RHYTHM METRICS PREDICT RHYTHMIC DISCRIMINATION

Laurence White, Sven L. Mattys, Lucy Series and Suzi Gage

Department of Experiment Psychology, University of Bristol
laurence.white@bristol.ac.uk

ABSTRACT

Metrics such as VarcoV (standard deviation of vocalic interval duration divided by the mean) and %V (proportion of total utterance duration comprised of vocalic intervals) provide empirical support for long-held notions about rhythmic distinctions between languages. Furthermore, listeners can discriminate languages with distinct rhythm metric scores purely on the basis of the durational information available in resynthesized monotone *sasasa* speech. However, some factors contributing to this durational variation, such as stress distribution and prosodic timing, are not directly reflected in rhythm scores. To test more precisely the predictive power of rhythm metrics, we used tightly controlled *sasasa* stimuli, eliminating stress distribution and prosodic timing cues to focus on the information directly quantified by rhythm metrics. We show that VarcoV and %V scores are predictive of listeners' discrimination within and between languages, even with these highly constrained stimuli.

Keywords: Rhythm, stress, rhythm metrics, discrimination.

1. INTRODUCTION

The perception of speech rhythm derives from the repetition of syllables, or stressed syllables in particular. Cross-linguistic differences in rhythm arise in part from the relative strength of stressed and unstressed syllables [2]. In English, for example, stressed syllables are more likely to have complex onsets and codas, and contain full vowels which are much longer than unstressed reduced vowels. Some stressed syllables are pitch-accented, which further increases vowel and consonant duration [14]. Spanish has much less durational difference between stressed and unstressed vowels, less centralization of unstressed vowels and little differentiation between stressed and unstressed syllables in terms of onset and coda complexity [2]. In addition, pitch accent is not consistently marked by lengthening in Castilian Spanish [8].

These patterns of variation in the duration of stressed and unstressed vowels and consonants have been exploited in rhythm metrics designed to quantify rhythmic differences between languages [6, 12]. Some metrics, such as ΔV and ΔC [12], are highly sensitive to speech rate [1], which makes comparison difficult in all but the most carefully controlled conditions. Recent studies have suggested that two rate-insensitive metrics provide effective discrimination between languages that have been held to differ rhythmically: VarcoV [3] and %V [12]. Together, VarcoV and %V provide a two-dimensional rhythm space which has been shown to discriminate English and Dutch from Spanish and French, and to distinguish first and second language English and Spanish [15].

Such results provide quantitative empirical support for the traditional distinction between stress-timed languages (e.g. English) and syllable-timed languages (e.g. Spanish) [10]. Patterns of rhythm scores indicate, however, that this distinction is far from categorical. Indeed, rhythmic differences are also found between accents of a given language. Accents with pitch-peak delay, such as Welsh Valleys (WV) and Orkney Islands (OR) English, which show levelling of the durational contrast between stressed and post-stress syllables, have been held to be more syllable-timed [7]. Rhythm scores confirm this subjective impression: Figure 1 shows VarcoV and %V scores for WV and OR intermediate between those of standard southern British English (SE) and those of Castilian Spanish (CS) [16].

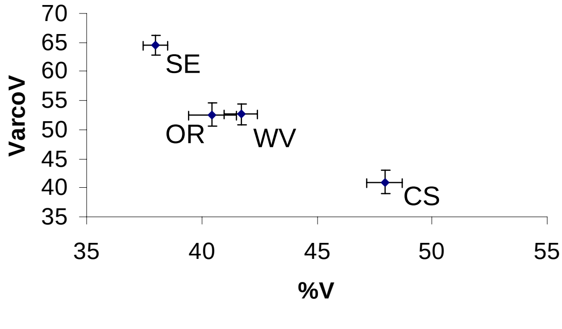

Figure 1: VarcoV and %V scores for three accents of English, plus Spanish, see text for key (from [16]).

