Long-distance rhythmic dependencies and their application to automatic language identification

Joseph Tepperman and Emily Nava

Rosetta Stone Labs
{jtepperman,enava}@rosettastone.com

Abstract

The perception of rhythmic differences among languages relies on varieties in periodicity within prominence groups. But the consensus in phonetic research on rhythm is that existing measures don't capture true rhythm by that definition - instead, they merely measure short-term timing. This work proposes a new rhythm measure, the Generalized Variability Index (GVI), that examines durational contexts over arbitrarily long linguistic distances. To evaluate this new measure, we conducted a set of experiments in automatic language identification using large amounts of data from 11 languages in the Globalphone and TIMIT corpora. When added to baseline rhythm measures, these new GVI features offer absolute improvement in 11-way language classification accuracy by as much as 12%. Moreover, the addition of wider and wider durational context in the GVI continues to contribute information useful for automatic language ID, abating in usefulness only at a distance of about 10 syllables.

Index Terms: rhythm, language identification, Pairwise Variability Index

1. Introduction

Rhythm is an organizing principle of speech that reflects the temporal factors of a given language [4], and more specifically the language-specific interaction of these factors. Researchers have long sought to formalize along principled lines the mental representation of periodicity in temporal events which gives rise to the perception of rhythmic differences across languages [2]. The nature of the processes that underwrite this regularity has been at the heart of research detailing cross-linguistic differences in rhythm, research that intends to uncover the mechanisms involved in this idealized distribution of regularity in temporal units. The standard metrics designed to capture these distributions have been largely based on segmental duration patterns.

While such metrics do get at certain phonological and phonotactic characteristics, they fail to give an adequate representation of rhythm due to the lack of necessary context of ongoing temporal sequences. In 2009, the journal *Phonetica* published a special issue on "Rhythm in Speech and Language" that included several articles offering what could be fatal indictments of the rhythm measures commonly used in phonetic research on rhythm (and, hence, of the measures also used in automatic analysis of rhythm by speech engineers). Arvaniti [2] challenged the theoretical foundations of the dominant rhythm measures, emphasizing the circular logic of first dividing languages into hypothesized stress-timed and syllable-timed categories and then devising metrics by which to support those distinctions, finally concluding that such measures are "unreliable predictors of rhythm which provide no more than a crude measure of timing." A similar study in the same issue by Barry et al. [3] showed that the common syllable-level Pairwise Variability Index (PVI) is in some ways sensitive to metrical differences in read poetry, but fails to capture the variance in prominence of beats, the prosodic foot-level variations that characterize perceived rhythm and exist above the level of inter-syllable timing. Both studies argued in favor of a new research paradigm, one that would shift analysis away from relative durational measures on the level of syllabic, consonantal, and vocalic intervals, toward a conception of rhythm that would account for variations in prominence patterns within larger rhythmic groupings. A third study [10] in the same special issue of *Phonetica* offered initial support of this new idea of rhythm, concluding that a novel foot-level PVI offered complementary information for distinguishing among languages of different rhythmic types, when used in combination with the syllable-level PVI.

The PVI measures the duration of adjacent regions and takes the average of those values across a given passage. Because of its limited scope and its failure to capture overall rhythmic composition [2], in this study a new metric is proposed, the Generalized Variability Index (GVI), that more adequately encompasses the relevant distribution of properties that constitute rhythm. The proposed GVI addresses the shortcomings of baseline rhythm metrics while maintaining duration as its measurement. The main contribution of the GVI is that it expands the window of analysis to capture a broader stretch of rhythmic dependencies between distant intervals, and in this it is a step toward harnessing the nature of variability as it relates to larger prominence groupings and hence more adequate definitions of perceptual rhythm. The need for greater context when describing the rhythmic environment responds to the understanding of how prominence patterns contribute to a speaker's message. Rhythm sets up an expectation about the intention of the message to be delivered, and the extent and direction of flexibility is relevant for language-specific differences. In as much as the GVI is able to encompass this, it affords predictions about how languages pattern together along a rhythmic dimension, as well as how to predict which language is being spoken, given input from one speaker.

