the need to reduce vowels and compress syllables in stress-timed languages. This hypothesis finds support in various empirical studies: Gutiérrez-Díaz (2001) found that advanced Spanish learners of English produced English with a stressed/unstressed syllable durational ratio, which was mid-way between the ratios for Spanish and English, suggesting that stress-timing posed a problem for those learners. Similarly, Trofimovich and Baker (2006) reported a significant difference between inexperienced and moderately experienced learners and native speakers of English in terms of their stressed/unstressed syllable ratio, again showing that learners had difficulty producing stress timing. Vowel reduction – a fundamental component of rhythm in English – has also been related to measures and perceptions of accentedness and comprehensibility in learners (Trofimovich & Isaacs, 2012).

Other studies have included rhythm as part of investigations focusing on differences in learner speech across proficiency levels. In this respect, Iwashita *et al.* (2008), in an important first attempt to provide an in-depth empirical analysis of learner speech, focused on distinguishing levels of proficiency through a range of linguistic features in the context of the TOEFL iBT speaking test. The researchers utilized a range of pronunciation measures (alongside measures of grammatical accuracy and complexity, vocabulary and fluency) and focused on the pronunciation of word and sub-word level features, on intonation and on rhythm using auditory coding of features (e.g. targetlike/non-targetlike) and not instrumental measurements. A significant difference between levels was found only with the production of targetlike syllables, but the authors reported high correlations between rhythm and proficiency level, with appropriate rhythm associated with higher proficiency learners.

Kang (2013) analyzed a range of linguistic features at CEFR Levels B1 to C2 in a set of Cambridge English speaking test performances. The speech analysis program *Praat* (Boersma & Weenink, 2010) was used alongside listener coding of measures to analyze objective pronunciation measures, such as proportion of words with prominent stress, number of prominent syllables per run, overall pitch range and a range of tones. The findings indicated that there are objectively measured differences between high- and low-proficiency learners, but not necessarily between adjacent levels. The author further showed that, as proficiency increased, the proportion of stressed words within a sentence decreased, thus supporting prior research indicating that low-proficiency learners stress items regardless of their function or importance (Kang, 2013). In a similar line of research, Isaacs and Trofimovich (2012) set out to produce an empirically based rating scale for pronunciation and identified a subset of features which best distinguish between three levels of L2 comprehensibility, using both auditory and instrumental measures. The authors noted a strong relationship between word stress and vowel reduction and raters' judgements in the sample, which, it needs to be noted, was limited to one L1 group (French).

The development of the rhythm metrics outlined in the previous section has given rise to investigations of L2 speech rhythm in a quantitative and systematic way. However, only a few studies have investigated rhythmic differences between L2 learners of different proficiency levels, or with different L1 backgrounds (e.g. Guilbault, 2002; Gut, 2009; see Li & Post, 2014, for review). These studies have provided inconsistent evidence for rhythmic differences in L2 speech depending on level of proficiency or L1.

Study Aim and Research Questions

As noted in the literature review, few studies have systematically examined how rhythm is displayed by L2 learners at different proficiency levels and from controlled L1 backgrounds, and even fewer studies to date have employed the rhythm metrics used here in the context of learner speech at different proficiency levels. The aim of this study, therefore, is to offer a more comprehensive empirically based investigation of rhythm in L2 speech than that given in previous studies in order to establish to what extent rhythmic measures can discriminate between proficiency levels. We do so through a small-scale quantitative investigation of objectively measurable micro-level prosodic rhythmic features in the speech of learners at different proficiency levels (reported as CEFR levels) and through controlling for learner L1 background. The study aims to provide more granularity in the analysis by moving from a judgement that a specific feature is not targetlike (e.g. Iwashita *et al.*, 2008), to an investigation of what makes it not targetlike at different CEFR levels. The main motivation is the need to better understand L2 pronunciation as a construct and to provide practical findings that can inform learning, teaching and assessment. A broader aim of the study is to contribute a cross-disciplinary perspective to L2 pronunciation through collaboration between L2 phoneticians and language assessment specialists. The following four research questions guide the study; the first two deal with pronunciation across proficiency levels while the last two deal with pronunciation across L1s.

(1) How reliably can levels of L2 pronunciation ability be discriminated across CEFR Levels A1–C2 using a set of rhythmic measures?
(2) Which rhythmic measures have the highest discriminative properties for particular proficiency levels?
(3) How far do rhythmic measures display different patterns for learners of different L1 backgrounds?
(4) Which rhythmic measures display the largest differences for particular L1s?

Methodology

Speech samples

Speech samples drawn from speaking test performances of 20 English learners from three typologically different L1 backgrounds and six CEFR levels were used in this study (see Table 9.2).

Approximately 60 seconds of speech were used per learner, taken from a Question-and-Answer task, where the learners responded to a series of questions and produced extemporaneous speech. The speech samples at CEFR Levels A2–C2 were taken from Cambridge English face-to-face speaking tests, whereas at Level A1 performances were extracted from a computer-delivered speaking test. In both test formats the same task type was used to minimize any differences due to a method effect.

The participants represented ‘average’ learners at each CEFR level based on their pronunciation score; that is, they were not borderline within their CEFR level. Borderline test takers with marks at the top or bottom of the scale would have been likely to show pronunciation features typical of the adjacent proficiency levels and were not deemed suitable for analysis. Rater effects were minimized by using Fair Average marks generated by Facets (Linacre, 1989) which were based on multiple marks from a group of accredited experienced examiners (marks were provided by Cambridge English).

