Applying Rhythm Metrics to Non-native Spontaneous Speech

Catherine Lai[1], *Keelan Evanini*[2], *Klaus Zechner*[2]

[1]School of Informatics, University of Edinburgh, Scotland, UK
[2]Educational Testing Service, Princeton, NJ, USA
clai@inf.ed.ac.uk, kevanini@ets.org, kzechner@ets.org

Abstract

This study investigates a variety of rhythm metrics on two corpora of non-native spontaneous speech and compares the non-native distributions to values from a corpus of native speech. Several of the metrics are shown to differentiate well between native and non-native speakers and to also have moderate correlations with English proficiency scores that were assigned to the non-native speech. The metric that had the highest correlation with English proficiency scores (apart from speaking rate) was rPVIsyl (the raw Pairwise Variability Index for syllables), with $r = -0.43$.

Index Terms: Rhythm metrics, non-native speech, fluency

1. Introduction

Various studies have investigated metrics for quantifying rhythmic differences between languages based on properties of segmental durations [1, 2, 3]. These metrics present summary statistics about the variation of consonantal and vocalic durations throughout utterances. For example, [1] consider the percentage of the speech that was vocalic (%V) versus the standard deviation of consonantal intervals (ΔC, similarly ΔV for vocalic segments) in a sentence, while [2] employ a Pairwise Variability Index to quantify differences in adjacent intervals. [3] propose normalization of ΔC and ΔV by dividing this measure by the mean durations of consonantal and vocalic intervals respectively (VarcoC, VarcoV).

The main motivation behind the development of these measures has been to investigate the idea that languages fall into discrete rhythm classes (i.e. syllable- vs. stress-timed). However, experimental evidence has cast doubt on the idea that cross-linguistic differences in these measures represent categorical 'rhythm class' differences [4, 5]. In general, it seems that languages may vary on several rhythmic dimensions. [4] examine how well these metrics discriminate five different languages using a large corpus of read data. They find that, while pairs of languages can be discriminated, no pair of metrics separates all pairs of languages. Similarly, [5] find speech rate to be a dominant factor in the perception of language differences, but still find that listeners can discriminate the original language of utterances with identical segmental and intonational content even when speech rate is controlled for.

In addition to categorizing speech from different languages with respect to their rhytmic properties, these types of rhythm metrics have also been used to measure the rhythmic closeness of non-native speech to a given target language. Recent studies have found some of these rhythm metrics, especially ones measuring vocalic durations, to be useful in characterizing differences between native and non-native speech. For example, [6] found VarcoV and %V to be the most discriminating with respect to L1 and L2 (English-Spanish, English-Dutch). Moreover, they found that the VarcoV of English from native Spanish speakers had an intermediate value between the values of native English and Spanish speech (and similarly for English speakers of L2 Spanish). This suggests that these metrics can be used to quantify the effect of L1 on non-native productions. [7] investigated several different rhythm metrics on Cantonese-accented English and Mandarin-accented English and found that some of the metrics grouped the non-native English speakers with English speakers whereas other metrics grouped them with Mandarin and Cantonese speakers. [8] attempted to discriminate L1 British English speakers and two classes of L2 English (L1 French) speakers using scores from all metrics. They also suggest that %V and VarcoC give the best discrimination between the L1 classes. Their best reported SVM-based classification score used %V, ΔV, VarcoV and nPVI-V and achieved a 67% accuracy rate. Finally, [9] defined a new rhythm measure, the Pairwise Variability Error, and used it, along with several standard rhythm measures, to classify native English speech and Japanese-accented English. They found that the Pairwise Variability Error was the single best-performing rhythm measure, with a classification accuracy of 69.4%.

In addition to studies that have discriminated between native and non-native speech using rhythm measures, some previous studies have also used rhythm measures to score the proficiency of non-native speech. [10] used several standard rhythm measures (in combination with additional features) to predict English proficiency scores for a set of Korean English learners. [11] studied the differences between Spanish-accented English and native English with regard to phrasal prominence. They defined a rhythm measure based on the differences in mean vowel durations for syllables with primary stress and secondary stress and found that this measure correlated with phrasal prominence scores at a rate of 0.683. Finally, [12] studied read aloud English produced by native Mandarin speakers and found that rhythm metrics improved the prediction accuracy of holistic English proficiency scores when they were added to regression models built on features assessing fluency, pronunciation, and reading accuracy. The vocalic measures had the best correlations with human scores; furthermore, the correlations were negative, which indicates that a higher relative percentage of vocalic intervals in the non-native speech led to lower scores.