Distinctions indicated by rhythm metrics have been shown to have perceptual correlates. Ramus *et al.* [13] used MBROLA resynthesis [4] to transform utterances into monotone sequences of *sasasa* syllables, while maintaining the duration of the original vowels and consonants. They demonstrated that French native speakers could perceive the difference between such utterances derived from languages with distinct ΔV and %V rhythm scores, such as English and Spanish or Polish and Catalan, while not discriminating languages with similar rhythm scores, such as English and Dutch or Catalan and Spanish.

Durational variation in vocalic and consonantal intervals is not, however, the only factor giving rise to the perception of cross-linguistic differences in rhythm. The distribution of stressed and unstressed syllables is also rhythmically important: Spanish permits long sequences of unstressed syllables, whereas the existence of secondary lexical stress in English, together with processes such as stress insertion and stress retraction, mean that shorter sequences of unstressed syllables (one, two or three) are the norm [2]. Thus, there are aspects of rhythmic variation between languages that are not directly quantified by rhythm metrics such as VarcoV and %V. Although Pairwise Variability Indices [6] were proposed to capture the sequential nature of rhythmic differences, in practice PVI scores for vowels are highly correlated with simple measures of variation in vocalic duration such as VarcoV [15].

Discrimination in Ramus et al.'s experiment may therefore have been facilitated by cross-linguistic differences in stress distribution, which have no clear correlates in rhythm scores. In addition, prosodic timing effects – in particular, utterance-final lengthening – are preserved in full *sasasa* utterances. This is important because languages differ in their durational marking of prosodic structure: for example, English has substantial utterance-final lengthening [17] but this appears less widespread in Spanish [5]. Patterns of utterance-edge timing effects are not, however, directly interpretable in terms of rhythm scores.

The experiment reported here modifies the methodology of Ramus et al. in critical ways. Firstly, we looked at contrasts within as well as between languages, eliminating the influence of stress distribution on discrimination judgements. Although rhythm metrics show Welsh Valleys English to have less differentiation between stressed and unstressed syllables than standard southern British English (Fig. 1), the *distribution* of stressed syllables is the same. Thus, discrimination between WV and SE on the basis of monotone *sasasa* speech could only arise from the durational balance of vocalic and intervocalic intervals rather than from differences in the sequencing of strong and weak syllables not directly captured by rhythm metrics. Secondly, we utilised *sasasa* utterances stripped of their initial and final stretches. Stimuli thus lacking cues to stress distribution and utterance-level prosodic timing allow a clearer test of the power of rhythm scores to predict rhythmic discrimination between and within languages.

2. METHOD

2.1. Materials

We used a subset of the utterances recorded for the study illustrated in Figure 1 [16]. There were four speakers each for three English accents (Welsh Valleys, standard southern British, Orkney Islands) and four speakers of Castilian Spanish. Each speaker read five sentences, which, to facilitate measurement of segment duration, were free of approximants. For each recorded utterance, the durations of vocalic and intervocalic intervals were measured by visual inspection of waveforms and spectrograms (see [15] for full details).

We used MBROLA resynthesis to convert each string of vocalic and intervocalic interval durations into a sequence of [sa] syllables of the same duration. The fundamental frequency of the synthesized utterance was 230 Hz throughout. To eliminate utterance-edge durational effects and to prevent cueing of language identity from utterance length, we truncated the synthesised utterance before the final stressed syllable and removed the initial syllable and sufficient additional syllables to leave just ten syllables in each utterance. Finally, to eliminate any effect of speech rate on listeners' perceptions of the different accents/languages (henceforth, for brevity, "languages"), we stretched or compressed each utterance uniformly to a constant 1900 ms. Thus the relative durations of vowels and consonants were as in the original utterances, but the listener could not use segmental information, pitch, initial/final durational effects, length or speech rate to distinguish utterances.

The 20 resynthesized *sasasa* utterances per language (four speakers and five sentences) were used in four pairwise discrimination experiments:

1. Castilian Spanish (CS) *vs* standard southern British English (SE).
2. Castilian Spanish (CS) *vs* Welsh Valleys English (WV).
3. Welsh Valleys English (WV) *vs* standard southern British English (SE).
4. Welsh Valleys English (WV) *vs* Orkney Islands English (OR).