Two studies served as the basis of the empirical investigation reported here. In Study 1, speech was used from 12 L1 Spanish speakers spanning Levels A1 to C2 (two learners at each level) to examine the variability in learner performances across proficiency levels, while controlling for L1 background; in Study 2, speech was used from L2 speakers from Korean, Spanish and German L1 (two each at Levels B1 and B2) to compare L1 effects. The three languages were chosen to be typologically different: German is stress-timed (Kohler, 1982); Spanish is syllable-timed (Pike, 1945); whereas Korean is generally considered to be neither stress-timed nor syllable-timed (Seong, 1995). CEFR Levels B1 and B2 formed the basis of Study 2.

Table 9.2 Dataset

Level	*L1*		
	German	*Spanish*	*Korean*
A1	–	2	–
A2	–	2	–
B1	2	2	2
B2	2	2	2
C1	–	2	–
C2	–	2	–

Speech measures

The 20 speech samples were analyzed for a number of measures which quantify different aspects of speech rhythm, and which were chosen based on previous research linking these measures to proficiency level or L1 differences. Table 9.3 provides a definition of these measures and their operationalization. Taken together, an analysis of learner speech across a range of proficiency levels using these measures should reveal how successful learners are at each level in producing speech that has stress-timing features.

Analysis

Syllable durations and durations of consonantal and vocalic intervals were extracted using *Praat* (Boersma & Weenink, 2010). An example of the segmental and prosodic labelling in *Praat* is given in Figure 9.2.

The first tier of the figure (fourth row down) contains orthographic transcription. The second tier (fifth row down) was used to calculate accentual and final lengthening. Each syllable is marked as unaccented and non-final (labelled as 's'), accented ('sa'), final ('sef'), accented and final ('safe') or hesitated ('sx'). Hesitated syllables were excluded from the analysis. The key segmentation criterion for syllabification of the speech produced by the study participants was Gussenhoven and Jacob's (2005) version of the Maximum Onset Principle (Pulgram, 1970), which would syllabify the word 'leader' as 'lea.der' instead of, for instance, 'lead.er'. Establishing this criterion was important to ensure consistency in the measurement of syllable durations across the different speech samples. The second tier was also used to calculate speech rate by counting the number of syllables that were realized in each 60-second speech sample. The third and fourth tier together provided the information required for the calculation of the rhythm metrics. In the third tier, each vowel and consonant was segmented primarily by visual inspection of the speech waveforms and wideband spectrograms with reference to standard criteria (e.g. Peterson & Lehiste, 1960; White & Mattys, 2007). The fourth tier contains the phrasing information, that is, the beginning and end of an intonation phrase. Within each intonation phrase, consecutive consonant/vowel intervals were merged into vocalic and consonantal intervals.

Inter-coder agreement was measured for a subset (15%) of the data (one speaker per language). The samples used for inter-coder estimation were segmented and labelled by three annotators. The inter-coder agreement was 97% (calculated as number of codes of agreement/total number of codes).

Once the relevant measures were derived, statistical analyses were performed using IBM *SPSS 20*. The data were analyzed using multivariate analysis of variance (MANOVA) with factors Language background (Spanish, German, Korean) and Proficiency level (A1, A2, B1, B2, C1, C2), and fixed factors Rhythm metrics (nPVI-V, Varco-V, Varco-C, %V, rPVI-C, nPVI-C).

Table 9.3 Speech measures used in this study

Measure		*Operationalization*
Speaking rate		Number of syllables in 60 sec sample.
Accentuation		Durational differences between accented and unaccented syllables. For example, in the sentence 'There are some CHILdren in this PHOto', the duration of the accented syllables (given in capital letters) should be relatively longer than the duration of the unaccented syllables. The appropriateness of accent placement was not considered; only the durational difference between accented and unaccented syllables was.
Boundary marking		Durational difference between phrase-final and non-phrase final syllables (in an intonation phrase). For example, in the sentence 'I LIKE playing drums', the duration of the final syllable (drums) in the intonation phrase should be longer than if it was followed by another word in the phrase, as in 'I LIKE playing drums at home.'
Rhythm metrics	%V	Proportion of vocalic material in speech. For example, the word 'stop' would have a lower proportion of vocalic material than the incorrectly pronounced /sətop/ with a schwa sound inserted to break up the consonant cluster. Function words, which typically have reduced vowels, may also add to the proportion of vocalic material in L2 speech if the vowels are not reduced.
	Varco-V and Varco-C	Variability in vocalic/consonantal interval duration (standard deviation divided by mean). In contrast to PVIs (below), this is averaged over all utterances. For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.
	rPVI-C	Durational differences between adjacent consonantal intervals in an intonation phrase (raw). In contrast to Varco, all variation here is measured locally, by comparing each interval with its successive interval. For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.
	nPVI-V	Durational differences between adjacent vocalic intervals in an intonation phrase (normalized for speaking rate). For example, in 'YESterday' the vocalic interval in 'yes' is much longer than in 'ter' (because of vowel reduction), followed by a longer vocalic interval in 'day'.
	nPVI-C	Durational differences between adjacent consonantal intervals in an intonation phrase (normalized for speaking rate). For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.

