Speech Rhythm Metrics for Automatic Scoring of English Speech by Korean EFL Learners[*]

Tae-Yeoub Jang (Hankuk University of Foreign Studies)

<Contents>

<Abstract>

Speech Rhythm Metrics for Automatic Scoring of English Speech by Korean EFL Learners

Tae-Yeoub Jang

Knowledge in linguistic rhythm of the target language plays a major role in foreign language proficiency. This study attempts to discover valid rhythm features that can be utilized in automatic assessment of non-native English pronunciation. Eight previously proposed and two novel rhythm metrics are investigated with 360 English read speech tokens obtained from 27 Korean learners and 9 native speakers. It is found that some of the speech-rate normalized interval measures and above-word level metrics are effective enough to be further applied for automatic scoring as they are significantly correlated with speakers' proficiency levels. It is also shown that metrics need to be dynamically selected depending upon the structure of target sentences. Results from a preliminary auto-scoring experiment through a Multi Regression analysis suggest that appropriate control of unexpected input utterances is also desirable for better performance.

* Keywords: Rhythm metrics, Non-native rhythm, Auto-scoring, Pronunciation diagnosis.

[*] This work was supported by Hankuk University of Foreign Studies Research Fund of 2007.

1. Introduction and Background

The rhythm-based binary characterization of languages into either *stress-timed* or *syllable-timed*, initiated by Pike [1] and reinforced by Abercrombie [2], has been challenged or supported by a number of succeeding investigations.[1] The disapproval of the idea is based on experimental studies showing no inter-syllable isochrony in alleged syllable-timed languages [3][4], or no inter-stress isochrony in alleged stress-timed languages [5][6][7][8][9]. On the other hand, language acquisition studies give evidence for the rhythm typology proving that infants are able to discriminate their native language from other languages, solely on the basis of differences in rhythm patterns [10][11][12].

Avoiding direct disputes, but approving differences in speech rhythm, more recent research has concentrated on discovering acoustic correlates of speech rhythm by which languages with diverse rhythm structures can be distinguished from each other. Based on the observation that stress-timed languages have a more complex and variable syllable structure than syllable timed languages, [13] proposes metrics (%V, the proportion of V; ΔV, and ΔC, the standard deviation of V and C, respectively) obtained by measuring temporal characteristics of vocalic and consonantal intervals. Modifying these metrics, two speech rate-normalized variability measures (VarcoC [14] and VarcoV [15]) are suggested. Grabe and her colleagues [16][17] attempt to capture the sequential temporal variability of vowels and consonants within the sentence, introducing pair-wise variability indices (nPVI, rPVI). These rhythm metrics have been cross-linguistically verified relatively successfully through a number of phonetic studies in various language environments (see [15], for review). In the mean time, Abercrombie's [2] strong version of dichotomy seems to have been moderated into a slightly softer version of typology describing that each language can be located at one point on a continuum from 'extremely syllable-timed' to 'extremely stress-timed.'

Given that those rhythm metrics provide meaningful results in identifying or, at least, differentiating languages in prosodic terms, the rhythm-based distinction between native and non-native speech also seems to be available as long as the target language (L2) of a learner is different in rhythm structure from his/her native language (L1). However, the results of previous investigations of L2 speech rhythm influenced by L1

[1] [18] defines yet another rhythm unit *mora*, introducing *mora-timed* as the third category of speech rhythm. In the current study, however, it is tentatively regarded as a sub-category of the syllable-timed rhythm, as this issue is largely irrelevant to processes in the current study. See [19], [20] and [21] for detailed discussions on *mora* as a rhythm unit.

are not uniform. [17] shows that Singapore English is affected by Mandarin Chinese speakers in terms of nPVI-V. [22] reports that ΔC and rPVI-C, but not %V, ΔV or nPVI-V, can be used to significantly distinguish between native Latvian speakers and Russian learners of Latvian or between learner groups of different proficiencies. On the other hand, [15] finds that such vowel-based metrics as %V and VarcoV are more useful for non-native speech rhythm detection than ΔV, ΔC, nPVI-V, or rPVI-C. A meaningful result from their investigation is that the values from L2 speakers--Spanish speakers of English and English speakers of Spanish--are located between the values of L1 speakers of each language, which provides evidence for the intermediate languages located between two contrastive rhythm patterns. Research by [23] reveals other metrics, including speech rate and syllable duration, can be a cue to detection of non-native Norwegian speech by various different L1 speakers. Diverse results of previous research are summarized in <Table 1>.

