Rhythm, Metrics, and the Link to Phonology

Jason Brown
Dept. of Applied Language Studies & Linguistics
University of Auckland
jason.brown@auckland.ac.nz

Sam Mandal
University of Western Sydney
and
MARCS Auditory Laboratories
s.mandal@uws.edu.au

Abstract

Since Ramus et al. (1999) a number of statistical *metrics* have been routinely employed by researchers (Ramus 2003, Grabe & Low 2002 etc.) in an effort to rhythmically classify languages. However, recent studies by Arvaniti (2009), Tilsen & Arvaniti (2013), Arvaniti & Rodriquez (2013) etc., have challenged both the validity of these metrics in reflecting speech rhythm, and the physical measurability of rhythm itself. The present study takes a comparative evaluative approach, and explores the applicability of the proposed metrics to a Papuan language (Urama) with a phonology quite different than traditional Western European (W.E.) languages. It is argued here that the statistical underpinning of the existing rhythm metrics is a direct outcome of an overt effort to capture the temporal durational characteristics of the phonotactics of W.E. languages. As such, these metrics are only capable of providing a crude measure of *timing*.

1 Introduction

Approaches to rhythm in language have traditionally viewed languages as falling into either strict or less precisely-defined categories that were based primarily on the notion of *timing* (Abercrombie 1967, Port et al. 1987). *Timing,* as used in these contexts, is more or less a blanket-term referring to all aspects of durational variation in speech. Originally linked to the notion of *isochrony* (Abercrombie 1967), the term has since been increasingly used to refer to the relative temporal durational variability of consonantal and vocalic intervals in running speech (Ramus et al. 1999). The cornerstone of subsequent studies (Grabe & Low 2002, Ramus et al. 2003, Dellwo 2006 etc.) have been a consistent focus on the relative variability of consonantal and vocalic durations across languages, with an effort to establish generalizations in patterns of durational variability to help classify languages into different *rhythm classes*. Metrics were developed as tools of statistical measurement of said durational variability in speech. The efficacy of such efforts, and the phonological basis of the rationale offered for their methodological choices, have been vigorously disputed in recent works. Arvaniti (2009, 2012), Tilsen & Arvaniti (2013), Arvaniti & Rodriquez (2013) present empirical evidence to illustrate that the aforementioned statistical metrics can neither classify non-prototypical languages, nor provide correlates of perceptual discrimination. These authors offer perceptual experimental data to prove that not only is *timing* affected by a multitude of factors (speaking rate, voice quality, stimuli type etc.), but perceptual discrimination is often achieved through attunement to duration-independent acoustic factors such as fundamental frequency (f_0).

This being the case, the present paper aims to investigate the phonological basis for these acoustic metrics, the interrelation between the mathematical formulae they employ and the acoustic correlates of rhythm they are supposed to measure. Our primary hypothesis is that these metrics are simply different statistical measures of how consonant or vowel-heavy a language (or a token) is, and while there might possibly be some correlation between perceptual discrimination abilities of listeners and metric scores, the metrics are rarely complete indicators of the causation of such perceptual abilities. We further argue that 'rhythm' is more of a psychological reality than an acoustic factor, an abstract realization that lacks a single physical/acoustic correlate. It is, rather, the perceptual effect produced in the mind by the internal interactions of the different phonological abstractions that constitute a language. The components that make up the phonology and interact with each other to induce the perceptual effect in the mind of the listener

Jason Brown and Sam Mandal. 2013. Rhythm, Metrics, and the Link to Phonology. In *Proceedings of Australasian Language Technology Association Workshop*, pages 112−117.

that is *rhythm in speech,* and being a psychological reality rather than an acoustic entity *rhythm* is likely to elude any physical/acoustic probing. Thus, there is not much of a basis for *rhythmic classes* in these metrics (cf. Arvaniti 2009); however, they do reflect the gross phonological properties of a given language. To the extent that these phonological properties are specific properties of individual rhythm classes remains rather unsubstantiated in terms of empirical evidence.

This paper presents an instrumental study of a unique (and under-documented) language, and an elaboration of both whether the traditional methodologies and metrics that have yielded dubious results even for most Western European (W.E.) languages can capture the dynamics of a phonotactically 'strict' language, and what methodological changes may be required in order to accommodate under-studied phonological types.