Results

Study 1: Differences across proficiency levels

We will first present the results of the 12 Spanish learners across six different proficiency levels, starting with the rhythm metrics. Of the six rhythm metrics of interest here, three (seen in Figure 9.3) were found to be discriminative for CEFR level: %V (proportion of vocalic material in speech), Varco-C (variability in consonantal interval duration), and nPVI-C (durational differences between adjacent consonantal intervals), as indicated in a MANOVA with fixed factor Level (A1–C2), which showed a significant main effect for %V, $F(5, 141) = 2.33$, $p < 0.05$; Varco-C, $F(5, 141) = 4.02$, $p < 0.01$; and nPVI-C, $F(2, 148) = 2.65$, $p < 0.05$. Post hoc tests (Scheffe) showed that for Varco-C, the effect was attributable to a marginal difference between A1 and A2 ($p = 0.078$), between A2 and C1 ($p < 0.05$), and A2 and C2 ($p < 0.05$); and for nPVI-C, A2 and C2 differed ($p < 0.05$); for %V none of the individual comparisons reached significance.

Overall, the consonantal metrics showed a steady increase over proficiency levels, with the exception of Level A1. This general upward trend indicates that the variability in consonantal interval durations as well as the durational differences between adjacent consonantal intervals increased, suggesting that as learners improved in proficiency, they became more adept at dealing with both single consonants and more complex consonant clusters, and showed a more or less steady progression in consonantal variability. Interestingly, the Varco-C measure was found to be stable at the higher C1/C2 levels, while nPVI-C increased at those levels (but not significantly). This indicates that C level L2 speakers are appropriately varying the durations of consonantal sequences in their speech globally (i.e. across the utterance).

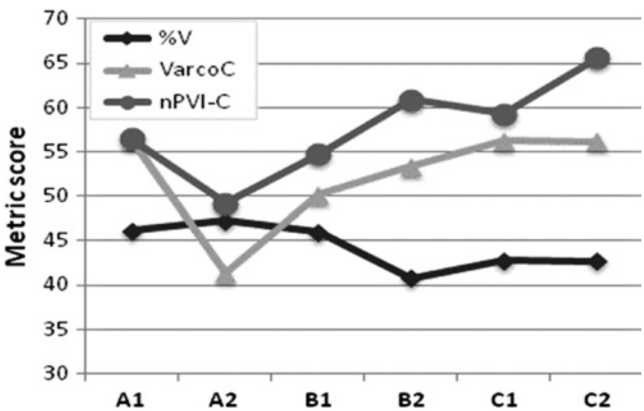

Figure 9.3 Mean metric values for %V, VarcoC and nPVI-C at 6 CEFR levels

This is likely to primarily reflect the adjustments that need to be made when consonantal sequences vary in complexity – and results in the same rhythmic profile at the two highest levels in Varco-C. However, when the variation in the duration of consonantal intervals is considered more locally (i.e. in terms of pairwise variability between consonants), the values are still changing, which could suggest that more localized variation in duration due to factors like accentuation and boundary marking is still developing further at this level. In the case of the vocalic metric %V (the proportion of vocalic material in speech), a shift occurred between B1 and B2, indicating that as learners developed in proficiency between these two levels, they became more adept at producing reduced vowels, and at decreasing the number of (incorrectly inserted) vowels.

Moving on to the prosodic lengthening measures, the durations of syllables overall and by syllable type were investigated. As a starting point, the mean number of syllables produced by learners at each level was calculated (Figure 9.4). Even though this measure is not strictly speaking an acoustic-phonetic measure, and is typically considered a measure of fluency, it is a useful initial gauge of the development of learners across levels. Findings indicated a clear progression of mean number of syllables across proficiency levels, with the exception of Levels A1 and A2. As the learners in the sample developed in proficiency, they produced more syllables in a 60-second stretch.

The duration of syllables – which reflects learners' speech planning and execution processes (Field, 2011; Levelt, 1989) – was also examined. Logically, if more syllables are produced in a constant stretch of speech (e.g. 60 seconds),

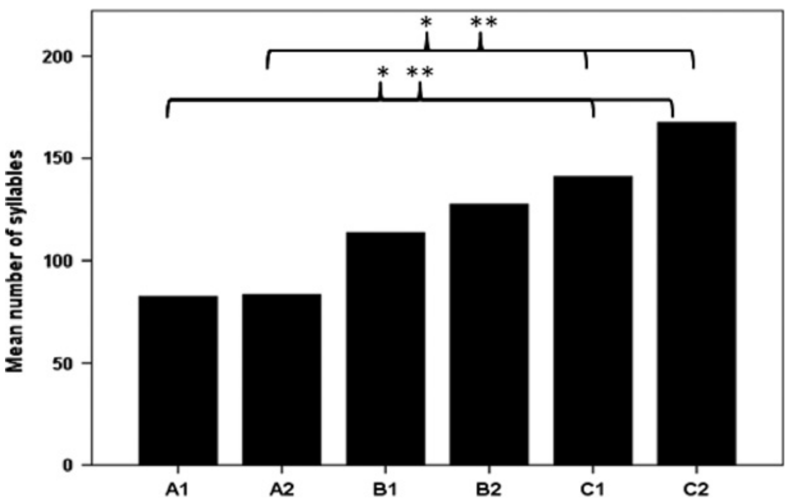

Figure 9.4 Speaking rate: Mean number of syllables (in 60-second sample)
Note: $*p < 0.05$; $** p < 0.01$.

then the syllables will become shorter. In less proficient learners, it is expected that the speech planning process will be slower as a result of lack of automatization of cognitive processes and limited lexico-grammatical resources (Levelt, 1989). The execution is further slowed down by the need for speech articulators to be moved into positions that could be unfamiliar for learners (e.g. ‘th’ or realizations of /r/ in different languages), which takes a longer time. As Levelt (1989: 413) noted, fluent articulation ‘involves the co-ordinated use of approximately 100 muscles’ – a tall order for a language learner.