<Table 1> Results of research on non-native speech rhythm

	Metrics Approved	Metrics of no significance	L1	L2
Low et al. (2000) [17]	nPVI-V		Mandarin Chinese	English
Stockmal et al. (2005) [22]	ΔC, rPVI	%V, ΔV, nPVI	Russian	Latvian
White and Matty (2007) [15]	%V, VarcoV	ΔV, ΔC, PVI-V, rPVI-C	Spanish, English	English, Spanish
van Dommelen (2006) [23]	nPVI-V, speech rate, syllable duration		Norwegian	7 languages incl. English

In such L1-L2 language contexts as Chinese-English, Spanish-English, or English-Spanish, V-based metrics appear to be more effective though specific metrics are still inconsistent. On the other hand, consonantal metrics work better in Rusian-Latvian. Diverse results lead to establish a conjecture that useful rhythm metrics to assess non-native rhythm need to be selected language particularly (either L1 or L2, or both) as is also implied in [24]'s Estonian rhythm characterization. It is also a general tendency that speech rate normalized rhythm metrics work better than raw metrics. These issues will be further addressed later in this paper.

Two studies [25][26] are interesting in that they investigate non-native English speech rhythm produced by Korean EFL learners. Based on the values of nPVI measured on stress-contrastive syllables/feet and vowels in stress-reduced syllables, they show that Korean EFL learners' speech can be characterized as compared to native speech in terms of the variability index.

The purpose of the current study is to investigate whether various rhythm metrics can be used to capture the non-native rhythm structure of the utterances spoken by Korean learners of English. As the ultimate goal is to utilize the valid metrics in assessing the pronunciation proficiency of non-native speech, automatic extraction of each metric from input speech is attempted. Based on the result of verification, a preliminary auto-scoring experiment is also carried out in order to demonstrate a practical method of employing metrics.

A basic assumption underlying the current approach is that the rhythm structure of spoken Korean is considered to be different from that of English. Although it is generally agreed that the spoken Korean is syllable-timed (see [27], however), and I also adopt this categorization, the current study is not critically affected by the validity of this issue. Rather, the results of the study are expected to provide clues to where the Korean language is located on a continuum between syllable-timing and stress-timing.

2. Definition of Rhythm Metrics

For the convenience of description, the rhythm metrics to be verified are grouped into four types: raw interval measures (RIM), rate-normalized interval measures (NIM), pair-wise variability indices (PVI) and above-word-level measures (AWM). Before delving into verification of their usability in non-native speech discrimination, it is necessary to define the metrics and recapitulate rationale behind their emergence.

2.1. Raw interval measures (RIM)

As observed in [9], stress-timed languages are likely to have a greater complexity in syllable structure (e.g., an onset or/and a coda with consonant clusters in English) than syllable-timed languages and unstressed or destressed syllables in stress-timed languages tend to become reduced to a great degree both in quality and quantity. Based on this observation, [13] introduces three interval measures, %V, ΔV, and ΔC that can be used to capture rhythm differences in languages, which can be summarized as follows:

(M1) %V: the proportion of vocalic intervals in the sentence
Expectation: Syllable-timed language > Stressed-timed language

(M2) ΔV: the standard deviation of vocalic intervals within the sentence
Expectation: Syllable-timed language < Stressed-timed language

(M3) ΔC: the standard deviation of consonantal intervals within the sentence
Expectation: Syllable-timed language < Stressed-timed language

In the current study, Korean EFL speakers' English utterances are expected to be affected by the persevering syllable-timed rhythm pattern of their native language and the degree of this negative transfer will vary dependent on each learner's proficiency level of their spoken English. Consequently, it is hypothesized that %V in Koreans' utterances is greater than that in native speakers' utterances while ΔV and ΔC are smaller than those in native speakers' utterances.

2.2. Rate-normalized interval measures (NIM)

Studies have found that ΔC varies considerably by speech rate at least in some languages including English and German [28][29]. If this is the case, speech rate normalization of target utterances seems inevitable when variability measures such as ΔC and ΔV are to be used. Alternatively, [14] suggests the rate-normalized version of consonantal variability, VarcoC. Later, VarcoV is added to list by [15], to fill up the metrics inventory.

(M4) VarcoC: rate-normalized standard deviation of consonantal interval duration
Expectation: Syllable-timed language < Stressed-timed language
Calculation formula: $VarcoC = \frac{\Delta C \times 100}{mean C}$

(M5) VarcoV: rate-normalized standard deviation of vocalic interval duration
Expectation: Syllable-timed language < Stressed-timed language
Calculation formula: $VarcoV = \frac{\Delta V \times 100}{mean V}$

2.3. Pair-wise variability indices (PVI)

Grabe and Low [16] aim to quantify durational variability between segments in successive syllables, on the assumption that duration of segments is more variable in stress-timed languages due to complex syllable structures and frequent vowel reductions. They claim that pair-wise variability measures are immune to speech rate variation while Ramus's metrics (ΔV and ΔC) are quite vulnerable in such a way that their values are inversely related to speech rate.

