A Comparison of Rhythm Metrics for L2 Speech

Kakeru Yazawa1, Mariko Kondo2

1Faculty of Humanities and Social Sciences, University of Tsukuba, Ibaraki, Japan
2School of International Liberal Studies, Waseda University, Tokyo, Japan

Abstract

A wide range of rhythm metrics (global metrics: %V, Δ, Varco, and segVarco; pairwise metrics: rPVI, nPVI, CCI, and ΔnCCI) was applied to L1 Japanese speakers’ L2 English speech data. Less proficient Japanese speakers of English are expected to show less durational variability for both vocalic and consonantal intervals (because of insufficient stress realization and transfer of CV syllable structure), although this pattern may be obscured by their slower speech rate (which increases interval durations in general). To test if the metrics can capture the L2 rhythmic characteristics, each metric was applied to read speech samples of “The North Wind and the Sun” by 183 Japanese speakers in the J-AESOP corpus. Only %V, VarcoV, and segVarcoVC were successful; other metrics yielded inconsistent or implausible results likely due to insufficient rate normalization. The overall results indicate that global metrics can effectively quantify L2 rhythm if speech rate is normalized by the mean duration of segments (which is a good predictor of tempo) rather than the mean interval duration (which is popular but susceptible to syllable complexity).

Index Terms: rhythm metrics, speech rate normalization, L2, Japanese, English

1. Introduction

1.1. Rhythm metrics

In the last few decades, researchers proposed various metrics to quantify the rhythmic differences in languages [1]. Ramus et al. [2] first proposed three metrics (%V, ΔV, and ΔC) based on the segmentation of speech sounds into alternating vocalic and consonantal intervals. %V is the percentage of the duration occupied by vocalic intervals within the utterance, and ΔV and ΔC are the standard deviations of the durations of the vocalic and consonantal intervals, respectively. Low et al. [3] then proposed the Pairwise Variability Index (PVI), which makes pairwise comparisons between successive vocalic or consonantal intervals. Raw PVI (rPVI) is the average durational difference between successive intervals:

\[ rPVI = \frac{\sum_{k=1}^{m-1} |d_k - d_{k+1}|}{(m - 1)} \]  

where \( m \) is the number of intervals within the utterance and \( d_k \) is the duration of the \( k \)th interval.

It was soon apparent that these metrics (except %V) are sensitive to speech rate. Grabe and Low [4] therefore introduced a rate-normalized version of PVI (nPVI), where the durational difference between successive intervals is divided by the mean duration of those intervals to control for tempo:

\[ nPVI = 100 \cdot \frac{\sum_{k=1}^{m-1} |d_k - d_{k+1}|}{\sum_{k=1}^{m-1} d_k} / (m - 1) \]  

Similarly, Dellwo [5] proposed the variation coefficient (Varco) for Δ, which uses the mean vocalic or consonantal interval durations (\( \mu \)) for rate normalization:

\[ \text{Varco} = 100 \cdot \frac{\Delta}{\mu} \]  

These metrics have been widely used in the literature, but susceptible to syllable complexity [6]. Thus, more recently proposed metrics consider the number of segments constituting the intervals. For example, our previous study [7] proposed a modified version of Varco called segVarco (formerly nVarco), where Δ is divided by the mean duration of vocalic or consonantal segments (\( \mu_{\text{seg}} \)) rather than intervals:

\[ \text{segVarco} = 100 \cdot \frac{\Delta}{\mu_{\text{seg}}} \]  

Further, Bertinetto and Bertini [8] proposed the Control/Compensation Index (CCI), which relativizes the rPVI formula to the number of segments composing successive vocalic or consonantal intervals:

\[ CCI = 100 \cdot \frac{\sum_{k=1}^{m-1} \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}}{(m - 1)} \]  

where \( n_k \) is the number of segments within the \( k \)th interval. A rate-normalized version of CCI (\( \Delta nCCI \)) was later proposed by adopting the nPVI normalization strategy [9]:

\[ \text{\( \Delta nCCI \)} = 100 \cdot \frac{\sum_{k=1}^{m-1} \left| \frac{n_k}{\mu_{\text{seg}}} - \frac{n_{k+1}}{\mu_{\text{seg}}} \right|}{(m - 1)} \]  