Conforming to expectations, the analyses revealed that syllable durations in less proficient learners are significantly longer and decrease with increasing proficiency, as seen in the downward trend in Figure 9.5. An ANOVA of mean syllable duration showed a significant effect for Level, $F(5, 1436) = 37.099$, $p < 0.01$. Post hoc tests (Scheffe) confirmed the three-way grouping that is visible in the figure (Levels A1/A2, B1/B2, C1/C2).

The specific durations of different prosodic syllable types (i.e. unaccented, accented, unaccented final and accented final) were additionally examined as a potential discriminating measure across proficiency levels. An ANOVA with CEFR Level (six levels) and Prosodic position (four levels) as fixed factors showed significant main effects for both, $F(5, 1418) = 10.93$, $p < 0.001$ and $F(3, 1418) = 146.09$, $p < 0.001$, respectively, as well as a two-way interaction, $F(15, 1418) = 2.68$, $p < 0.001$. Post hoc tests (Scheffe) showed that accented and unaccented final syllables generally formed three

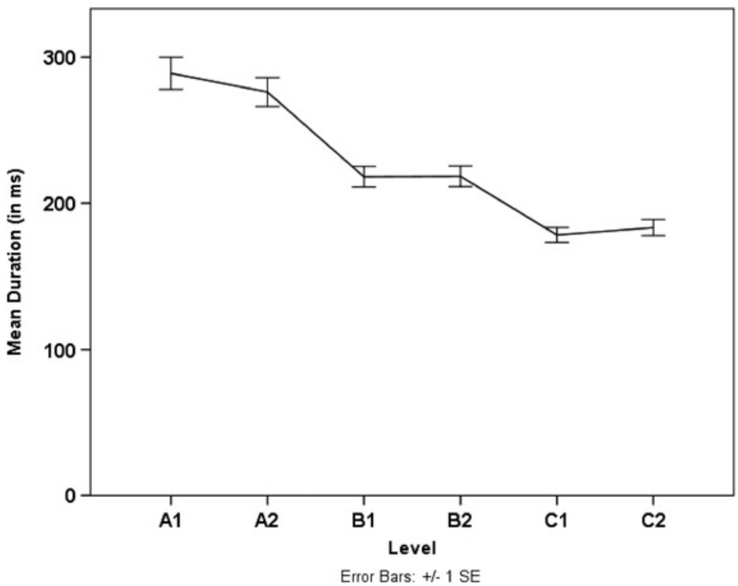

Figure 9.5 Speaking rate: Mean duration of syllables

homogeneous subsets: Levels A1/A2, B1/B2 and C1/C2, as the measurements in Figure 9.6 indicate.

Figure 9.6 indicates that, with the exception of unaccented final syllables, the duration of all other syllable types generally decreases with increasing proficiency, which is a reflection of developing speech planning and execution processes. However, importantly, the durations of the different syllable types do not change to the same extent. Instead, unaccented and accented durations differ more at advanced C levels than at beginner A levels. This suggests a progression towards the more English realization of the syllable types where consistent durational differences are present between the two syllable types. Additionally, the lengthening of unaccented final syllables, a very characteristic property of stress-timed languages such as English, becomes longer than unaccented syllables at the C levels but is indistinguishable from accented syllables at the A levels. Therefore, while the durations of syllable types clearly overlap at beginner levels and are thus indistinguishable, at C2 learners have implemented different 'categories' for all prosodic syllable types that are marked by distinct durational patterns. This shows a progression from a more syllable-timed realization of syllable types to a more stress-timed English realization.

It is also worth noting the error bars in Figure 9.6, which indicate variability within proficiency levels and overlap between scale bands – a finding

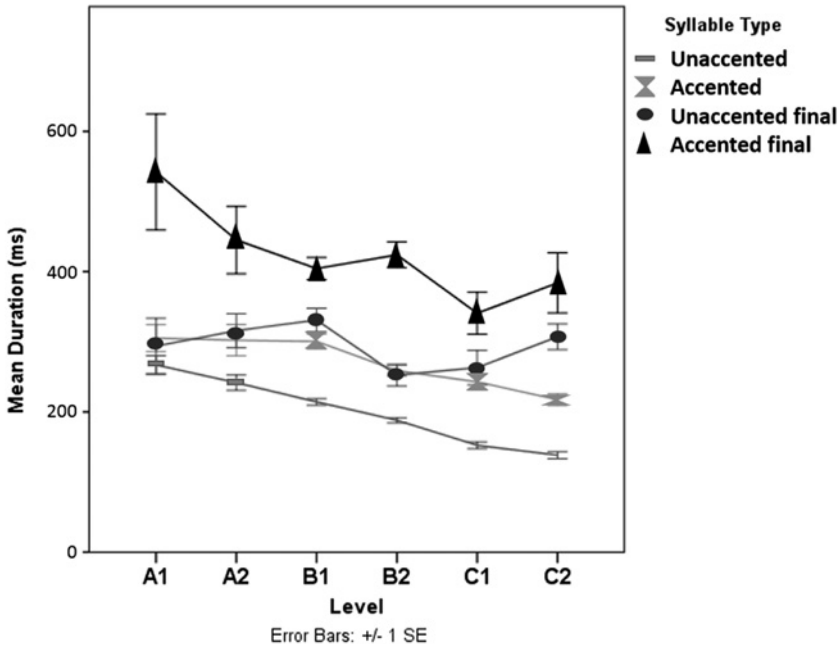

Figure 9.6 Mean prosodic lengthening across proficiency levels by syllable type

that echoes results reported in Kang (2013) and Iwashita *et al.* (2008). Since this study is an exploratory study with only two speakers per CEFR level, more research is needed to further validate the robustness of the effects between different proficiency levels. Nevertheless, we can conclude that the measures tested here are useful in discriminating between levels of proficiency, as demonstrated with the Spanish learners of English participating in the study.