The methodology of using rhythm features to evaluate a non-native speaker's speaking proficiency in this study is similar to the approaches taken in these previous studies. However, most previous studies were based on restricted speech (read aloud or repeat aloud tasks) and only included non-native speakers from a single language background. In the current work we examine what these measures can tell us about spontaneous speech from speakers representing a large range of L1 backgrounds. This study will thus provide more direct evidence of the usefulness of rhythm measures for assessing a non-native speaker's communicative competence in English, since a more naturalistic speaking task is examined. Furthermore, the inclusion of speakers from many different L1 backgrounds in this study means that conclusions about more general test taker populations can be drawn.

This paper is organized as follows: first, Section 2 describes the data sets and the methodology used for extracting the rhythm metrics; Section 3 presents the results of the experiments as follows: Section 3.1 shows how the individual rhythm metrics are correlated with proficiency ratings in non-native speech, Section 3.2 compares the rhythm metrics for non-native speakers to native speakers, Section 3.3 investigates the robustness of the syllable-level rhythm metrics by comparing two different non-native corpora to native speech, and Section 3.4 examines how the metrics vary based on the L1 of non-native speakers; finally, Section 4 summarizes the main contributions of this paper and discusses directions for future research.

2. Data and Methodology

In this study, we examined spoken responses from three data sets related to the TOEFL iBT assessment, an international assessment of English proficiency. The non-native speech sets were drawn from two sources: 1) responses to the TOEFL Practice Online assessment (henceforth TPO) and 2) responses to the TOEFL Academic Speaking Test (henceforth TAST). The native speech was drawn from a study in which native English speakers responded to TOEFL test questions in a laboratory setting (henceforth TOEFL-NS). In all of these data sets, the speakers responded to open-ended prompts that elicited spontaneous speech on a variety of topics. All of the responses are either 45 or 60 seconds in duration. Table 1 summarizes the sizes of these three data sets by listing the number of responses and speakers contained in each, as well as the number of different L1s represented (for the two non-native data sets).

Data set	# Responses	# Speakers	# L1s
TPO	1019	239	50
TAST	87	60	23
TOEFL-NS	182	34	N/A

Table 1: *Summary of the three data sets used in the study*

The TPO responses were each subsequently provided with holistic scores of English proficiency by two independent, expert raters using scoring rubrics that reflected a speaker's delivery (fluency, pronunciation, intonation, etc.), language use (vocabulary, grammatical accuracy, etc.), and content appropriateness. The raters gave each response a proficiency score on a scale of 1 - 4, with 4 indicating the highest proficiency level.

Table 2 summarizes the rhythm metrics that are investigated in this section. The metrics were calculated over consonantal (C), vocalic (V) and syllabic (syl) intervals except for speech rate, which is defined only in terms of syllables. In addition, we use the proportion of utterance medial silence as a feature (%sp); this metric is equivalent to the inverse of the proportion of syllabic intervals in a response (i.e., %sp = 1 - %syl). Phone boundaries were derived automatically using the Penn Phonetics Lab Forced Aligner [13] on manual transcriptions of the spoken responses. The phones were grouped into syllables using a rule-based, onset maximization approach.[1] Disfluencies (such as filled pauses) were not removed from the data sets before the calculation of rhythm measures.

Metric	Description
ΔX	Standard deviation of X intervals.
%X	Percentage of X speech.
VarcoX	ΔX × 100/ mean(X)
nPVI-X	Normalized Pairwise Variability Index:
	$100 \times \sum_{k=1}^{n-1} \lvert x_{k+1} - x_k/(x_{k+1} + x_k/2)\rvert/n - 1$
rPVI-X	raw PVI.
	$\sum_{k=1}^{n-1} \lvert x_{k+1} - x_k\rvert/n - 1$
srate	Syllables per second.