2 Rhythm Metrics

The search for the proper acoustic metrics to capture the durational variability patterns thought to be indicators of rhythmic typology, Ramus et al. (1999) claim, was based on the observations regarding certain phonotactic regularities in syllable structure within Romance and Germanic languages, as elucidated in Dauer (1983). However, Arvaniti (2009) finds that these metrics are only partially based on the eight parametric criteria elaborated by Dauer (1983), and further that Dauer's (1983) own study contradicts the predictions one would make based on her criteria for languages such as Greek and Spanish (Dauer 1983:58). As Arvaniti points out, the main source of complication is two-fold; (a) Dauer's (1983) criteria have not been rigorously tested with a wide enough cross-linguistic focus, and (b) while Dauer's criteria combine factors that directly reflect phonetic timing as well as ones with no direct link to timing (e.g. function of f_0 in language), the design philosophy employed for the metrics only takes into account *those specific criteria that relate directly to timing* while excluding others. This is inherently problematic given that duration of segments, the main target of these statistical metrics, is affected by a multitude of factors like consonant gemination, phrase-final lengthening, syllable-position, focus-oriented lengthening, etc., all of which fail to be accounted for in the these metrics. As such, it becomes logically evident that these metrical measurements are very loosely based on a small subset of Dauer's (1983) criteria and can, at best, provide a very crude measurement of durational variability in speech.

Despite such obvious shortcomings, Ramus et al. (1999), for example, claims that a combination of %V and ΔC provide the best correlates for acoustic rhythm. Their study was limited to mostly W.E. languages, and Grabe and Low (2002) rightly point out that using different metrics on a large sub-set of languages yield confusing results with the effect of classifying the same language into different rhythmic types. For example, a PVI-based measure classifies Thai as stress-timed and Luxembourgish as syllable-timed, while a combination of %V and ΔC classify the same languages as being syllable-timed and stress-timed, respectively. Similarly, White and Mattys (2007a, 2007b) compared the efficacy of different metrical measurements using different varieties of English, and concluded that a combination of %V and VarcoV yields the most effective results. Other such attempts at arriving at the *perfect metric* abound in the literature, however one significant contribution made by Grabe and Low (2002) is the revelation that none of these metrical measurements fares very well when applied to (prosodically) non-prototypical, non-W.E. languages. One might wonder whether these metrics, and by extension Dauer's (1983) criteria, were a result of a focus on the phonology of these well-documented W.E. languages.

Arvaniti and Rodriquez (2013) point out that not only have rhythm discrimination experiments been conducted on a very small sub-set of languages, but the languages typically used for such experiments differ in other perceptual factors than timing, such as inherent speaking rate. While Germanic languages are typically spoken with a lower speaking rate, Romance languages employ a much faster rate (cf. Arvaniti & Rodriquez 2013). These *non*-rhythmic factors potentially lead to perceptual discrimination, thus rendering the conclusion that discrimination is due to rhythmic differences moot. In fact, Ramus et al. (2003) report that in their experiments Polish was discriminated from both English (stress timed) and Spanish (syllable timed), even though in that study it is classified as a stress-timed language. Clearly, rhythm (as captured by these metrics) cannot be the sole perceptual cue to inter-language discrimination.

The metrics under discussion here are:

%V: Proportion of vocalic intervals within an utterance, an indicator of overall syllable complexity, obtained by calculating the total duration of the utterance that is taken up by the vowels,

i.e. [(sum total of all the vocalic intervals in the utterance) / (duration of the utterance)] x 100. The basic problem with this approach is that it was developed with languages like English and German in mind, where Vs and Cs are either present in approximately equal amounts, or Cs slightly outnumber Vs, but where this is balanced out by the fact that vowels get lengthened or shortened regularly due to phonotactics, while consonants remain relatively unaffected. Speaking rate, likewise, affects vowel duration much more than consonant durations.

PVI: The pairwise variability index is calculated by taking into account the durational difference between pairs of successive intervals, then taking the absolute value |x| of the difference and dividing it by the mean duration of the pair. For rPVI, the division step is omitted. The division is done to normalize for speaking rates, and is applied to vowels only. Stress-timed languages like English tend to display high scores for nPVI, as they use full vowels as well as reduced vowels.

Varco: Coefficient of variation (of C and V), i.e. [(the standard deviation of vocalic/consonantal interval durations) / (mean of vocalic/consonantal duration)] x 100.

ΔC & ΔV: Standard deviation of the consonantal and vocalic duration of the utterances.

3 Methods

The present study seeks to apply the various methodologies discussed in the preceding sections to an under-documented language, and test whether they are capable of providing a stable account of durational variability. The language considered for this study is Urama, a Papuan language of New Guinea. Urama is ideal as a test case, as its phonotactics are more 'strict' than W.E. languages: all syllables are open, no consonant clusters are allowed, there exists no vowel reduction, and there is no vowel length contrast. Thus, Urama tolerates long strings of vowels, but not of consonants.