(M6) nPVI-V: rate-normalized pair-wise variability index of vocalic intervals

Expectation: Syllable-timed language < Stressed-timed language

Calculation formula: $$nPVI = 100 \times \left(\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| \right) / (m-1)$$

d_k: duration of k^{th} interval

m: number of intervals

(M7) rPVI-C: raw pair-wise variability index of consonantal intervals

Expectation: Syllable-timed language < Stressed-timed language

Calculation formula: $$rPVI = \left(\sum_{k=1}^{m-1} |d_k - d_{k+1}| \right) / (m-1)$$

2.4. Above-word-level measures (AWM)

The metrics proposed by previous research are employing segmental information based on the assumption that syllable structure complexity and vowel reduction rate are significantly affected by rhythm patterns of a language.[2] Features above that level will also have clues to rhythm patterns.

Unaccented function words in stress-timed languages are believed to be reduced to a considerable extent in duration as well as in quality (i.e., spectral features), resulting in a great difference in duration as compared to content words. Consequently, I propose that the proportion of function words within a sentence be employed as another rhythm metric. Although definition of function words may vary, I regard words in the following categories as function words: *articles*, *pronouns* (except utterance initial subjective pronouns), *prepositions*, *conjunctions*, *auxiliary verbs*, *expletives*, and *particles*.

(M8) %FW: proportion of function words within the sentence

Expectation: Syllable-timed language < Stressed-timed language

[2] Or the other direction may be the case: i.e., the rhythm pattern of a language is formulated depending on the syllable structure complexity or vowel reduction rate.

Calculation formula: $\frac{\sum d_{function\,word}}{\sum d_{word}}$

Through perception tests of Russian and German, [30] shows that slower speech and more utterance internal pauses make listeners perceive more foreign accents. They calculated *speech rate* by counting *the number of phonemes per second* in their terms. In the current study, however, I propose the method of counting syllables per unit time in order to alleviate the difficulty in counting when a phoneme is severely reduced or virtually deleted phonetically, as is often the case.

(M9) Speech rate: Speed of utterance excluding internal pauses[3)]
Expectation: Syllable-timed language < Stressed-timed language
Calculation formula: Number of syllables / sec

It is quite often the case that language learners' hesitation is expressed in the form of short pause or silence. Thus, I propose to use the number of utterance-internal silences/pauses as a cue to non-native speech assessment.

(M10) Number of silence/pause intervals within the sentence
Expectation: Syllable-timed language < Stressed-timed language
Calculation formula: Number of silence/pause

To the best of my knowledge, the two novel metrics M8 and M10 have not been properly investigated with respect to automatic pronunciation scoring of non-native speech.

3. Verification of Metrics

In this section, the usefulness of ten metrics are verified to see whether they are useful to capture Korean speakers' native rhythm pattern carried over to their English utterances. Although hand-adjusted phonetic labels will be mainly used for accurate extraction of metrics, automatic methods of calculation will also be adopted for the purpose of later application for automatic scoring.

[3)] Also known as *articulation rate* in definition by [32].

3.1. Data

A total of 360 speech tokens read by 36 speakers are extracted from a corpus constructed to develop an auto-pronunciation assessment system [31]. Ten sentences are spoken by each speaker whose proficiency level falls on one of four proficiency levels: natives speakers (NS), advanced learners (AL), intermediate learners (IL) and lower-level learners (LL). The proficiency levels are pre-determined on the basis of each speaker's *years of English learning* and *years of stay in English speaking countries*. For parallel comparison, each group is designed to consist of nine speakers.

Data include utterance tokens of ten short sentences, as in <Table 2>, of which the number of words ranges from 4 to 13. Design of sentences in heterogeneous shapes are intended to find extra factors of speech rhythm that might be useful for characterization. Each sentence is made to include at least one function word in order to properly exhibit characteristics of the stress-timed rhythm. Instead of choosing sentences that exaggerate the English rhythm characteristics, relatively plain sentences that may appear in natural conversation are selected. Infrequent words are also avoided to prevent the (low-level) learners' hesitation, stammering or stuttering that may disrupt the valid metric extraction.

<Table 2> Data sentences

no	sentence	Number of		
		words	f-words	syllables
1	We took a vacation	4	1	6
2	Janice bought the Christmas present	5	1	8
3	Ironically, her husband is my friend	6	2	10
4	I didn't break the window deliberately	6	2	11
5	Jackson will testify against his brother	6	3	11
6	We couldn't stop the overflow from the tank	8	4	11
7	Cell phones are not permitted in the classroom	8	4	11
8	Fortunately, he didn't hurt his finger in the accident	9	4	16
9	She holds the world record for the one hundred meters	10	3	13
10	The firm needed to outlay a large amount of capital of the project	13	6	19

3.2. Labelling

Phonetic annotation, using modified TIMITbet[4] phone symbols, is conducted

[4] A keyboard-friendly phonetic alphabet system used in annotating the DARPA Acoustic

through two steps. First, automatic labels are generated through an automatic phone aligner which is basically a phone recognizer constructed by ASR techniques. The procedure to build this aligner can be summarized as: 41 three-state phoneme-like unit HMMs with 39-dimension acoustic features (MFCC, Energy and their first and second derivatives), trained with a native-speech database comprising 15935 tokens with their hand-demarcated phoneme boundary annotation files. As the aligner is trained with a separate speech database, its acoustic parameters have been re-estimated for enhanced performance using a portion of speech corpora from which the current data are also extracted.