The current study aims to show the benefit of the GVI's long-distance rhythmic dependencies in a speaker-level automatic language identification study, similar to the one conducted in [9]. With data from 11 languages and about 100 or more speakers per language, we hope to conclusively show that this new rhythm measure that accounts for periodicity within and beyond the level of the prosodic foot can offer substantial improvements in automatic rhythmic analysis above those afforded by the standard measures.

language	*abbrev.*	*speakers*	*hours*
Chinese (Mandarin)	CH	132	16.57
English (US)	EN	462	2.70
French	FR	100	23.71
German	GE	77	15.06
Japanese	JA	143	25.32
Korean	KO	99	18.05
Portuguese (Brazil)	PT	101	22.46
Russian	RU	115	21.63
Spanish (Costa Rica)	SP	96	17.19
Swedish	SW	98	17.29
Turkish	TU	100	14.87

Table 1: Statistics for all the data used in this study.

feature	*length*	*definition*
ΔC	1	std. of consonantal intervals
ΔN	1	std. of syllable nuclei
ΔV	1	std. of vocalic intervals
%V	1	proportion of vocalic intervals
varcoC	1	ΔC / mean consonantal interval dur.
varcoN	1	ΔN / mean syllable nuclei dur.
varcoV	1	ΔV / mean vocalic interval dur.
nPVI-C	1	normalized PVI of consonantal intervals
nPVI-N	1	normalized PVI of syllable nuclei
nPVI-V	1	normalized PVI of vocalic intervals
GVI-C$_{2:10}$	9	GVI of consonantal intervals w/ $M = 2, \ldots, 10$
GVI-V$_{2:10}$	9	GVI of vocalic intervals w/ $M = 2, \ldots, 10$
GVI-N$_{2:10}$	9	GVI of syllable nuclei w/ $M = 2, \ldots, 10$

Table 2: Features used in this study. All are baselines except for the GVI features in the last 3 rows. Note that the GVI with context $M = 1$ reduces to 0.5 times the nPVI.

2. Speech Data

The Globalphone Corpus [13] consists of read newspaper speech and corresponding word-level transcripts from 19 languages, with about 100 (or more) native speakers per language. All recordings were collected in the participants' native countries to avoid cross-linguistic effects in pronunciation, and the speech was elicited in noise-free, laboratory-quality conditions. Ten of these languages were selected for the data set in this study. The eleventh language, English, came from comparable recordings in the TIMIT corpus. Statistics about these eleven languages can be found in Table 1. Note that there are roughly four times as many English speakers than any other language - this is because there was much less data per speaker in TIMIT than in Globalphone. By the standards of most past studies in cross-linguistic rhythm comparisons, this is a tremendous amount of data. Consider [7], which examined eighteen languages but with only one speaker of each, or [9], which used 41 speakers total, but only across five languages.

3. Rhythm Measures

3.1. Baselines

Pike [11] and Abercrombie [1] were among those who championed the notion that inter-stress intervals for so-called "stress-timed" languages were of equal length, i.e. isochronous, and likewise for syllable durations in so-called "syllable-timed" languages. But numerous subsequent studies failed to provide acoustic evidence for strict measures of isochrony. Subsequent analyses of these claims suggest that in formulating their original hypotheses, Pike and Abercrombie were responding to the perceptual differences in the distribution of vowel identity across languages (vowel durations, spectral quality, etc.) [8], but that this is not corroborated by actual production data in the same terms of their proposal.

It wasnt until the work of Dauer [5] that the complexity behind rhythmic classification was detailed with any success. Following her widely-accepted analysis, languages were then considered to be organized along a continuum, ranging from more or less syllable-timed to more or less stress-timed based on the language-specific distribution of the following phonotactic properties: vowel reduction, syllable structure inventory, and physical correlates of word-level stress. This inventory of properties was subsequently used to inform the development of metrics designed to incorporate these structural differences.

Among the most common of these measures are as follows: ΔC and ΔV [12], defined as the standard deviations of consonantal and vocalic intervals, respectively; proportion of vocalic intervals (%V), defined as the proportion of speech occupied by vowels; and the rate-normalized variability of consonantal and vocalic intervals (varcoC and varcoV), defined as the above ΔC and ΔV divided by the mean durations of consonantal and vocalic intervals, respectively [6].

By far the most widely-used rhythm measurement is the Pairwise Variability Index (PVI), which calculates the mean durational difference between adjacent intervals (past studies have defined these in terms of vocalic intervals, consonantal intervals, syllables, or feet) [6, 9, 10]. Formally, the nPVI, normalized for speaking rate, is defined as

$$nPVI = \frac{1}{N-1} \sum_{n=2}^{N} \left| \frac{v_n - v_{n-1}}{(v_n + v_{n-1})/2} \right|$$

where v_n is the duration of unit n in a sequence of length N.