Table 2: *Summary of rhythm metrics. We calculate these measures over V=Vocalic and C=consonantal intervals, as well as Syl=syllables.*

3. Results

3.1. Rhythm Metrics and Proficiency Scores

Metric	r1	r2
ΔV	-0.22	-0.20
ΔC	-0.18	-0.17
Δsyl	-0.27	-0.24
%V	-0.20	-0.26
%C	0.20	0.26
VarcoV	*n.s.*	*n.s.*
VarcoC	-0.15	-0.16
Varcosyl	-0.16	-0.11
nPVIV	*n.s.*	*n.s.*
rPVIC	-0.30	-0.25
nPVIsyl	-0.22	-0.16
rPVIsyl	-0.44	-0.36
%sp	-0.38	-0.27
srate	0.41	0.40

Table 3: *Correlation with scores (r1, r2) on the TPO data set (N=1019)*

Table 3 shows the Pearson correlations between the various rhythm metrics and the holistic English proficiency scores from two different raters (r1 and r2) for the TPO data set (all correlations reported in the table are significant at $p < 0.05$). The C- and V-based metrics that correlated best with the human scores were rPVIC, ΔV and %V. This indicates that a greater amount of variability in segment lengths or a greater proportion of vocalic speech was associated with lower scores. However the correlations associated with these metrics are relatively low compared to those associated with the syllable-based metrics. In particular, speaking rate (r=(0.41, 0.40)) and rPVIsyl (r=(-0.44, -0.37)). This level of correlation is not too far off the inter-rater correlation for this data set (r=0.50). This suggests that lower proficiency scores correlate with slower speech and with greater duration changes from syllable to syllable. Note, these two metrics are also correlated (r=-0.49). The proportion of the utterance medial silence, %sp, was also quite highly correlated with the pronunciation scores. We can take this to be a more traditional measure of fluency.

All of the metrics were significantly correlated with the human scores ($p < 0.001$) except the normalized vowel measures:

[1] `https://p2tk.svn.sourceforge.net/svnroot/p2tk/python/syllabify/syllabifier.py`

VarcoV and nPVIV. This is somewhat unexpected given that [6] found VarcoV to be the most useful metric for discriminating L2 speech (Spanish, English). This may be due to the larger range of L1's associated with the L2 speech in this data set. However, the fact that neither of the two rate-normalized vowel measures correlated with the scores suggests that the rhythmic information provided by these metrics is dominated by influence of speaking rate on the pronunciation scores. So, while these metrics may highlight L1 features of L2 speech, this may not actually be very important for how fluent the speech is perceived to be by human raters.

%sp appears to be more or less independent of the other rhythm metrics: it was only significantly correlated with ΔV, VarcoV, and ΔSyl and all correlations were reasonably low. Interestingly, %sp did not have as high correlations with the scores as speaking rate. All metrics were significantly correlated with speaking rate except Varcosyl. Given that speaking rate had a much higher correlation with human scores, it appears that segment level rhythm measures are not so useful for automated scoring. However, they may still be useful for understanding the underlying rhythm differences between different L2 speakers.

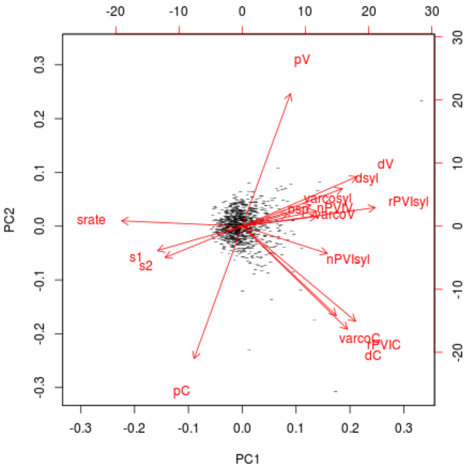
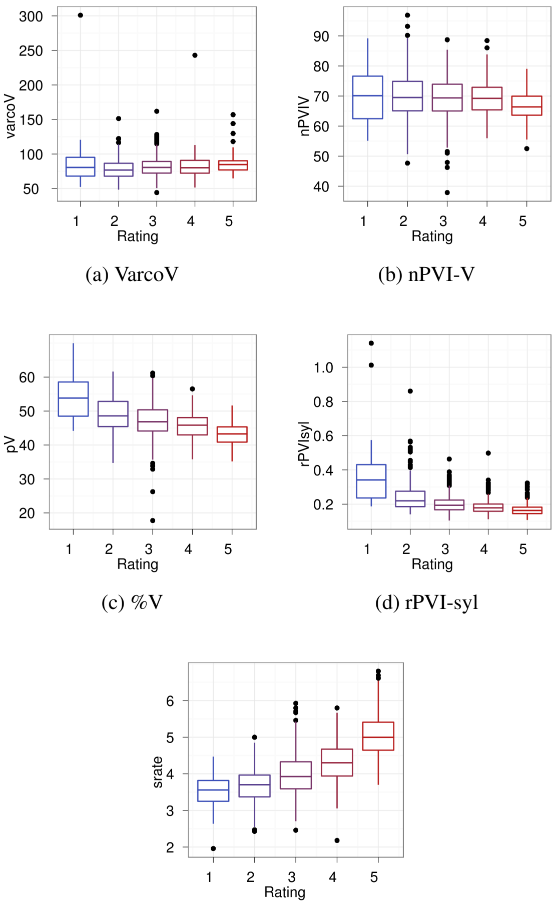