After autolabels are obtained, they are hand-adjusted to produce a refined version of labels. Considering that automatic labels themselves are supposed to be used for auto-scoring, hand-adjustment is minimized by focusing on re-demarcating boundaries caused by unexpected errors such as filled pauses, stuttered words and non-linguistic noises.

While the metrics verification is based on the refined hand-adjusted labels, the autolabeling technique is directly applied to the autoscoring process which will be demonstrated in Section 4 later in this paper.

3.3. Analysis Method

Metrics are automatically calculated based on hand-adjusted annotation information. Then, the values are averaged over the tokens produced by nine speakers in the same proficiency group.

Whether each metric can discriminate proficiency groups or not is judged through correlation coefficients between proficiency level scores (i.e., 5, 3, 2 and 1 point for NS, AL, IL and LL, respectively) and corresponding speaker-averaged metric scores.[5] Subsequent statistical significance tests help interpret the results.

3.4. Results and discussion

<Table 3> presents each value of the rhythm metrics calculated by averaging 90 tokens (9 speakers x 10 sentences) for each proficiency group. According to one-way ANOVA tests, five metrics %V, VarcoV, nPVI-V, rate and Num-sil turn out to be

Phonetic Database which was collected by Texas Instruments and labelled by MIT [33].

[5] Native speakers' (NS) speech is given five points instead of four, due to a substantial gap between the groups NS and AL.

effective for distinguishing at least one proficiency level from the others (F(3, 356) = 4.61, 4.04, 7.16, 24.32, 45.65, respectively; $p<0.01$). Note that metrics with the diacritic † (i.e., ΔV, ΔC, VarcoC, rPVI-C) appear to be statistically significant but against expectation. For example, ΔV is expected to be greatest in native speakers' due to durational differences between stressed syllables and reduced syllables. However, the mean value of ΔV is the smallest in NS speech (45.2), increasing as the proficiency level gets lower. This phenomenon will be further discussed later in this section.

<Table 3> Overall rhythm metrics: obtained by averaging over all utterance tokens in each proficiency group; N=90 for each value. Rows for statistically significant metrics are shaded. The last column shows pairs that are distinguished from each other.

	Metric	Natives (NS)	Advanced (AL)	Intermediate (IL)	Low (LL)	Significant distinction
RIM	%V *	38.3 (5.4)	40.6 (5.4)	40.9 (5.9)	41.9 (5.7)	NS-LL
	ΔV †	45.2 (15.2)	47.4 (13.1)	52.4 (13.3)	54.7 (15.7)	
	ΔC †	36.0 (8.3)	40.0 (9.8)	42.5 (10.1)	45.3 (11.6)	
NIM	Varco-V *	54.2 (13.0)	49.1 (10.5)	49.9 (10.3)	48.4 (10.9)	NS-AL, NS-LL
	Varco-C	43.7 (9.5)	45.5 (8.8)	44.6 (8.6)	45.3 (8.8)	
PVI	nPVI-V *	59.2 (12.7)	51.2 (11.6)	54.44 (11.2)	51.3 (10.7)	NS-AL, NS-LL
	rPVI-C †	43.7 (10.3)	50.3 (12.6)	51.4 (15.0)	56.1 (19.4)	
AWM	%FW	19.2 (9.2)	20.9 (10.0)	22.2 (9.5)	22.6 (10.4)	
	rate *	4.7 (0.6)	4.3 (0.6)	3.9 (0.5)	3.7 (0.5)	all but IL-LL
	Num-sil *	0.4 (0.6)	0.4 (0.6)	0.9 (0.9)	1.5 (1.4)	all but NS-AL

The overall results only outline the general picture of each metric's discriminability. Although some metrics seem to be more effective than others, no categorical distinction is available to determine what metrics to accept or discard for automatic scoring. It is likely that a metric valid for one sentence may not work at all for another with different syntactic structure and/or vocabulary. For instance, the metric '%FW' will be especially effective for a sentence composed of a relatively large number of function words. On the other hand, VarcoC will work better at scoring sentences composed of words with consonant clusters. Consequently, it is necessary to analyze and validate metrics with respect to specific sentences.

<Table 4> shows the correlation between pre-defined proficiency levels and specific-sentence-based rhythm metrics. Each coefficient is calculated from 36 tokens for each sentence type with nine speakers for each of four proficiency levels. Cells are shaded to represent the statistical significance at the level of $p<0.01$ (*) or $p<0.001$ (**).