1.2. The current study

This study tests whether the non-pairwise (henceforth “global”) and pairwise metrics presented above can accurately capture the rhythmic characteristics of L2 English spoken by L1 Japanese speakers. Unlike stress-timed English, Japanese is a mora-timed language without lexical stress or vowel reduction. Thus, vocalic variability metrics like ΔV, nPVI-V, and CCI-V show smaller values for Japanese than for English [10]. The syllable structure of Japanese is fundamentally (CV), simpler than that of English with complex consonant clusters, meaning consonantal variability metrics like ΔC, rPVI-C, and CCI-C show smaller values for Japanese than for English [10]. These cross-linguistic differences likely affect L1 Japanese speakers’ L2 English rhythm, the extent to which would depend on their L2 proficiency levels, as summarized below:

(a) Less proficient speakers show less durational variability for vocalic intervals (because of insufficient stress realization and vowel reduction [11–13]).

(b) Less proficient speakers show less variability for consonantal intervals (because the L1 CV syllable structure is repaired via vowel epenthesis or consonant elision [14, 15]).
However, not all metrics may successfully capture (a) and (b) because of the influence of speech rate. Here, less proficient L2 learners speak more slowly [16], exhibiting more durational variation in general. The rate effect may thus counterbalance or even outweigh (a) and (b), making some metrics ineffective. Kawase et al. [17] confirmed this by applying $\Delta$, Varco, and $r$-PV1 to "experienced" and "inexperienced" Japanese speakers' English. In their study, rate-normalized VarcoV and nPV1-V showed results consistent with (a), but unnormalized $\Delta$C did not show a significant group difference. Unnormalized $\Delta$C showed larger values for the "inexperienced" group contrary to (b), while rate-normalized VarcoC showed a non-significant result. However, these results must be interpreted carefully because only four (two "experienced" and two "inexperienced") speakers were tested. The current study therefore examines a larger number of speakers, with additional metrics such as segVarco and $n$-CC1, to test which metric can best capture the L2 rhythmic characteristics despite the adverse effect of speech rate.

2. Method

2.1. Speakers and materials

Read speech samples of "The North Wind and the Sun" [18] by 183 Japanese learners of English and 25 native English speakers were obtained from the J-AESOP corpus [15]. To assess the learners’ levels of English proficiency, 16 phonetically trained judges assessed each speakers' sample using four criteria (segmental accuracy, prosody, fluency, and nativelikeness) on a scale from 1 to 10 each. This study focuses on the prosody score, where lexical stress realization, overall speech rhythm, and lack of epenthesis and elision were evaluated. Inter-rater correlations of the prosody score were generally high ($r = 0.73 \text{ – } 0.92$). The learners were assigned to seven groups according to their mean prosody scores so that each group is comparable to the native group ($n = 25$) in size: "1" ($n = 26$, score $< 3.86$), "2" ($n = 26$, score $< 3.86$), "3" ($n = 26$, score $< 4.50$), "4" ($n = 26$, score $< 5.11$), "5" ($n = 26$, score $< 5.62$), "6" ($n = 26$, score $< 6.46$), and "7" ($n = 26$, score $< 7.78$). A larger number indicates higher proficiency.

2.2. Segmentation and measurements

The speech samples were segmented into vocalic and consonantal intervals using the corpus' standard annotation data. The annotations are the output of automatic forced alignment using [19] and [20] with subsequent hand corrections by trained phoneticians in the J-AESOP team. A vocalic interval consisted of /i, e, æ, o, Ə, u, oʊ, o, ɔ, ɔɪ, ǝ, ɤ/ and epenthetic vowels. A consonantal interval consisted of /p, t, k, b, d, g, f, v, θ, ð, s, z, j, ʃ, h, ɬ, ʃʃ, j, w, l, m, n, nj/ and epenthetic consonants. Rhotic vowels were considered vocalic, and the closure duration of stops and affricates consonantal. Successive segments of the same type were merged (e.g., [spa] as one consonantal interval) unless there was a temporal pause in between.

This study adopts the inter-pausal unit (IPU) as the basic unit of analysis. An IPU is defined as a series of speech segments surrounded by two pauses no shorter than 200 ms [21]. For each IPU, the global ($\%$V, $\Delta$, Varco, segVarco) and pairwise ($r$-PV1, nPV1, CCI, and $n$-CC1) metrics were calculated following the procedures stated in Section 1.1. Extraneous phenomena like misreading, word repetition, word insertion, disfluency, and fillers were excluded.