From the vowel and consonant durations of the *sasasa* utterances, we derived VarcoV and %V scores for each language group, shown in Figure 2.

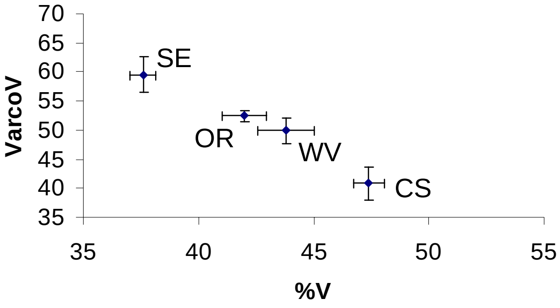

Figure 2: Mean VarcoV and %V scores for the resynthesised *sasasa* utterances.

Two-tailed t-tests showed that the pattern of differences in rhythm scores for the *sasasa* utterances was equivalent to that of the larger set of intact utterances from which they were derived (Fig. 1). For VarcoV scores: SE was significantly higher than WV [p < .05] and CS [p < .001]; WV was significantly higher than CS [p < .05]; WV and OR were not significantly different. The pattern of contrasts was parallel for %V scores: SE was significantly lower then WV [p < .001] and IS [p < .001]; WV was significantly lower than CS [p < .05]; WV and OI were not significantly different.

2.2. Procedure

Following the protocol of Ramus *et al.* [13], we told participants that they would hear modified speech from four languages: Sahatu (SE), Moltec (WV), Eboda (CS) and Ventish (OR). We used an AAX paradigm for presentation of utterances, participants having to respond by pressing a key to indicate that the language of the test utterance – X – was either the same as or different to the example language – A. Within each AAX trials, the three utterances were always from different speakers and different original sentences. The order of presentation of example languages was counterbalanced between participants, as was the order of the four pairwise discrimination experiments. Each experiment comprised 40 trials, 20 with each language as the exemplar.

2.3. Participants

We tested 24 native English speakers, undergraduates or paid volunteers, with no speech or hearing problems. All participants took part in all four discrimination experiments in one session.

3. RESULTS

In accordance with recent guidelines [9], we used the signal detection theory measure *d'* to assess participants' sensitivity to differences between pairs of languages. This is derived from the hit rate, H – proportion of correct "same" trials – and the false alarm rate, FA – proportion of incorrect different trials. We calculated *d'* thus:

(1) *d'* = normsinv (H) – normsinv (FA).

To assess participants' ability to discriminate the test languages, we compared their d' scores for each experiment to 0, the chance level, using two-tailed t-tests. Table 1 shows percentage correct, *d'* scores and significance level for each experiment.

	% correct	*d'*	*p*
1. CS *vs* SE	53.4	0.20	< .05
2. CS *vs* WV	54.5	0.25	< .05
3. WV vs SE	53.8	0.23	< .05
4. WV *vs* OR	52.8	0.13	> .10

Table 1: Percentage correct, *d'* scores and significance level for pairwise discriminations.

As the percentage correct scores indicate, this was a very difficult task for listeners. In all cases, however, participants reliably discriminated pairs of languages with significantly different VarcoV and %V scores: (1) Castilian Spanish *vs* standard southern British English; (2) Castilian Spanish *vs* Welsh Valleys English; and (3) Welsh Valleys English *vs* standard southern British English. Experiments 1 and 2 show that discrimination is possible in the absence of utterance-initial or utterance-edge durational information. Experiment 3 shows, in addition, that distributional information is not required for discrimination: stress distribution is the same in the utterances from the different varieties of English, and only information about variation in the durations of stressed and unstressed vowels and consonants, direct measured by VarcoV and %V, was available for listeners.

In Experiment 4 – Welsh Valleys English *vs* Orkney Islands English – VarcoV and %V scores are not significantly different and accordingly participants are unable to discriminate these accents. Thus, the information directly reflected in rhythm scores is not only sufficient, but also necessary, for discrimination within languages.