Study 2: Differences across first languages

Figure 9.7 plots the proportion of vocalic material (%V) produced by the German, Spanish and Korean learners against the variability in their consonantal intervals (Varco-C); rPVI-C is omitted from the figure, since its pattern of results resembles that for Varco-C. The data in the figure show that the rhythm metrics differ for the three L1 learner groups, with the highest Varco-C values for German and the highest %V values for Spanish, as would be expected under L1 transfer – that is, a high number of consonants and low number of vowels in German as a stress-timed language and vice versa in Spanish as a syllable-timed language.

A MANOVA, which included all six rhythm metrics and the fixed factors Language (German, Spanish, Korean) and Level (B1, B2), revealed a significant main effect of Language for Varco-C, $F(2, 148) = 3.36$, $p < 0.05$, and rPVI-C, $F(2, 148) = 4.70$, $p = 0.01$, and a significant interaction between Language and Level for %V, $F(2, 148) = 3.45$, $p < 0.05$, and rPVI-C, $F(2, 148) = 3.28$, $p < 0.05$, but no main effect for Level. Post hoc tests (Scheffe)

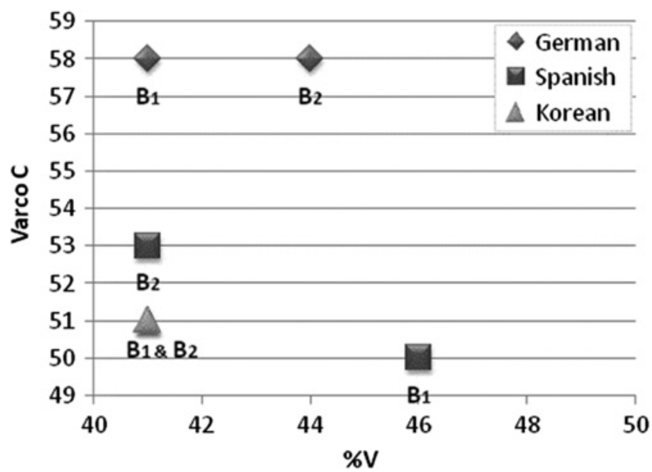

Figure 9.7 Mean metric values for %V and Varco-C for German, Spanish and Korean learners of English at Levels B1 and B2

showed that German and Korean learners differed marginally for Varco-C ($p = 0.071$), and German and Spanish learners did so for rPVI-C ($p < 0.05$).

Figures 9.8 and 9.9 show the effect of L1 background on accentual and phrase-final lengthening for Levels B1 and B2, respectively. Figure 9.8 reveals an L1 background effect, with German learners better at distinguishing between the prosodic lengthening of accented and unaccented final syllables than the other learners, at least at Level B1. At Level B2 (Figure 9.9) the picture changes, with German learners differentiating less between the different syllable types, while the Koreans are doing better, and the Spanish are

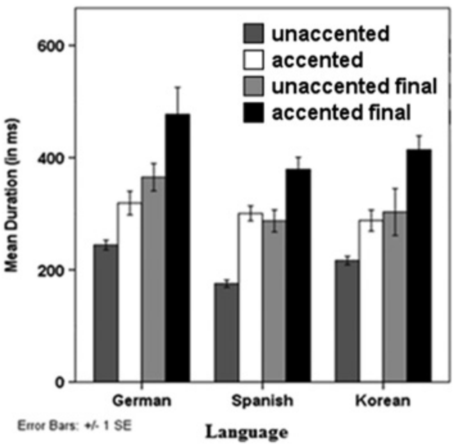

Figure 9.8 Lengthening across languages by syllable type (Level B1)

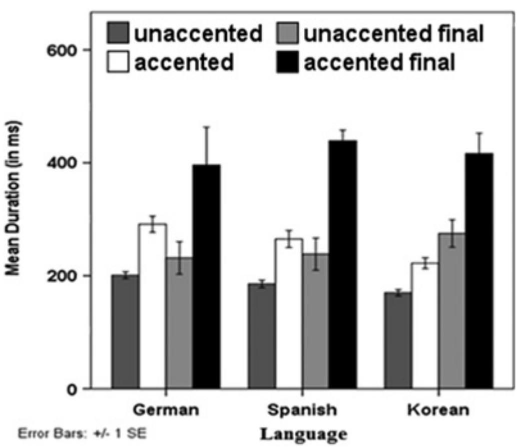

Figure 9.9 Lengthening across languages by syllable type (Level B2)

essentially continuing as before. Interestingly, although German learners with the most typologically similar language background to the target language showed an early advantage for some measures, they were often still as far off target as the other L1 groups at the intermediate B levels.

An ANOVA with fixed factors L1 Language (German, Spanish and Korean), Level (B1 and B2), and Lengthening condition (unaccented, accented, unaccented final, accented final) showed significant main effects for all three factors respectively, $F(2, 1683) = 3.02$, $p < 0.05$; $F(1, 1683) = 11.99$, $p = 0.001$; $F(3, 1683) = 116.84$, $p < 0.001$, and a significant interaction between Language and Level, $F(2, 1683) = 3.22$, $p < 0.05$, confirming the pattern of results sketched above. Therefore, we can conclude that the measures tested here can successfully distinguish between the spoken productions of learners from different L1 backgrounds.