<Table 4> Correlation between rhythm metrics and learners' proficiency level: Shaded cells denote statistically significant correlations (either p<0.01* or p<0.001**). '+' and '-' signs below each metric name indicate whether the correlation is expected to be positive or negative. For cells with the marks † or † †, see descriptions in the text.

Class	Metrics	Sentence number									
		1	2	3	4	5	6	7	8	9	10
RIM	%V (-)	-0.16	-0.48 **	-0.57 **	-0.34 *	-0.27	-0.29 *	-0.12	-0.03	-0.22	-0.20
	ΔV (+)	-0.15	-0.37 † †	-0.47 † †	-0.13	-0.19	-0.42 † †	-0.29 †	0.31 *	-0.29 †	-0.31 †
	ΔC (+)	-0.23	-0.38 † †	-0.09	-0.54 † †	-0.40 † †	-0.20	-0.45 † †	-0.35 †	-0.56 † †	-0.25
NIM	VarcoV (+)	0.18	0.16	-0.28 †	0.57 **	0.28 *	-0.21	0.28 *	0.11	0.30 *	0.29
	VarcoC (+)	-0.16	-0.15	-0.05	-0.36 †	-0.18	0.10	0.01	-0.05	-0.03	0.27
PVI	nPVI-V (+)	0.08	0.07	0.01	0.34 *	0.37 **	0.30 *	0.23	0.15	0.04	0.45 † †
	rPVI-C (+)	-0.28 †	-0.41 † †	-0.19	-0.33 †	-0.37 † †	-0.17	-0.58 † †	0.35 *	-0.53 † †	-0.24
AWM	%FW (-)	-0.14	-0.37 **	-0.45 **	-0.05	-0.51 **	-0.57 **	-0.43 **	-0.20	-0.58 **	-0.54 **
	rate (+)	0.37 **	0.68 **	0.36 **	0.68 **	0.62 **	0.51 **	0.66 **	0.52 **	0.66 **	0.65 **
	Num-sil (-)	-0.16	-0.43 **	0.00	-0.41 **	-0.36 *	-0.56 **	-0.30 *	-0.58 **	-0.41 **	-0.53 **

Again, the cells with diacritics † and † † are not shaded even though their correlation coefficients are statistically significant. As mentioned above, this is because those figures represent the wrong direction of correlation. For example, the correlation between the proficiency level and ΔV for the sentence 2 is -0.37 and statistically significant († †, denoting p<0.001). However, it has been hypothesized that their correlation is supposed to be positive, as marked with '+' after the metric name, in the sense that learners with upper level proficiency will have greater vocalic interval variability in their utterances than lower proficiency learners who will be affected to a greater degree by the syllable-timed rhythm. The reason for the wrong direction of correlation is straightforward when we take the speech rate into account. The variability brought about by lower level learners' slow speech is severe enough to entirely mask the effect caused by the relevant metric. This speculation is further supported by the fact that the phenomenon is mostly relevant to the raw metrics without speech rate normalization (ΔV, ΔC, and rPVI-C), but not to the corresponding rate-normalized metrics (VarcoV, VarcoC). Consequently, the current results are in line with [29] which claims that control of speech rate in target sentences is mandatory when ΔC is to be used as a rhythm metric.

In general, metrics regarding vocalic intervals are more effective than consonantal interval measures. The roles of consonantal metrics such as ΔC, VarcoC and rPVI-C appear to be virtually vacuous. It is conjectured that acquiring consonantal phonotactics is not so difficult for Korean learners as vocalic variability in English. This result is in agreement with [34] suggesting that "... deficits of speech rhythm competence in a second language are mostly revealed on a vocalic level." Otherwise, consonantal characteristics may not crucial at all in rhythm classification of at least some languages including Czech, as has recently been reported by [35].

The negligible role of ΔV may also be attributed to the speech rate effect as has been indicated by [29]. Since non-native speech tends to be produced in a slower rate than native speech, standard deviation of vocalic, as well as of consonantal, intervals will increase extensively, preempting the effect of non-native rhythm. This leads to a speculation that the speech rate itself can be used as a reliable measure for assessing non-native prosody, as is verified in the present experiment. On the contrary, another RIM metric %V is found to be useful as it is known to be fairly free from speech rate effects, which also agrees with Dellwo and Wagner's [29] observation.

As clearly shown in <Table 4>, The AWM metrics are more consistent in discriminating proficiency levels than the other metrics. Especially, the 'rate' is effective for all ten sentences including the shortest four-word sentence for which no other metrics are operative. The number of silences/pauses within the utterance is also a reliable measure capturing degrees of learners' hesitation and disfluency. It should be mentioned that this feature is not directly associated with speech rate, as interval durations of within-utterance silences are excluded from calculation of speech rate.