Figure 1: *Biplot of TPO rhythm metrics: rhythm data projected onto the first two components from a PCA.*

Principal components analysis on this data give us more information about the relationship between these measures. Figure 1 shows rhythm data projected onto the first two components from a Principal Components Analysis. The red arrows represent the original metrics with respect to these components. We see that syllable and vowel based measures cluster together and point in the opposite direction to the scores. Consonantal measures, on the otherhand, appear orthogonal to the score vectors. %V appears independent of the other vowel measures.

3.2. Comparison to native speech

In order to investigate how the correlations between rhythm metrics and pronunciation scores relate to differences between native and non-native speech, we also applied the rhythm measures to the TOEFL native speaker forced alignment data. Figure 2 shows how the rhythm metrics differ for different score groups. In these figures, human scores from the first set of raters (r1 in Table 3) are used for the TPO responses, and the TOEFL-NS responses were each given an arbitrary rating of 5



(a) VarcoV

(b) nPVI-V

(c) %V

(d) rPVI-syl

(e) srate

Figure 2: *TPO and TOEFL-NS data: TPO data is rated between 1-4. TOEFL-NS data is rated 5 to highlight the differences between corpora.*

to highlight the differences between the two corpora. We previously saw that neither VarcoV nor nPVI-V were significantly correlated with TPO scores. In Figure 2a we see that VarcoV (standard deviation/mean vowel duration) doesn't appear to really differ across scores or corpora. The similarity in the means for VarcoV seems to reflect the fact that what mattered for the scoring was not the variance in vocalic intervals, but rather the fact that the less fluent speakers spoke slower. In Figure 2b, we can see that nPVI-V distributions are not significantly different across the TPO data. However, the nPVI-V values for TOEFL-NS are significantly lower. This suggests that adjacent vocalic durations are more similar in L1 English, but this was a rhythmic factor not captured by the non-native speakers in this task. Figures 2c-2e confirm that %V, rPVIsyl, and speaking rate are good indicators of closeness to L1 English: L1 English speakers have a lower percentage of vocalic segments, lower syllable-to-syllable duration differences and a faster speaking rate.

In order to compare the performance of all of the features across the two data sets, Figure 3 shows the correlations between the various metrics and speaking rate.[2] Looking at the

[2] In the figures, the symbol Δ in the rhythm metrics is represented by *d*, and % is represented by *p*; for example, dC represents ΔC and

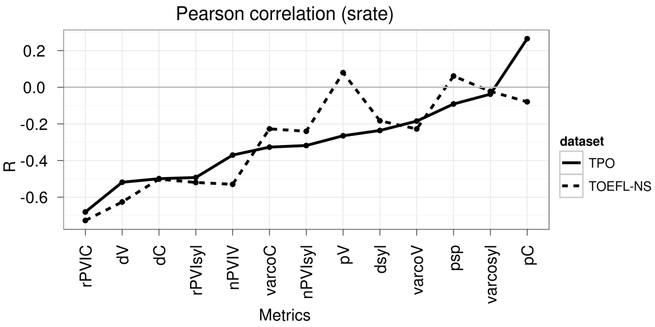
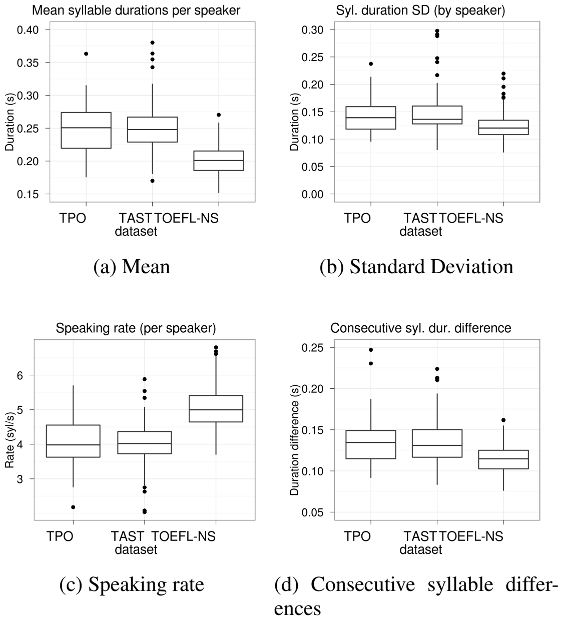