Discussion

Overall, this investigation has shown that some of the prosodic measures under investigation here can provide useful micro-level prosodic measures for consideration in L2 teaching and assessment contexts. The results of Study 1, which focused on differences across proficiency levels, indicated that the learner speech observed at the different proficiency levels signalled progression from a more syllable-timed realization of speech to a more stress-timed realization, in line with the prosodic requirements of English. More specifically, the higher level learners in the sample used here were found to:

(1) have a higher speech rate, and produce more frequent and shorter syllables as reflected in the speech rate measurements, likely as a result of higher automaticity of speech planning and execution processes (Field, 2011);
(2) differ durationally between unaccented/accented and non-final/final syllables, indicating that at higher levels of proficiency learners are likely to have implemented different prosodic categories for different syllable types and display these with appropriate durational patterns;
(3) be more adept at producing appropriate durations for vowels and consonants as seen in the %V metric which showed a downward shift between Levels B1 and B2, and in the Varco-C and nPVI-C metrics which showed a more-or-less steady increase in consonantal variability from A2 to C2; this most likely reflects increased mastery of language-specific properties like vowel reduction and syllable structure complexity, and also accentual and final lengthening.

These findings are in line with Low's (2006) assertion that the stress-timing rhythm patterns of English speech would present problems for learners. They also support previous research which has shown an increase in syllables in learner speech as proficiency increases (Kang, 2013; Kang & Wang, 2014) and the difficulty of lower level learners with managing stress-timed speech (Iwashita *et al*., 2008). The current findings extend earlier research through pinpointing micro-level features of rhythm, which cause challenges for learners.

The results of Study 2, which focused on learners from three typologically different L1 backgrounds at intermediate-level proficiency, indicated that learners with different language backgrounds show different prosodic patterns:

(1) learners from a stress-timed language such as German showed the highest ability to deal with consonantal variation (as seen in Varco-C and rPVI-C), and learners from a syllable-timed language such as Spanish displayed the highest proportion of vocalic material (%V), suggesting that transfer effects that can be predicted on the basis of the rhythmic properties of their respective L1s still take effect at the B levels;
(2) German learners were better at distinguishing the durations between different syllable types than the Korean and Spanish learners at B1, but the Korean learners showed a clear progression at B2.

These findings suggest that there may be a prosodic basis to the results reported by Crowther *et al.* (2015), who found that speakers' L1 plays an important role in listener judgements of L2 comprehensibility.

A finding common to both studies was the variability within CEFR levels, with differences found mainly between Levels A1 and C2, and no meaningful differences between adjacent levels. This is in line with results reported in other studies, which have found that pronunciation features are not clearly distinct between adjacent levels (Isaacs *et al*., 2015; Iwashita *et al*., 2008; Kang, 2013; Kang & Wang, 2014).

Implications

Although the implications of these findings are only tentative due to the small-scale nature of the study, there are nevertheless useful insights for learning, teaching and assessment. In L2 pronunciation, global intelligibility and local (phonetic) precision are key concerns, and teachers and examiners must, therefore, have knowledge of the components of pronunciation and an awareness of when, how and why a pronunciation feature is problematic for learners. Levis (2006: 247) rightly argues that 'the first thing that teachers must learn is to give more than global impressions of pronunciation. They need to become aware of relevant phonological categories and be able to name important errors.' What this study has revealed are the rhythmic challenges that learners might face, such as distinguishing between different syllable types in terms of duration, the difficulty of reducing the amount of vocalic material in their speech, and the challenge of varying consonantal intervals. Important prosodic features such as syllable variability and the crucial role of vowel reduction could, as such, become a focus in the classroom, as also suggested by Liang (2003) and Low (2006). Such awareness would allow teachers to include prosodic features in their instruction, and would assist them to focus on the form-meaning relationships which are fundamental to learning (Isaacs, 2009). Understanding what learners can and cannot do at a specific proficiency level could also support learners to notice key aspects of their speech that would aid learning through conscious noticing and awareness (Schmidt, 1990). Clarity of pronunciation features that matter and affect their speech can additionally help learners to become more autonomous. The findings also indicate that different pronunciation features could be fruitfully emphasized in teaching and learning based on the L1 background of learners.

Investigating measures which 'count' in pronunciation also has implications for the assessment of pronunciation. As noted at the beginning of this chapter, raters have reported lower confidence in assessing pronunciation as compared to other skills (e.g. Brown & Taylor, 2006) and are more comfortable making global judgements of intelligibility as opposed to nuanced judgements about features of pronunciation. These findings may reflect the 'undemanding nature of judgements [of global intelligibility] in terms of technical expertise' (Yates *et al.*, 2011: 36), as opposed to the precise technical knowledge needed to deconstruct pronunciation into its constituent parts, and they signal the need for a comprehensive and in-depth training of examiners. The role of L1 background in learner speech has also confirmed the importance of supporting examiners in developing an 'international ear' through exposure to test takers from different L1s and levels of intelligibility, since a German and a Spanish learner, for example, would present different profiles that examiners would need to evaluate – as underlined by our findings. Raters' increased familiarity with L2 speech by learners from different L1 backgrounds and ability levels would, in addition, minimize the effect that rater familiarity with a test taker's pronunciation plays in assessment – empirically shown to play a role (e.g. Ballard & Winke, this volume; Carey *et al.*, 2011; Saito *et al.*, this volume) – and would ultimately result in more reliable and valid assessment. The role of L1 background in learner speech and its possible effect on rater comprehension of that speech also indirectly lends support to Wagner and Toth's (this volume) argument that pronunciation needs to be considered as part of both the speaking and listening constructs.