To summarize, AWM metrics are robust enough to be used in any utterances but too short sentences. Metrics for vocalic intervals are more effective than those for consonantal metrics. Besides, different metrics need to be selected for scoring sentences of different types while speech rate appears to be a universal measure for non-native rhythm proficiency.

4. Autoscoring of Rhythm

Based on the correlation analysis above, an experiment of automatic scoring has been performed. Considering the relatively small size of data used, this experiment is not so much a verification of metrics for automatic rhythm scoring as a demonstration of a method by which metrics can be applied to practical auto-scoring tasks.

4.1 Procedures

When an input test speech is provided, the automatic aligner (described in Section 3.2) demarcates the phone and word boundaries. Subsequently, metrics that have been determined to be useful for the rhythm characterization are extracted sentence-specifically. For example, if the input utterance is Sentence 5 (*Jackson will testify against his brother*), each value for metrics such as VarcoV, nPVI-V, %FW, rate, and Num-sil is calculated. Sentence specific features are selected based upon the verification results described in the previous section. That is, features that are found to play a significant part in distinction between proficiency groups are selected for each sentence. The shaded cells in <Table 4> denotes the useful features.

The key procedure is to generate one overall rhythm score by combining individual metric scores after assigning an optimal weight to each feature. This procedure is performed by the statistical technique named Multiple Regression which can be represented as:

$$rhythm\,score = \beta_0 + \beta_1 fea_1 + \beta_2 fea_2 + \ldots + \beta_n fea_n \qquad (1)$$

where β_0 denotes the intercept, and β_n is the weight assigned to n^{th} feature.

When features extracted from training data are provided, regression coefficients (*β's)* are estimated in terms of the linear least squares function. Once the equation is established, an overall score for each test token can be generated. This technique has already been employed in automatic assessment of essay writing tasks [36][37] and begun to be applied to wider areas including automatic scoring of non-native speech [38].

The 360 tokens used in the previous experiment are randomly broken into training tokens (240) and test tokens (120). Training tokens are provided for estimation of regression coefficients as well as the intercept value (β_0), while the test set are for verification of auto-scoring performance, whose result follows in the next section.

4.2. Results and discussion

The overall correlation coefficient between automatic scores and pre-defined proficiency levels is 0.368. Sentence specific correlation coefficients are presented in

<Table 5>.

<Table 5> Sentence specific correlation between autoscores and pre-defined proficiency levels (*p<0.05, **p<0.01)

	Sentence number									
	1	2	3	4	5	6	7	8	9	10
r	0.47	0.56	0.50	0.77	0.43	0.20	0.49	0.37	0.70	0.52
p	0.062	0.028	0.048	0.002	0.084	0.270	0.055	0.117	0.006	0.043
significance		*	*	**					**	*

Although the overall correlation is statistically significant (p<0.001) and some sentences are found to be suitable for autoscoring, the results in general fall short of expectations.

The mean scores for each proficiency level are: LL (2.98), IL (2.94), AL (3.61), and NS (4.22). A one-way ANOVA and a subsequent Post Hoc test (the Tukey's test) reveal that differences between groups (F(3, 116)=8.02, p<0.001) do exist. However, the significant distance is confirmed only between groups: LL-NS and IL-NS. Especially, lower groups (LL and IL) do not appear to be distinguishable from each other. The reasons for less-than-expected discriminability are many folds.

First, autosegmentation errors are quite likely to have hampered reliable feature extraction. As mentioned before, autosegmentation performance by the aligner directly influences the autoscoring reliability as most metrics are extracted on the basis of phoneme or word boundary information. However, it does not mean that the aligner currently used is poor in quality. Rather, more elaborate input control seems to be necessary. Many speech tokens by learners are quite frequently filled with unexpected non-speech parts within an utterance, including stuttering, hesitation, meaningless interjection. This tendency is more severe in speech by lower level learners, as they are more inclined to struggle to pronounce words they encounter. These errors will cause disruption of autosegmentation, unless separate sophisticated techniques are implemented to track down those errors before the segmentation module operates. Although a few methods of resolving this problem, including producing instant request for repetition when a duration or acoustic quality of speech token is above designated threshold, more fundamental solution in signal processing stage will be necessary.

Second, the validity of pre-defined proficiency levels need to be questioned. Determined by relatively crude external information on *duration of stay in English speaking countries* along with *duration of English learning*, they do not seem to most appropriately represent the optimal target scores autoscorer aims to achieve. The most practical alternative reference scores might be obtained by human scorers by expert EFL educators, as adopted in recent autoscoring experiments [39][40]. Nevertheless, it is still pretty much an open question how to objectively evaluate auto scoring performance.

Finally, data sparsity could have kept optimal regression coefficients for each metric from being estimated. Only a total of 240 training tokens are used for Multiple Regression modelling. For each sentence type, features from only 24 (6 speakers x 4 levels) tokens are counted in. This is obviously too little amount of data from which to build robust acoustic models, even before taking into account high variability of prosodic parameters and language learners' diverse pronunciation.