Grabe and Low (2002) have pointed out that the proposed metrics are incapable of handling non-prototypical languages, and fail to classify these languages into any fixed rhythmic category (hence *non-prototypical)*. However, to the best of our knowledge, no one has tested how the durational contrasts of more exotic languages are captured by a system built almost entirely upon data from W.E**.** languages. Arvaniti (2012) suggests that in order to tap into the true timing pattern of a language, metric scores must be derived for controlled and uncontrolled data. She suggests two types of control-data: type-1 designed to emulate syllable-timing by eliminating consonant clustering, vowel reductions etc. as much as permissible within the language's phonology, and type-2, designed to do just the opposite and emulate stress-timing. Such methodologies, however, fail to account for languages like Urama, which employs a strict (C)V template for its syllable-structure, while lacking any contrastive lengthening of vowels.

In this study, we employ the three most popular metric-combinations (%V-ΔC, CrPVI-VnPVI and %V-VarcoV) and test their effectiveness in capturing the timing patterns of Urama. We compare the scores to other languages in order to establish a cross-linguistic contrast with an effort to test the extent to which these metrics can reflect the differences in the phonological and phonetic properties of these languages.

3.1 Participants and Stimuli

The participant is a female native speaker of Urama. There were two contexts in which speech data was collected: controlled speech contexts, where the participant was instructed to read and/or repeat sentences, spoken at a moderate rate, and spontaneous speech contexts.

For the controlled speech data, the participant was instructed to read/repeat a declarative sentence that was between 12-19 syllables long, and on average approximately 4-5 seconds in duration. In traditional metric-based rhythm studies the standard practice is to use declarative utterances because they are expected to most accurately approximate running speech (Ramus 1999, Grabe & Low 2002). However, in order to test whether clause type has an impact on the metrics, both interrogative and exclamative versions of the declarative sentence were also recorded for this study. There were 5 sentences constructed in this fashion, yielding 15 (3 conditions x 5 base sentences) sentences total. For the spontaneous contexts, a short (approximately 1.5 minute) narrative was collected, spoken at a rate appropriate for this kind of speech style. It is important to note here that if the metrics indeed capture *rhythm in speech*, a property of the inherent prosody of the language, the scores should be independent of both the type and the duration of the utterances used for analyses.

4 Results

The comparisons of each of the metrics for Urama, including interrogatives (Q) and exclamatives (!) vs. English, Dutch, French, and Spanish (from Arvaniti 2012) are presented below for controlled sentences.

	English	Dutch	French	Spanish	Urama	Urama Q	Urama !
%V	40.1	42.3	43.6	43.8	51.45	54.68	57.1
ΔC	0.054	0.053	0.044	0.047	0.016	0.027	0.031
ΔV	0.046	0.042	0.038	0.033	0.025	0.023	0.035
CrPVI	5.6	6.2	4.8	5.25	0.017	0.019	0.021
VnPVI	67	59.8	44.8	42.5	26.711	26.64	23.68

Table 1: Controlled speech metric scores

The scores for spontaneous speech are compared with English, Spanish, and Italian in Table 2.

	English	Spanish	Italian	Urama
varcoV	61.5	67.6	63.1	67.102
VarcoC	58.1	50.9	52.3	35.299
%V	51.9	53.2	54.7	54.16
ΔC (x100)	63.4	47.3	43.1	3.3
VnPVI	62.9	57.2	51.8	60.547
CrPVI	73.8	51.6	46.1	0.039

Table 2: Spontaneous speech metric scores

What can be seen here are extremely low C-scores, especially CrPVI in spontaneous speech.

With respect to %V, it is predicted that it is languages like Urama where this measure would be most likely to fail. Urama vowels, in any given utterance, outnumber consonants significantly. Hence, the longer the utterance, the more vowels there will be; with an increase in total data, the increase in the amount of Cs and Vs is far from equal. Given the controlled data above, the value ranges from 51.4 (for declaratives) to 57.1 (for exclamatives), which is a larger difference than is present between stress-timed English (40.1) and syllable-timed French (43.6).

There are similar problems with PVI values. In Urama vowel reduction is not a factor, not unlike French. The scores however are far greater than French, which is most likely due to the fact that a very low presence of consonants eliminates durational variability in vowels. Similarly, complete absence of consonant clusters contributes to significantly lower CrPVI scores. With respect to Varco, once again, the syllable structure employed by Urama explains the scores. While Spanish and Urama receive similar Varco V scores, the Varco C scores for Urama are substantially lower than any other language. This is again due to an imbalance in Vs vs. Cs.