The findings have further implications for assessment in terms of scale development, scale descriptors and number of scale bands. Regarding the number of bands, the findings reported here indicate that with some of the measures there were three distinct homogenous subsets in the data – A1/A2, B1/B2, C1/C2, suggesting that at least in the small-scale study here, only three broad levels were observed. This finding is, of course, based on a small subset of measures for pronunciation, but it provides some confirmation of research that has reported difficulty in scaling pronunciation ability across six or more levels. For example, in an older version of the IELTS test, pronunciation was marked on a four-point scale, in contrast to the nine-point scale used for other traits such as lexical and grammatical resource, coherence and fluency, largely as a result of the different Many-Facet Rasch analysis findings for pronunciation (Taylor, personal communication). Research on the current nine-point IELTS scale – even though it was reported by examiners to be easy to use – has, nevertheless, indicated that listener-rated measures were difficult to discriminate between nine levels (Isaacs *et al.*, 2015). This finding also echoes the difficulties reported by North and Hughes (2003) in developing a six-level pronunciation scale, and indicates that pronunciation ability may be more meaningfully measured with fewer scale levels, or at least that further empirical work is needed focusing on the scaling of pronunciation.

The micro-level features identified here could also impact on the development of descriptors in assessment scales. They indicate that suprasegmental features, and specifically rhythm, play a distinguishing role at both higher and lower levels of L2 proficiency, and should therefore be included at all levels (as also argued by Harding, this volume). Not all features identified as important in this study can be captured in assessment scales, since scales are driven by a need for conciseness, as well as a need for usability and the avoidance of vague terminology, as Harding contends in an earlier chapter in this volume. This presents a case for more explicit references in scales to components of rhythm. Examples of this approach can be seen in the IELTS scale, which includes descriptors such as 'can sustain appropriate rhythm' and 'rhythm may be affected by lack of stress-timing' (http://www.ielts.org). Such specific references to the components of prosody in assessment scales are positive examples of how pronunciation can be deconstructed in scales, and how it could potentially make examiners more reliable through training and awareness raising of important pronunciation features. For example, in a survey of examiner views on the IELTS speaking scales, 83% and 76% of examiners reported that 'rhythm' and 'stress timing', respectively, were salient pronunciation features for them when they assess (Galaczi *et al.*, 2012).

Even if not included in assessment scales, micro-level features can provide useful guidance for examiner training, since they provide explicit information about what matters in pronunciation across CEFR levels, and can be beneficial in developing a shared understanding of pronunciation and its constituent parts. As Yates *et al.* (2011: 36) argue, identification of important micro-level features could offer examiners 'a discourse that they can use to articulate what they have noticed' and 'a framework within which to talk about the same aspects of a performance'.

The findings reported here could also have implications for automated speech recognition (ASR) and assessment systems of speech in which pronunciation measurement plays a central role. Most current systems are based on detailed taxonomies of pronunciation features (e.g. articulation and duration of phonemes, pauses, use of pitch, and mean duration between stressed syllables) and associated weightings of those features in computer algorithms (e.g. Evanini & Wang, 2013; van Moere, 2012; Xi *et al.*, 2012). Such systems could potentially include some of the rhythm measures identified here as a means of improving speech recognition accuracy and enhancing the assessment of prosody. The findings on the effect of L1 background on rhythm can also be useful for ASR, as they support the development of ASR systems targeted at specific L1s or language groups.

Future Research and Conclusion

This exploratory study is constrained by its small-scale nature and limited generalizability. A larger scale investigation would address this limitation and be useful in further exploring the pronunciation measures that were found to play a role. Such an investigation could extend not just to a larger sample of learners and L1s, but also to task types, since research has revealed interesting findings about differences in pronunciation measures across monologic and interactional task types (Kang & Wang, 2014). A mixed-methods integration of quantitative and qualitative findings would provide further useful insights, and could, for example, extend the present study to an investigation of the degree to which raters attend to the instrumentally derived measures identified here. The need for longitudinal work also needs to be borne in mind, since it could complement the cross-sectional snapshots provided in the current study. The use of data from two different assessment modes (computer and face-to-face) – a potential limitation of the study – is a further area to explore. Even though from an assessment perspective there are clear differences between the two test modes, this is not considered to be a major threat to this study, since in both modes learners provided extemporaneous speech which illustrated their mastery of a range of phonological features. Nevertheless, an exploration of the effect of the face-to-face versus computer-based mode on pronunciation could reveal potential pronunciation differences and implications for assessment. Notwithstanding these limitations, a systematic empirical investigation of the research questions guiding this study has contributed to an empirically based understanding of the pronunciation construct, which can inform scale development and rater and teacher competence, and contribute to the development and assessment of learner pronunciation.

References

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

ACTFL (American Council on the Teaching of Foreign Languages) (2012) *ACTFL Proficiency Guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Anderson-Hsieh, J., Johnson, R. and Koehler, K. (1992) The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. *Language Learning* 42, 529–555.

Bachman, L. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Boersma, P. and Weenink, D. (2010) *Praat: Doing Phonetics by Computer*, Version 5.2 [Computer program].

Bolinger, D.L. (1965) Pitch accent and sentence rhythm. In I. Abe and T. Kanekiyo (eds) *Forms of English: Accent, Morpheme, Order* (pp. 139–180). Cambridge, MA: Harvard University Press.

Brown, A. and Taylor, L. (2006) A worldwide survey of examiners' views and experience of the revised IELTS speaking test. *Research Notes* 26, 14–18.

Carey, M.D., Mannell, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28, 201–219.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Crowther, D., Trofimovich, P., Saito, K. and Isaacs, T. (2015) Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly* 49, 814–837.