Despite prematurity in modeling, autoscoring seems to be useful at least for discriminating learners' rhythm manifestation from the native speakers'. However, further experimental investigations with more data and elaborate design will be necessary in order to clarify the behavior of individual metrics and to make the prosody autoscorer fit for practical use.

5. Summary and Future Directions

The traditional vowel metrics, such as %V, VarcoV, and nPVI-V and new metrics, such as %FW, Number of silences, Speech Rate, are approved as useful cues in one way or another for distinguishing the speech rhythm of Korean EFL learners from that of native speakers of English. On the other hand, consonantal metrics do not seem to carry valid information for capturing rhythmic deficit of non-native speech.

Autoscores obtained from dynamically selected metrics are significantly, but coarsely, correlated with learners' English proficiency levels derived from their English learning experiences.

In order to use those verified metrics in practical pronunciation education systems, improvements in various aspects are desirable including: robust modeling with more first and second language data, developing efficient scoring algorithm, and improving performance of phoneme boundary aligners or ASR in general.

Though duration metrics are robust, practical, and easy to extract, they are not exhaustive in any sense for characterizing speech rhythm. Sonority, as in [41], has already been suggested to be a reliable feature for non-native rhythm scoring. Various other characteristics based on F0 and/or intensity analyses also need to be employed after through phonetic and technical verifications.

Acknowledgement

I am grateful to Professor Jan van Santen at Oregon Graduate Institute for precious comments and suggestions. I also appreciate three anonymous reviewers for their appropriate critiques.

References

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

[2] D. Abercrombie, *Elements of General Phonetics,* Edinburgh: Edinburgh University Press, 1967.

[3] G. Pointon. "Is Spanish really syllable-timed?", *Journal of Phonetics,* Vol. 8, pp. 293-305, 1980.

[4] B. J. Wenk, F. Wioland, "Is French really syllable-timed?", *Journal of Phonetics,* Vol. 10, pp. 193-216, 1982.

[5] D. L. Bolinger, "Pitch accent and sentence rhythm", In Abe Isamu & Tetsuya Kanekiyo (eds.), *Forms of English: Accent, Morpheme, Order,* Havard University Press, 1965.

[6] W. L. Lea, "Prosodic aids to speech recognition: IV--A general strategy for prosodically-guided speech understanding", *Univac Report* PX10791, St. Paul, Minnesota: Sperry Univac, 1974.

[7] J. D. O'Connor, "Perception of time intervals", *Progress Report,* Vol. 2, Phonetics Laboratory, University College London, pp. 11-15, 1965.

[8] Y. Shen, G. G. Peterson, "Isochronism in English." *Occasional Papers,* Vol. 9, University of Buffalo, Studies in Linguistics, pp. 1-36, 1962.

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

[10] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, C. Amiel-Tison, "A precursor of language acquisition in young infants", *Cognition,* Vol. 29, pp. 143-178, 1998.

[11] T. Nazzi, J. Bertoncini, J. Mehler, "Language discrimination by newborns: toward an understanding of the role of rhythm", *Journal of Experimental Psychology,* Vol. 24, No. 3, pp. 756-766. 1998.

[12] F. Ramus, J. Mehler "Language identification with suprasegmental cues: a study based on speech resynthesis", *Journal of the Acoustical Society of America,* Vol. 105, No. 1, pp. 512-521. 1999.

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

[14] V. Dellwo, “Rhythm and speech rate: A variation coefficient for delta C”, in P. Karnowski and I. Szigeti (eds.), *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium*, pp. 231-241, Piliscsaba 2003. Frankfurt: Peter Lang, 2006.

[15] L. White, S. L. Mattys, “Calibrating rhythm: first language and second language studies”, *Journal of Phonetics* Vol. 35, pp. 501-522. 2007.

[16] E. Grabe, E. L. Low, “Durational variability in speech and the rhythm class hypothesis”, *Papers in Laboratory Phonology* Vol. 7. pp. 515-546, Berlin: Mouton, 2002.

[17] E. L. Low, E. Grabe, F. Nolan, “Quantitative characterisations of speech rhythm: ‘syllable-timing’ in Singapore English”, *Language and Speech* Vol. 43, pp. 377-401, 2000.

[18] M. Han, “The feature of duration of Japanese”, *Onsei no Kenkyu (Study of Sounds)*, Vol. 10, pp. 65-75, 1962.

[19] C. Hoequist Jr., “Syllable duration in stress-, syllable-, and mora-timed languages”, *Phonetica* Vol. 40, pp. 203-237, 1983.

[20] R. F. Port, J. Dalby, M. O’Dell, “Evidence for mora timing in Japanese”, *Journal of the Acoustical Society of America* Vol. 81, pp. 1574-1564, 1987.