Figure 3: *Correlation with srate: TPO (non-native) and TOEFL-NS (native speakers) data sets*

correlations, we note that, unlike for the TPO data, %V was not significantly correlated with speaking rate in the native speaker data again suggesting that %V reflects a different aspect of speaker competence. The negative correlation with %V may reflect the presence of more filled pauses in the lower scored speech, which tend to have an extended duration.

3.3. Native/Non-native Syllable level differences



(a) Mean

(b) Standard Deviation

(c) Speaking rate

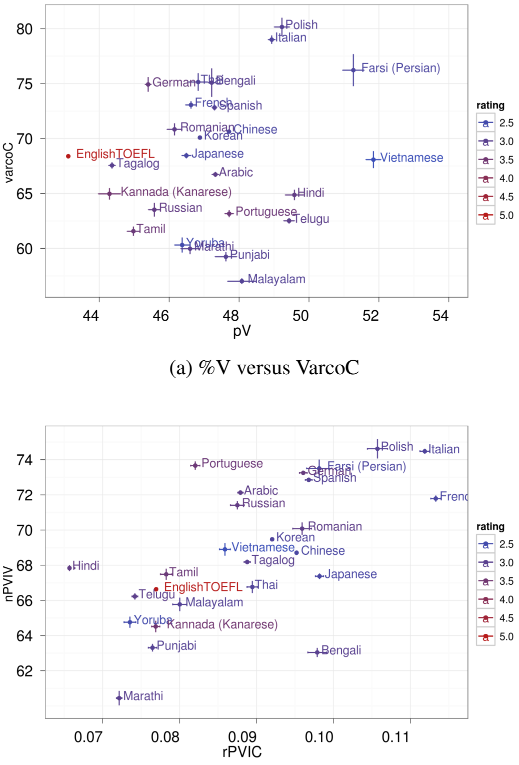
(d) Consecutive syllable differences

Figure 4: *Duration differences across corpora*

In the results reported above, syllable level measures had the highest correlation with pronunciation scores, particularly speaking rate and syllable to syllable variability. In order to test the robustness of these features across different data sets of non-native spontaneous speech, we compared timing data between the TPO and TAST data sets.

Figure 4d shows mean syllable duration, standard deviations of syllable durations, speaking rate, and mean syllable-to-syllable differences, calculated for each speaker for the TAST, TPO and TOEFL-NS data sets. We see significant differences for these features between the native and non-native sets, but not between the non-native sets. As expected from the TPO results, the graphs show that non-native speakers speak slower, in terms of syllables per second, and have more variable syllable durations (excluding short pauses) in both the non-native speech corpora. So, it seems these relatively high level durational features are useful in the broad classification of native versus non-native speech.

3.4. The Effect of L1



(a) %V versus VarcoC

(b) rPVI-C versus nPVI-V

Figure 5: *L1 differences: Means and standard errors of the means in the TPO and TOEFL-NS data sets. The color scale represents mean proficiency scores for each language.*

Studies such as [1, 2] have suggested that rhythm metrics reflect consistent differences between L1 speech, as well as between L1 and L2 speech [6], based on read speech. To see how well these findings extend to spontaneous speech, we examined the distribution of rhythm measures values by L1. Figure 5 shows means for different L1 language groups in the TPO data set (for languages with more than 10 samples). Figure 5a mirrors the approach of [1] (%V versus VarcoC), while Figure 5b reflects that of [2] (nPVI-V versus rPVI-C). Neither graphs matches the previous studies in the terms of the ordering of languages. For example, English (typically categorized as a *stress timed* language) is expected to have a higher nPVI-V score than the Romance languages French and Spanish (which are categorized as *syllable timed*). Similarly, [6] find Spanish speakers of English to have lower VarcoV than native speakers, which we do not find in our data. This casts doubt on the stability of these metrics on different corpora, e.g., when looking at

psp represents %sp.