3. Results

3.1. Global metrics

3.1.1. $\%$V and $\Delta$

Figure 1 shows the mean $\%$V ±1 standard errors for the learner groups ("1" – "7") and the native group ("N"). The learner groups show a larger $\%$V than the native group, where $\%$V and proficiency levels are inversely related. This was expected because less proficient speakers insert a greater number of epenthetic vowels to preserve the CV syllable structure [14].

Figure 1: By-group mean $\%$V ±1 standard errors.

Of more relevance to the current study is the unnormalized variability metric $\Delta$. Figure 2 shows the mean $\Delta$V and $\Delta$C ±1 standard errors for each group. Contrary to (a) and (b), the learner groups exhibit larger $\Delta$V and $\Delta$C than the native group, where $\Delta$ and proficiency levels are inversely related. The result partially replicates [17], which found the same tendency for $\Delta$C but not for $\Delta$V. The sensitivity of $\Delta$ to speech rate is evident when comparing Figure 2 and Table 1, which summarizes the groups’ speech rates as measured by the number of segments per second. The fewer the segments per second (i.e., the slower the rate), the larger the $\Delta$.

Figure 2: By-group mean $\Delta$ ±1 standard errors.

Table 1: By-group number of segments per second.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>s.e.</th>
<th>Group</th>
<th>Mean</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.03</td>
<td>0.08</td>
<td>5</td>
<td>8.88</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>8.86</td>
<td>0.08</td>
<td>6</td>
<td>9.09</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>8.58</td>
<td>0.07</td>
<td>7</td>
<td>9.67</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>8.91</td>
<td>0.07</td>
<td>N</td>
<td>10.34</td>
<td>0.09</td>
</tr>
</tbody>
</table>
3.1.2. Varco

We now turn to Varco, in which $\Delta$ is divided by the mean duration of intervals for rate normalization. Figure 3 shows the mean VarcoV and VarcoC ±1 standard errors per group. The result for VarcoV aligns with (a), where less proficient groups show smaller values, replicating [17]. In contrast, VarcoC yielded an inconsistent result, where the least proficient learner group (“1”) and native group (“N”) show similar values, as do the other learner groups (“2” to “7,” perhaps except for “6”). The large standard errors of VarcoC indicate that the mean duration of consonantal intervals (i.e., $\mu$ in (3)) fluctuates greatly, likely as a function of syllable complexity. For example, the consonantal interval duration would be inevitably longer for [spa] than for [s] even when spoken at the same tempo.

Figure 3: By-group mean Varco ±1 standard errors.

3.1.3. segVarco

Figure 4 shows the result for segVarco, which extends Varco by using the mean duration of segments rather than intervals. Both segVarcoV and segVarcoC are proportionate to learners’ proficiency levels as per (a) and (b). The most proficient learner group (“7”) shows a nearly identical segVarcoV to the native group, similar to a previous finding that Japanese speakers can acquire a native-like durational realization of English stress [11]–[13]. The result for segVarcoV is somewhat comparable to VarcoV in Figure 3, because a vocalic interval often consists of a single segment, in which case $\mu_{seg}$ equals $\mu$. In contrast, the result for segVarcoC differs drastically from VarcoC, because a consonantal interval tends to vary in the number of constituting segments.

Figure 4: By-group mean segVarco ±1 standard errors.

3.2. Pairwise metrics

3.2.1. rPVI

Regarding pairwise metrics, Figure 5 shows the results for rPVI-V and rPVI-C for each group. Although rPVI is usually not applied to vocalic intervals (nPVI is used instead; see [4] for an explanation), it is done here for the sake of comparison. The results show a striking resemblance to $\Delta V$ and $\Delta C$ in Figure 2, where less proficient learner groups show larger values for both vocalic and consonantal intervals, contrary to (a) and (b). This confirms the influence of speech rate and indicates the necessity of rate normalization.

Figure 5: By-group mean rPVI ±1 standard errors.

3.2.2. nPVI

Figure 6 presents the results for nPVI, which normalizes rPVI by using the mean duration of successive intervals. Although nPVI is usually not applied to consonantal intervals (rPVI is used instead; see [4]), it is done so for comparison sake. The result for nPVI-V appears random: learner groups “1,” “5,” “6,” and “7” show smaller values than learner groups “2,” “3,” and “4,” with the “N” group placed in between for no clear reason. The result for nPVI-C remains similar to that for rPVI-C in Figure 5, where less proficient learner groups show larger values, reflecting their slower speech rate. The overall result raises the question as to whether the normalization method adopted in nPVI is effective. The weakness of the nPVI method was also pointed out by Bertini et al. [9]: “Suppose that, among two syllables, one is stressed: by normalizing over this sequence, one disruptively wipes out the stress effect.”