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

Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for ΔC. In P. Karnowski and I. Szigeti (eds) *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium, Piliscsaba 2003* (pp. 231–241). Frankfurt am Main: Peter Lang.

Evanini, K. and Wang, X. (2013) Automated speech scoring for non-native middle school students with multiple task types. *Proceedings of INTERSPEECH, Lyon, France, 2013*. Baixas: ISCA.

Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39, 399–424.

Field, J. (2011) Cognitive validity. In L. Taylor (ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Studies in Language Testing (Vol. 30, pp. 65–111). Cambridge: UCLES/Cambridge University Press.

Galaczi, E.D., Lim, G. and Khabbazbashi, N. (2012) Descriptor salience and clarity in rating scale development and evaluation. Paper presented at the Language Testing Forum, Bristol.

Grabe, E. and Low, E.L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven and N. Warner (eds) *Laboratory Phonology 7* (pp. 515–546). Berlin: Mouton de Gruyter.

Guilbault, C.P.G. (2002) The acquisition of French rhythm by English second language learners. Doctoral dissertation, University of Alberta.

Gussenhoven, C. and Jacobs, H. (2005) *Understanding Phonology* (2nd edn). London: Hodder Education.

Gut, U. (2009) *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt am Main/Oxford: Peter Lang.

Gutiérrez-Díaz, F. (2001) The acquisition of English syllable timing by native Spanish learners of English. An empirical study. *International Journal of English Studies* 1 (1), 93–113.
Hahn, L.D. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38, 201–223.
Isaacs, T. (2009) Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal* 27, 1–12.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment*. Hoboken, NJ: John Wiley.
Isaacs, T. and Trofimovich, P. (2012) 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.
Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.
Iwashita, N., Brown, A., Mcnamara, T. and O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29 (1), 24–49.
Jenkins, J. (2002) A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics* 23, 83–103.
Kang, O. (2010) Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System* 38 (2), 301–315.
Kang, O. (2013) Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Research Notes* 52, 40–48.
Kang, O. and Wang, L. (2014) Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes* 57, 40–49.
Kang, O., Rubin, D.L. and Pickering, L. (2010) Suprasegmental measures of accentedness and judgements of English language learner proficiency in oral English. *Modern Language Journal* 94, 554–566.
Kohler, K. (1982) Rhythmus im Deutschen [Rhythm in German]. *Arbeitsberichte, Institut für Phonetik der Universität Kiel* 19, 89–106.
Levelt, W.J.M. (1989) *Speaking*. Cambridge, MA: MIT Press.
Levis, J.M. (2006) Pronunciation and the assessment of spoken language. In R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics* (pp. 245–270). New York: Palgrave Macmillan.
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 (2), 223–255.
Liang, W.X. (2003) Teaching weak forms. *Forum* 41, 32–36.
Linacre, J. (1989) *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.
Low, E.L. (1998) Prosodic prominence in Singapore English. Unpublished doctoral dissertation, University of Cambridge.
Low, E.L. (2006) A review of recent research on speech rhythm: Some insights for language acquisition, language disorders and language learning. In R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics* (pp. 99–125). New York: Palgrave Macmillan.
Low, E.L., Grabe, E. and Nolan, F. (2000) Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43 (4), 377–401.
Major, R. (2008) Transfer in second language phonology: A review. In J.G.H. Edwards and M.L. Zampini (eds) *Phonology and Second Language Acquisition* (pp. 63–94). Amsterdam: John Benjamins.
Munro, M.J. and Derwing, T.M. (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45, 73–97.

North, B. and Hughes, G. (2003) CEF illustrative performance samples. See www.eaquals.org/wp-content/uploads/Documentation-to-English-2004-DVD-with-adults_speaking.pdf.

Peterson, G.E. and Lehiste, I. (1960) Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32 (6), 693–703.

Pickering, L. (2001) The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly* 35 (2), 233–255.

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

Post, B., Payne, E., Prieto, P., Astruc, L. and Vanrell, M.D. (2010) A multisystemic model of rhythm development: Phonological and prosodic factors. BAAP Colloquium, London, 29–31 March.

Prieto, P., Vanrell, M.d.M., Astruc, L., Payne, E. and Post, B. (2012) Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English, and Spanish. *Speech Communication* 54, 681–702.

Pulgram, E. (1970) *Syllable, Word, Nexus, Cursus*. The Hague: Mouton.

Ramus, F., Nespor, M. and Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292.

Ramus, F., Dupoux, E. and Mehler, J. (2003) The psychological reality of rhythm classes: Perceptual studies. In M. Solé, D. Recasens and J. Romero (eds) *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 337–342). Barcelona: Universitat Autonomá de Barcelona.

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

Schmidt, R. (1990) The role of consciousness in second language learning. *Applied Linguistics* 11, 129–150.

Seong, C. (1995) The experimental phonetic study of standard current Korean speech rhythm: With respect to its temporal structure. PhD thesis, Seoul National University.

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.

Trofimovich, P. and Isaacs, T. (2012) Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15 (4), 905–916.

Van Moere, A. (2012) A psycholinguistic approach to oral language assessment. *Language Testing* 29 (3), 325–344.

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

Xi, X., Higgins, D., Zechner, K. and Williamson, D. (2012) A comparison of two scoring models for an automated speech scoring system. *Language Testing* 29 (3), 371–394.

Yates, L., Zielinski, B. and Pryor, E. (2011) The assessment of pronunciation and the new IELTS Pronunciation Scale. In J. Osborne (ed.) *IELTS Research Reports* 12 (pp. 1–46). Melbourne and Manchester: IDP IELTS Australia and British Council.