read vs. spontaneous speech. Additionally, recent studies have called into question the notion that languages can be clearly differentiated based on rhythm metrics, since the degree of inter-speaker variability in these metrics for a single language is often similar in magnitude to the degree of variability between languages [4, 14].

Nevertheless, we do observe that the L2 speech roughly groups around language families: for example, French, Spanish, Portuguese and Romanian are relatively close. So, while these measures do not produce the same topology as the original native speaker studies, they still may be useful for characterising L1 transfer effects from the different language families.

4. Conclusions

In this study we investigated a variety of rhythm metrics in two corpora of non-native spontaneous speech and compared their distributions to a corpus of native speech. Several of the metrics resulted in large group-level differences between native and non-native speakers and also showed moderate correlations with holistic proficiency scores assigned to the non-native spoken responses. These two findings indicate that these types of metrics should be incorporated into applications that provide automated assessments of spontaneous spoken English, such as [15], in addition to the more commonly used fluency and pronunciation features.

In prior studies, duration based measures were shown to be useful for distinguishing native and non-native speech in terms of fluency, e.g. lower pause durations and higher speaking rate correlate with higher pronunciation scores [16, 17, 18]. This study replicated this finding by demonstrating that the *srate* feature had the highest correlation with proficiency scores among all of the rhythm metrics. However, this study also demonstrated that several additional segment-based rhythm metrics have significant correlations with proficiency scores with a magnitude close to the *srate* feature. This finding indicates that these rhythm metrics can be useful additional indicators of a non-native speaker's fluency in spontaneous speech.

Future research will integrate these rhythm features into a system for automated assessment of non-native speech in order to see how much of an impact these features can have on the prediction of holistic English proficiency scores. In addition, we will investigate the distributions of the rhythm features on a data set that contains both spontaneous and restricted speech in order to determine whether these two speaking styles have different rhythmic characteristics in non-native speech.

5. Acknowledgments

The authors would like to thank Su-Youn Yoon, Lei Chen, Anastassia Loukina, and three anonymous reviewers for their feedback about this paper.

6. References

[1] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 75, no. 1, 2000.

[2] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515—546, 2002.

[3] V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC," *Language and language-processing*, pp. 231–241, 2006.

[4] A. Loukina, G. Kochanski, B. Rosner, and E. Keane, "Rhythm measures and dimensions of durational variation in speech," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3258–3270, 2011.

[5] L. White, S. L. Mattys, and L. Wiget, "Language categorization by adults is based on sensitivity to durational cues, not rhythm class," *Journal of Memory and Language*, vol. 66, no. 4, 2012.

[6] L. White and S. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.

[7] P. Mok and V. Dellwo, "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English," in *Proceedings of Speech Prosody*, 2008.

[8] A. Tortel and D. Hirst, "Rhythm metrics and the production of English L1/L2," in *Proceedings of Speech Prosody*, 2010.

[9] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental English through parroting," in *Proceedings of Speech Prosody*, 2010.

[10] T.-Y. Jang, "Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers English pronunciation," in *Proceedings of the 2nd International Conference on East Asian Linguistics*, 2009.

[11] E. Nava, J. Tepperman, L. Goldstein, M. Zubizarreta, and S. Narayanan, "Connecting rhythm and prominence in automatic ESL pronunciation scoring," in *Proceedings of Interspeech*, 2009.

[12] L. Chen and K. Zechner, "Applying rhythm features to automatically assess non-native speech," in *Proceedings of Interspeech*, 2011.

[13] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

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

[15] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[16] J. Liscombe, "Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency," Ph.D. dissertation, Columbia University, 2007.

[17] J. Yuan, Y. Jiang, and Z. Song, "Perception of foreign accent in spontaneous L2 English speech," in *Proceedings of Speech Prosody*, 2010.

[18] C. Cucchiarini and H. Strik, "Automatic assessment of second language learners' fluency," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.