[21] K. Tajima, *Speech Rhythm in English and Japanese: Experiments in Speech Cycling*, PhD dissertation, Indiana University, 1998.

[22] V. Stockmal, D. Markus, D. Bond, “Measures of native and non-native rhythm in a quantity language”, *Language & speech* Vol. 48. pp. 55-63, 2005.

[23] W. A. van Dommelen, “Quantification of speech rhythm in Norwegian as a second Language”, in *Centre for Languages & Literature Working Papers* Vol. 52, pp. 33-36, Lund University, 2006.

[24] E. L. Asu, F. Nolan. Estonian rhythm and the Pairwise Variability Index. *Proceedings of Fonetik 2005: The XVIII’th Swedish Phonetics Conference*, pp. 29-32, Gothenburg, 2005.

[25] O.-H. Lee, J.-M. Kim, “Syllable-timing interferes with Korean learners’ speech of stress-timed English”, *Speech Sciences* Vol. 12, No. 4, pp. 95-112, The Korean Association of Speech Sciences, 2005.

[26] J.-M. Kim, S. Flynn, M. Oh, “Non-native speech rhythm: a large-scale study of English pronunciation by Korean learners”. *Studies in Phonetics, Phonology and Morphology* Vol. 13, No. 2, pp. 245-275, 2007.

[27] Hyun-Bok Lee, “A phonetic study on Korean rhythm”, *Malsori* Vol. 4, pp. 31-48 (in Korean), 1982.

[28] W. J. Barry, B. Andreeva, M. Russo, S. Dimitrova, T. Kostadinova, “Do rhythm measures tell us anything about language type?” *Proceedings of the 15th international congress of phonetics sciences*, pp. 2693-2696, Barcelona, 2003.

[29] V. Dellwo, P. Wagner, “Relations between language rhythm and speech rate”, *Proceedings of the 15th international congress of phonetics sciences*, pp. 471-474, Barcelona, 2003.

[30] B. Wendt, I. Bose, M. Sailer, H. Scheich, H. Ackermann, “Speech rhythm of a woman

with foreign accent syndrome (FAS)", *Proceedings of International Congress on Phonetic Science*, pp. 2009-2012, Saarbrücken, Germany, 2007.

[31] H. Chung, T.-Y. Jang, W. Yun, I. Yun, J. Sa. "A study on automatic measurement of pronunciation accuracy of English speech produced by Korean learners of English", to appear, *Language and Linguistics*, Vol. 42, 2008.

[32] T. H. Crystal, A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech", *Journal of the Acoustical Society of America* Vol. 88, No. 1, pp. 101-112, 1990.

[33] S. Seneff, V. W. Zue, "Transcription and alignment of the TIMIT database", DARPA TIMIT CD-ROM Documentation, 1988.

[34] F. G. Diez, V. Dellwo, N. Gavaldà, S. Rosen, "The development of measurable speech rhythm during second language acquisition", *Journal of Acoustical Society of America* Vol. 123, No. 5, Pt 2, p. 3886 (Abstract), May 2008.

[35] J. Dankovicovâ, V. Dellwo, "Czech speech rhythm and the rhythm class hypothesis", *Proceedings of International Congress on Phonetic Science*, pp. 1241-1244, Saarbrücken, Germany, 2007.

[36] E. B. Page, "Computer grading of student prose, using modern concepts and software." *Journal of Experimental Education* Vol. 62, No. 2, pp. 127-142, 1994.

[37] E. B. Page, N. S. Petersen, "The computer moves into essay grading: Updating the ancient test", *Phi Delta Kappan* Vol. 76, pp. 561-565, (first developed in mid-1960s by Ellis Page), 1995.

[38] K. D. Zechner, D. Higgins, and X. Xiaoming, "SpeechRaterTM: a construct-driven approach to score spontaneous non-native speech", *Proceedings of the 2007 Workshop of the International Speech Communication Association (ISCA) Special Interest Group meeting on Speech and Language Technology in Education (SLaTE)*, pp. 128-131. Farmington, PA, October. 2007.

[39] S. M. Witt, S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech Communication* Vol. 30, pp. 95-108, 2000.

[40] L. Neumeyer, H. Franco, V. Digalakis, M. Weintraub, "Automatic scoring of pronunciation quality", *Speech Communication* Vol. 30, pp. 83-93.

[41] A. Galves, D. Duarte, J. Garcia, C. Galves, "Sonority as a basis for rhythmic class discrimination", *Proceedings of the first International Conference on Speech Prosody*, pp. 323-326, Laboratoire Parole et Langage, Aix-en-Provence, 2002.

접수일자: 2008년 5월 23일
게재결정: 2008년 6월 9일

▶ 장태엽(Tae-Yeoub Jang)
주소: 130-791 서울특별시 동대문구 이문동 270번지
소속: 한국외국어대학교 영어학과
전화: 02) 2173 3119
E-mail: tae@hufs.ac.kr