All of the aforementioned studies report statistically significant groupings of the languages under study using these metrics, such that values from "stress-timed" languages (English, Dutch, German) group together, as do the values from "syllable-timed" languages (Spanish, French), with "mora-timed" Japanese constituting a third group. Additionally, the use of these measures has extended to studies investigating the acquisition of rhythm by non-native speakers [15]. All the aforementioned measures essentially calculate the same thing: durational differences corresponding to phonological and phonotactic differences. And they all suffer from the same shortcoming: a failure to capture the durational distributions of longer sequences, never extending into measures of variation in prominent beat periodicity.

A summary of the baseline measures used in this study can be found in Table 2. To approximate syllabic durations in data that is not annotated with syllable boundaries, we also calculated measures relative to the syllable nucleus (obtained from all individual vowel boundaries in the transcripts). Many other related measures have been proposed for analysis of rhythmic differences among languages, e.g. the raw (unnormalized) PVI (or rPVI), the rate of speaking (ROS), and the ratio of vocalic to consonantal duration [9]. But, as attempts to capture true rhythm in terms of prominence groupings, many of these measures (e.g. the ROS) are even more flawed than the most common baselines chosen for comparison here.

3.2. The Generalized Variability Index

The Generalized Variability Index (GVI), proposed for the first time in this study, is based on the PVI but allows for comparisons between intervals beyond the PVI's strict adjacencies. Though it is still essentially a pairwise comparison, the GVI is "generalized" in terms of its ability to analyze durational differences on scales of arbitrary length.

Formally, the GVI is defined very similarly to the PVI:

$$GVI = \frac{\sum_{n=2}^{N} \sum_{m=1}^{\min(M,n-1)} \left| \frac{v_n - v_{n-m}}{v_n + v_{n-m}} \right|}{\sum_{i=1}^{M} (N-i)} \qquad (1)$$

In the numerator, the outer sum counts over the whole sequence of N durations, and the inner sum compares all pairs of intervals within the case-defined context variable M. Note that m can be as large as $M = N - 1$, but is constrained by the length of the sequence, hence the $\min(M, n-1)$ limit. Otherwise the numerator is identical to that of the normalized PVI except for the absence of the divisor 2, which is not essential for rate-of-speaking normalization. The GVI's denominator calculates the mean difference over all pairs by dividing by the number of pairs within a given combination of N and M. Note that the GVI is limited to the range $[0, 1)$ - if all durations are equal, the GVI is 0; as the differences in duration pairs approach infinity, the GVI approaches 1. At $M = 1$, the GVI reduces to the ordinary nPVI, but multiplied by 0.5 - this linear factor of 0.5 should not affect the discriminative qualities of the measure.

To reiterate, essentially the main advantage of using the GVI over the traditional PVI is in its ability to incorporate long-distance durational dependencies in calculating a measure of rhythm. The context variable, M, defines the size of the window over which pairwise comparisons are made. Hence the GVI can, for example, compare pairs of syllables within a prosodic foot consisting of many syllables, or between two distant prosodic feet. The traditional PVI only looks at pairs of durations that are adjacent in the sequence, hence it can rarely capture the variability in prominence within a larger rhythmic group. Long-distance dependencies offer measures of the variability that account for perception of rhythmic distinctions, and the GVI is suited for measuring that variability.

4. Experiments

In our previous work in [14], we evaluated an earlier version of the GVI for template-based scoring of nonnative speakers - this was the Pairwise Variability Error (or PVE). There we used long-distance syllable-level durational differences for improved discrimination between native and nonnative English, with the best classification accuracy at $M = 3$ context. Here we are more interested in showing the usefulness of long-distance durational context in distinguishing among multiple languages. To that end, we conducted a set of language ID experiments designed to answer two questions. First, how do the new GVI measures compare with standard rhythm baselines from the literature? And, what is the effect of the GVI's long-distance context on language ID accuracy? For example, is there an ideal length for this context? (Since there is no reference template in this case, the PVE is not applicable.)