4. DISCUSSION

Work on rhythm metrics [15] has shown VarcoV and %V to be the most successful at distinguishing so-called stress-timed and syllable-timed languages: thus, French and Spanish have higher %V and lower VarcoV than English and Dutch. In the same study, these metrics also produced scores for second language speakers that were distinct from those of both the target language and of the speakers' native language. Finally, VarcoV and %V have shown distinctions between accents of English in line with what has been previously asserted about the greater syllable-timing of accents such as Welsh Valleys English [16].

Rhythm is fundamentally a perceptual phenomenon, however, and perceptual validation of rhythm metrics is necessary. Ramus et al. started this process, using monotone *sasasa* speech to focus on durational cues, thereby showing discrimination by native French listeners that followed the pattern predicted by rhythm metrics [13]. Information about stress distribution and prosodic timing effects – not directly reflected in rhythm scores – were, however, also present in their materials, and could account, at least in part, for their discrimination results

Here we have shown that native English listeners can discriminate languages with different VarcoV and %V scores even in the absence of utterance-level prosodic timing cues. In addition, listeners are able to discriminate varieties of English – with the same stress distribution – simply on the basis of durational variation between stressed and unstressed vowels and consonants.

The fact that metrics such as VarcoV and %V are not directly informative about stress distribution is a limitation. A full quantification of linguistic rhythm will clearly also require some statistical metric of the arrangement of stressed and unstressed syllables. It is worth noting, however, that even in this very difficult task, with minimal cues available, listeners are capable of attending to and interpreting precisely the information – variation in vowel and consonant duration – that is encapsulated by rhythm metrics VarcoV and %V.

This version of the *sasasa* experiment thus represents a significant extension of the findings of Ramus *et al.* [13]. These results provide strong support for the use of rhythm metrics in classifying perceptually salient aspects of speech rhythm.

5. REFERENCES

[1] Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? *Proc. 15th ICPhS Barcelona*, 2693-2696.

[2] Dauer, R.M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51-62.

[3] Dellwo, V. 2006. Rhythm and speech rate: A variation coefficient for ΔC. *Language and Language Processing: Proc. 38th Linguistic Colloquium Piliscsaba*, 231–241.

[4] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vreken, O. 1996. The MBROLA project… *Proc. ICSLP '96 Philadelphia*, 1393-1396.

[5] Frota, S., D'Imperio, M., Elordieta, G., Prieto, P., Vigário, M. 2007. The phonetics and phonology of intonational phrasing in Romance. In Prieto et al. 2007.

[6] Low, E.L., Grabe, E., Nolan, F. 2000. Quantitative characterisations of speech rhythm: 'syllable-timing' in Singapore English. *Language and Speech* 43, 377-401.

[7] Mees, I.M., Collins, B. 1999. Cardiff: a real-time study of glottalization. In Foulkes, P., Docherty, G. (eds.), *Urban Voices…*, 185-202. London: Arnold.

[8] Ortega-Llebari, M., Prieto, P. 2007. Disentangling stress from accent in Spanish: production patterns of the stress contrast in deaccented syllables. In Prieto et al. 2007.

[9] Pastore, R.E., Crawley, E.J., Berens, M.S., Skelly, M.A. 2003. "Nonparametric" *A'* and other modern misconceptions about signal detection theory. *Psychonomic Bulletin and Review* 10, 556-569.

[10] Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[11] Prieto, P., Mascaró, J., Solé, M.-J. (eds.). 2007. *Segmental and Prosodic issues in Romance Phonology*. Amsterdam: John Benjamins.

[12] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265-292.

[13] Ramus, F., Dupoux, E., Mehler, J. 2003. The psychological reality of rhythm classes: Perceptual studies. *Proc. 15th ICPhS Barcelona*, 337-342.

[14] Turk, A.E., White, L. 1999. Structural influences on accentual lengthening in English. *Journal of Phonetics* 27, 171-206.

[15] White, L., Mattys, S.L. In press. Calibrating rhythm: First and second language studies. *Journal of Phonetics.*

[16] White, L., Mattys, S.L. 2007. Rhythmic typology and variation in first and second languages. In Prieto et al. 2007.

[17] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.J. 1992 Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91, 1707-1717.