Considering the mathematical rationale behind the different metrics employed in rhythm studies, it can be readily observed that Δ-values being simply *standard deviation* of vocalic/consonantal intervals remain unaffected by the *sequential patterning* of durational variability of segments- a key element underlying the perceptual effects of *speech rhythm*. The PVI, however, captures this *sequential patterning* by averaging the durational difference between successive vocalic or consonantal intervals:

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

However, there are a couple of discrepancies present in the way in which PVI measures are usually applied. First, for vocalic intervals a *normalized* version of the PVI measure is used in order to *supposedly* correct for speaking rate and *tempo fluctuations*. This is achieved by relating the difference between two consecutive intervals to the mean of the two durations.

$$n\,PVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1)\right],$$

The effect, however, is a very *local normalization* that actually ends up reducing length differences caused in running speech due to stress, accent and other phonotactic factors. Second, while it may still be argued that the PVI does indeed capture *some of the sequential patterning effects of duration* it still calculates vocalic and consonantal variations separately, and thus fail to capture any perceptual effects of vocalic and consonantal structure on the auditory rhythmic patterns of languages.

A third surprising result is the higher consonantal variability for all languages in spontaneous speech measures. Such results have been reported elsewhere (Barry & Russo 2003), with spontaneous speech from Italian reportedly exhibiting higher CrPVI values. In rhythmic terms, then, such results would suggest that spontaneous speech from the tested languages is *more* stress-timed than controlled utterances. Such differences between controlled and spontaneous speech data is presumably a direct result of seg-

mental lengthening of vowels and sonorants in running speech, and is likely to exhibit variation as a function of syntactic-lexical structure of phrases, focus, speech style, tempo, etc., all of which occur with greater variability and lesser predictability in *undersigned* and *uncontrolled* speech.

Otherwise, Urama follows the pattern of changes in scores exhibited by other stress vs. syllable-timed languages in the tables, such as higher %V scores than stress-timed languages, lower PVI scores for vowels, etc. It tends to follow the syllable-timed languages in its scores when compared to English, with the only difference being that the difference in scores for Urama is substantial, an effect of the extremely V-heavy nature of the syllable-structure.

5 Conclusion

W.E. languages tend to get grouped according to rhythm classes in metrical analysis, because these metrics were *specifically designed with their syllable structure and phonotactics* in mind. They do not reflect rhythm, only co-incidentally their scores for W.E. languages tend to correlate with rhythmic typology because the mathematic underpinnings of the metrics reflect phonotactic properties. The results reported for Urama illustrate how the variation in metric scores correlates with variation in phonotactics. Thus, these metrics only provide a *very crude* measure of timing, illustrated by the confusing inter-language scores. This has obvious implications for speech technology incorporating rhythmic properties, including automatic recognition of emotion (Ringeval et al. 2012), spoken language identification (Timoshenko & Höge 2007), Zhang & Glass 2009), and clinical applications (Selouani et al. 2012).

References

Abercrombie, D. 1967. *Elements of general phonetics* (Edinburgh University Press, Edinburgh).

Arvaniti, A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66:46-63.

Arvaniti, A. 2012. "The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40, 351–373.

Arvaniti, A. & Rodriguez, T. 2013. The role of rhythm class, speaking rate and *F0* in language discrimination. *Laboratory Phonology* 4: 7-38.

Barry, W.J. and Russo, M. 2003. "Measuring rhythm. Is it separable from speech rate?", Proceedings of the International AAI Workshop "Prosodic Interfaces", Nantes 27-29 mars, 2003

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

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

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

Port, R.R., J. Dalby & M. O'Dell. 1987. Evidence for mora-timing in Japanese. *Journal of the Acoustical Society of America* 81:1574-1585.

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

Ramus, F. 2002. Acoustic correlates of linguistic rhythm: Perspectives. Proc. Speech Prosody, Aix-en-Provence 2002:115–120.

Ringeval, F., Chetouani, M., & Schuller, B. 2012. Novel metrics of speech rhythm for the assessment of emotion. *Interspeech 2012*, pp. 2763-2766.

Selouani, S.A., Dahmani, H., Amami, R. & Hamam, H. 2012. Using speech rhythm knowledge to improve dysarthric speech recognition. *International Journal of Speech Technology* 15:57-64.

Tilsen, S. & Arvaniti, A. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America* 134: 628-639.

Timoshenko, E. & Höge, H. 2007. Using speech rhythm for acoustic language identification. *Interspeech 2007*, pp. 182-185.

White, L. & Mattys, S. L. 2007a. Calibrating

rhythm: First language and second language studies. *Journal of Phonetics* 35: 501–522.

White, L. & Mattys, S. L. 2007b. Rhythmic typology and variation in first and second languages., in *Segmental and Prosodic Issues in Romance Phonology,* P. Prieto, J. Mascaró and M. J. Solé (John Benjamins, Amstredam), pp. 237-257.

Zhang, Y. & Glass, J.R. 2009. Speech rhythm guided syllable nuclei detection. *International Conference on Acoustics, Speech and Signal Processing*, pp. 3797-3800.