All recordings were force-aligned using the corpora's word-level transcripts and Rosetta Stone's proprietary speech recognizer, for which acoustic models and phoneme-level pronunciation dictionaries exist for some 30 languages (including the 11 investigated here). Vocalic and consonantal intervals were defined as unbroken regions of vowels or consonants in the forced alignment. GVI scores were not calculated across pause or sentence boundaries. Each speaker had one set of scores - a 37-dimensional vector corresponding to the scores enumerated in Table 2 - and all classifications were an 11-way choice on the speaker level. All experiments were done on the entire data set using a 10-fold crossvalidation. The classifier chosen was the Support Vector Machine (SVM) implementation in the Weka toolkit. Since we are not interested in absolute accuracy but only the in comparative performance among different feature sets, the default SVM settings were used throughout.

	feature set	*% accuracy*
(1)	GVI-V$_{2:10}$ + GVI-N$_{2:10}$ + GVI-C$_{2:10}$	61.65
(2)	*baselines*	67.04
(3)	*baselines* + GVI-V$_{2:10}$	67.96
(4)	*baselines* + GVI-N$_{2:10}$	71.70 *
(5)	*baselines* + GVI-C$_{2:10}$	74.92 *
(6)	(1) + (2)	79.05 *

Table 3: Language ID performance of various feature sets.

CH	*EN*	*FR*	*GE*	*JA*	*KO*	*PO*	*RU*	*SP*	*SW*	*TU*	↘
110	1	2	0	0	8	5	0	2	4	0	*CH*
1	433	2	4	2	0	1	0	5	13	1	*EN*
3	0	75	0	0	3	0	5	7	0	7	*FR*
0	3	0	41	0	3	1	5	0	19	5	*GE*
0	0	0	0	138	0	2	0	1	0	2	*JA*
11	5	10	1	0	54	1	1	6	4	6	*KO*
4	0	1	1	2	0	76	0	1	3	13	*PO*
1	0	4	1	0	2	3	95	0	4	5	*RU*
4	11	8	0	0	1	0	0	59	0	13	*SP*
1	9	0	6	0	5	0	3	0	74	0	*SW*
4	1	5	2	0	9	3	6	19	2	49	*TU*

Table 4: Confusion matrix of the best language ID classifier. Columns are targets, rows are classification results. See Table 1 for a key to the abbreviations.

To evaluate the GVI's performance against the baseline measures, we conducted the language ID classification using the following sets of features: the baselines, all GVI measures alone, and the full set of combined features. Then, to compare among consonantal, vocalic, and nuclear GVIs, we added each one of these subsets of features to the baseline set. The performances of these feature combinations are given in Table 3. The language ID confusion matrix for the best feature set (the complete set of features) is given in Table 4.

To examine the effets of context in the GVI measure, we looked at the contribution to language ID accuracy when adding each successive level of context to the feature set. Starting with just the consonantal, vocalic, and nuclear GVI features corresponding to context $M = 1$ in Eqn. 1, we added GVI features from context $M = 2$, then added $M = 3$, and so on, up to $M = 10$. Recall that at $M = 1$, the GVI reduces to the PVI times a factor of 0.5. Results of these classification experiments are listed in Table 5. As a further illustration, Figure 1 shows the distribution of these 11 languages according to GVI-N and context length.

5. Discussion

It is clear from Table 3 that the proposed GVI features are valuable for speaker-level language identification (all scores marked with * were significantly better than the baseline with $p < 0.01$). Adding all GVIs to the baselines resulted in a 12% absolute improvement in classification accuracy. Considering wider and wider contexts in the GVI continued to improve classification accuracy in Table 5, up to and including the very wide context of $M = 9$. For syllable nucleus GVIs, $M = 9$ is like comparing durations between two distant syllabic nuclei that lie 8 syllables apart. Capturing long-distance dependencies does indeed improve automatic language ID, by spanning many prosodic feet and hence whole groups of prominent beats.

One cannot conclude the baseline features to be useless: in

GVI context	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
% accuracy	44.65	50.43	54.83	58.83	60.80	61.79	62.90	64.35	67.17	66.58

Table 5: Language ID performance as a function of GVI context. Feature sets and results are cumulative from left to right.

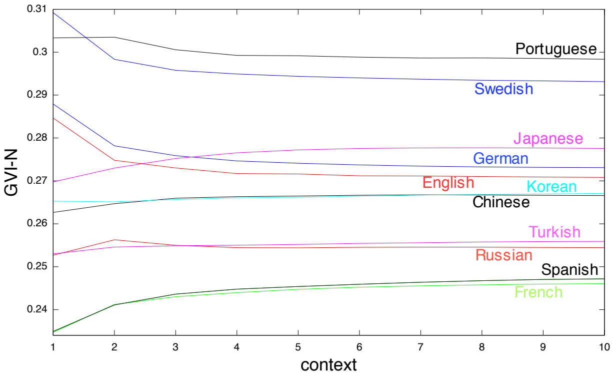

Figure 1: Language distribution by GVI-N and context.

isolation, they outperform the isolated GVIs, and are of smaller cardinality - 10 baseline measures compared to the 27 GVIs. Among the GVI features, it appears that the consonantal (GVI-C) have the most explanatory power and offer the greatest improvement in accuracy - they are potentially capturing phonotactic distinctions in syllable structure that lead to differing perceived rhythms across these 11 languages.