Figure 6: By-group mean nPVI ±1 standard errors.
CCI is a modification of rPVI, which considers the number of intervals within successive intervals. Figure 7 shows the mean CCI-V and CCI-C for each group (the standard errors are too small to draw properly). CCI must be interpreted differently from the other metrics. Specifically, “controlling” (syllable- or mora-timed) languages like Japanese should show values along the bisecting line because the variability of vocalic and consonantal segments should be comparable. “Compensating” (stress-timed) languages like English should show values below the bisector because the variability should be larger for vocalic segments than for consonantal segments. A previous study [10] confirmed that Japanese was along or above the bisector (depending on whether devoiced vowels were labeled as vocalic or consonantal) and English below the bisector. However, in Figure 7, all groups show an equally “compensating” rhythm, which is implausible given the expected influence from the L1 Japanese “controlling” rhythm. This may be because CCI does not control for speech rate.

3.2.4. \( \mu nCCI \)

The final metric is \( \mu nCCI \), which adopts the nPVI normalization strategy for CCI. Figure 8 shows the mean \( \mu nCCI-V \) and \( \mu nCCI-C \) per group (the standard errors are again too small to draw properly). The general tendency remains similar to the unnormalized CCI in Figure 7, indicating the insufficiency of the nPVI normalization method. Furthermore, less proficient groups show even more “compensating” rhythm in Figure 8, which is again not very plausible.

4. Discussion

This study compared the performance of various rhythm metrics as applied to L1 Japanese speakers’ L2 English speech. The results have important implications for the applicability of these metrics to L2 speech, as discussed below.

The results for the global metrics were consistent with the prediction that (a) and (b) will be masked by the effect of speech rate. Unnormalized \( \Delta \) was heavily influenced by rate differences, as can be seen in Figure 2 and Table 1. Although rate-normalized Varco was expected to solve the problem, only VarcoV was successful. This is attributable to the nature of vocalic and consonantal intervals, where the former typically consists of one segment while the latter varies in the number of constituting segments. The mean interval duration (i.e., the normalizing factor) is therefore prone to syllable complexity in VarcoC but not in VarcoV. The overall success of segVarco, which considers the number of segments constituting vocalic and consonantal intervals, supports this claim. Taken together, global metrics seem effective if speech rate is normalized by the mean duration of segments instead of intervals.

The results for the pairwise metrics were similar in that unnormalized rPVI was dependent on rate (Figure 5 and Table 1) and that the nPVI normalization using mean interval durations was unsuccessful. However, (\( \mu \))CCI that considers the number of segments also diverged from the expectations, as all learner groups showed a native-like “compensating” rhythm. This does not mean CCI is incapable of capturing L2 speech rhythm, though. There is another normalization method for CCI called \( \mu nCCI \) which controls for speech rate “by referring the individual segment’s duration to the mean duration of the members of the corresponding natural phoneme class” [9]. Because \( \mu nCCI \) performs better than \( \mu nCCI \) [9], the metric may yield a different result for the current dataset. A practical obstacle to applying \( \mu nCCI \) to L2 speech, though, is that it is unclear to which natural class to assign each segment. For example, Bertini and Bertinetto [9] proposed classifying vowel segments based on phonological height, but is it L1 or L2 height? What about other features such as backness, roundedness, tenseness (English), and length (Japanese)? As stated in [9], \( \mu nCCI \) is an exploratory attempt, and more research is needed on how the metric can be extended to L2.

In conclusion, the overall results suggest that existing rhythm metrics can detect and capture the rhythmic characteristics of L2 speech, but only if given adequate rate normalization. The popular normalization method using the mean interval duration is not sufficient because interval durations are prone to syllable complexity. Using the mean segmental duration seems to be a promising approach for global metrics, while pairwise metrics may need to additionally consider more detailed information such as each segment’s natural class. Although the validity of rhythm metrics has been generally questioned [6], insufficient rate normalization in previous metrics may be at fault. Future research can test the applicability of segVarco and \( \mu nCCI \) to a different dataset (e.g., different L1-L2 combinations).

5. Acknowledgments

This research was supported by JSPS KAKENHI Grants 19K21638 and 20K13144 and Waseda University Research Grant BARDO12073. The authors express their gratitude to Yumi Ozaki, the first author of an earlier study, for stimulating our interest in this issue.
References