What can the GVI tell us about about the Rhythm Class Hypothesis and the notion of isochrony? In analyzing the confusion matrix in Table 4, we see that many of the most commonly confused pairs lie within traditional language categories: stress-timed (e.g. German is often confused here with Swedish) or syllable-timed (e.g. Spanish is often confused here with Turkish). The plot of GVI-N vs. context over all 11 languages (Fig. 1) also illustrates a spectrum from languages previously hypothesized as syllable-timed to ones hypothesized to be stress-timed - the syllable nucleus is an approximation of the entire syllable's length and hence the GVI-N can potentially capture syllable-level isochrony if it exists. Languages traditionally termed syllable-timed (French, Spanish, Turkish) cluster toward the bottom of the plot, with lower pairwise variability, while languages thought to be stress-timed (Swedish, German, English, and debatably Portuguese) cluster at the top with higher pairwise variability - though there are a few exceptions to this general observation (e.g. Russian, Japanese).

Also notable is the shape of the curves along the context axis in Fig. 1: languages with low initial GVI (at context $M = 1$) tend to increase in variability with the addition of wider contexts; the opposite is seen for languages with relatively high variability at $M = 1$. This can be interpreted as evidence that the GVI is sensitive to durational contrasts within and across the prosodic foot. High variability at short contexts indicates a long-short distinction between adjacent durations, as seen in languages in which the basic rhythmic unit is the foot - this was part of the reasoning behind the original PVI measure. With the addition of wider context, high adjacent variability is diluted by comparisons between syllables of similar lengths that may be quite distant. For languages in which adjacent syllables have low variability (i.e. ones hypothesized to be syllable-timed), widening the context adds noise to the variability measurement by comparing distant syllables that may differ in length due to non-rhythmic reasons like speaking rate.

6. Conclusion

In the task of Automatic Language Identification, the new rhythm measure proposed in this paper has proven to provide complementary information when paired with existing baseline measures. With the GVI's wide context of distant durational comparisons, we demonstrated absolute improvements in 11-way classification by up to 12%. The GVI measure is also a step toward a new paradigm in phonetic research on rhythm, one in which the measures used to describe rhythm will reflect its proper definition in terms of periodicities in prominence groups, rather than as variations in short-term timing. Essentially this study took a short-term timing measure (the PVI) and adapted it to rhythmic analysis on an arbitrarily large scale. Future work is needed to develop better measures that will not only compute long-distance pairwise durational variability, but also characterize distributions in stress patterns beyond pairwise comparisons.

7. References

[1] D. Abercrombie, *Elements of General Phonetics*, Chicago: Aldine, 1967.

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

[3] W. Barry and J. Koreman, "Do Rhythm Measures Reflect Perceived Rhythm?" *Phonetica*, 66:1-17, 2009.

[4] E. H. Buder, "Dynamics of speech processes in dyadic interaction," in *Dynamic Patterns in Communication Processes*, Ed. J. H. Watt and C. A. Vanlear, Sage: Thousand Oaks, CA, 1996.

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

[6] V. Dellwo, A. Fourcin and E. Abberton, "Rhythmical classification based on voice parameter," in *Proc. of IICPhS*, 2007.

[7] E. Grabe and E.L. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis," *Papers in Laboratory Phonology 7*, Ed. N. Warner and C. Gussenhoven, Berlin: Mouton de Gruyter, 2002.

[8] I. Lehiste, "Isochrony reconsidered," *Journal of Phonetics*, 5:253-263, 1977.

[9] A. Loukina, G. Kochanski, C. Shih, E. Keane, and I. Watson, "Rhythm measures with language-independent segmentation," *Proc. of Interspeech*, Brighton, 2009.

[10] F. Nolan and E.L. Asu, "The Pairwise Variability Index and Coexisting Rhythms in Language," *Phonetica*, 66:64-77, 2009.

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

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

[13] T. Schultz, M. Westphal and A. Waibel, "The GlobalPhone Project: Multilingual LVCSR with JANUS-3," *Proc. of 2nd SQEL Workshop*, Plzen, Czech Republic, 1997.

[14] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing Suprasegmental English Through Parroting," in *Proceedings of Speech Prosody*, Chicago, 2010.

[15] L. White and S.L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, 35:501-522, 2007.