Applying Rhythm Features to Automatically Assess Non-Native Speech

Lei Chen, Klaus Zechner

Educational Testing Service, Princeton, NJ

{LChen,KZechner}@ets.org

Abstract

Speech rhythm measurements have been used in a limited number of previous studies on automated speech assessment, an approach using speech recognition technology to judge non-native speakers' proficiency levels. However, one of the most problematic issues of these previous studies is a lack of a comparison of these rhythm features with other effective non-rhythm features found in decade-long previous research. In this paper, we extracted both non-rhythm and rhythm features and compared them with respect to their performances to predict proficiency scores rated by humans. We show that adding rhythm features significantly improves the performance of the scoring model based only on non-rhythm features.

Index Terms: automated speech assessment, speech rhythm, prosody

1. Introduction

Speech rhythm is a way of describing the regularity of certain language elements in speech that are perceptually similar, e.g., sequences of stressed syllables. Languages around the world have been suggested to be grouped into three categories based on rhythm characteristics, including stress-timed (e.g., English, German, and Russian), syllable-timed (e.g., Italian, French, and Spanish), and mora-timed (e.g., Japanese). A large amount of previous phonetics and psycholinguistics studies suggested that different languages could be differentiated based on a variety of rhythm metrics (cf. some of widely used ones in Section 2).

Non-native speakers' production of second language (L2) is strongly influenced by their native language (L1). Therefore, if the rhythm structure of a non-native speaker's native language is different to his/her target language, it is reasonable to conjecture that non-native speakers' rhythm patterns on the target language could be deviating from the native speakers' rhythm patterns. Hence, rhythm metrics measuring deviations to native language rhythm norms could be used to measure non-native speakers' proficiency levels. In fact, recently, a limited number of recent studies, which will be reviewed in Section 2, have utilized rhythm metrics to distinguish native and non-native speakers and to score non-native speakers in an automated way.

In this paper, we will report our research work of applying rhythm metrics in automated speech assessment on a large-sized non-native speech corpus. Unlike the previous studies, we will explore the question whether and how much the addition of rhythm features can benefit a scoring model to predict non-native speakers' proficiency scores as provided by human raters.

The paper is organized as follows: Section 2 reviews the previous research of rhythm measurements and the applications on non-native speech; Section 3 describes the speech corpus we collected for our experiments; Section 4 describes the automated speech assessment system, the speech features representing various dimensions of speaking skills, as well as the rhythm features under investigation; Section 5 reports the experimental results, including these rhythm features' prediction abilities and assessment performance using these rhythm features; finally, Section 6 discusses the findings from the experiments and plans for future research works.

2. Previous Research

Ramus et al. [1] proposed the three metrics based on speech unit internals, including ΔV, the standard deviation of vocalic internals (Vs); ΔC, the standard deviation of consonantal internals (Cs); and $\%V$, the relative duration of vocalic intervals withinin the total uterance. When these metrics were applied to distinguish stress-timed languages, syllable-timed langauge, and mora-timed language. ΔC and $\%V$ showed significantly different values among these three language groups [1]. Low et al. [2, 3] proposed the pairwise variability index (PVI) metrics, which will be described in detail in Section 4.3, to capture the sequential nature of rhythmic contrasts. For stress-timed languages, such durational variation between successive speech units (such as syllables, vowels, and consonants) tends to be great. Dellwo [4] utilized a rate-normalized metric, including $VarcoC$, which is the standard deviation of consonantal interval durations (Cs) divided by the mean consonantal durations (and multiplied by 100). He found that $VarcoC$ produced clearer discrimination than ΔC at all rates between stress-timed English and German and syllable-timed French. In addition, several researchers utilized the mathematical apparatus of coupled oscillators to model speech rhythm [5].

Recently, speech rhythm metrics have also been applied to automated speech assessment. For example, Liscombe measured distances among stressed syllables and syllables with boundary tones and computed several statistics of these distances, such as the mean distance of successive stressed syllables or syllables with boundary tones, to be the features for scoring speaking proficiency [6]. Therefore, his method requires knowing the locations of stressed syllables, which are typically determined by using a classifier (e.g., decision tree) trained on a data set with stress/no-stress labels annotated by human. However, obtaining such annotated data is a time-consuming and costly process. In addition, accurate stress/tone annotation is a challenging task and often suffers from low inter-annotator agreement.

In other related research, the rhythm metrics that do not require knowing stress or tone information, have been utilized. Jang [7] studied effective rhythm features that can be used to assess Korean speakers' English skills. He found that the metrics regarding vocalic intervals are more effective than consonantal interval measures. Tepperman et al. [8] conducted a research on Japanese learner's parroting of English to judge a speaker to be native or non-native. In his continuing work, Jang [9] conducted

research on using effective rhythm features found in [7] to build an automated system to predict Korean English learners' proficiency levels on a five-point scale from poor (1) to native-like (5). In particular, he used the rhythm features plus three non-rhythm features, including proportion of function words, speech rate (number of syllables per second), and the number of pauses and silences within an utterance, to predict a rhythm score. However, no results were provided regarding the gain of adding rhythm features.

[7, 8] used rhythm features to distinguish English learners from native speakers. [9] utilized rhythm features to distinguish different non-native skill levels. However, [9] only used a small data set that just contained one sentence read by hundreds of Korean speakers in the experiment. In addition, these recent studies required the reading content to be manually transcribed before using a forced alignment method to get temporal information of phonemes. This is not practical for fully automated speech assessment. Finally, [9] did not investigate the contribution of using rhythm features on the basis of using non-rhythm features. In this paper, we will address these issues.

3. Speech Corpus

The speech corpus used in our experiments was collected from non-native English learners who speak Mandarin, a syllable-timed native language. About $1,500$ speakers with a range of proficiency levels, from beginning learners to advanced learners, took part in the data collection. Speech audio was collected by using either land-line telephones or cell phones.

The non-native speakers were asked to read four paragraphs randomly selected from a pool of 12 reading paragraphs. Each paragraph consists of several sentences and typically takes a normal speaker about 45 seconds to finish.

From all collected audio data, about 40.0 hours of audio were used to train and evaluate the speech recognition system. The remaining part was used for building and testing the automated speech assessment system. The audio responses in this set were scored by experienced human raters. They were asked to score these reading responses according to *pronunciation* and *intonation* aspects using a three-point scale each, 1 for low-level, 2 for medium-level, and 3 for high-level. Then, an averaged score from the pronunciation score and the intonation score was computed; yielding a 5-point scale (i.e., 1.0, 1.5, 2.0, 2.5, 3.0). Thereafter, these averaged scores will be referred to as *human scores* in the paper and they are denoted as $SC1$, $SC2$, $SC3$, $SC4$, and $SC5$, respectively. Table 1 reports the basic statistics of the number of responses of two sets (train and eval) for our experiment. The human score distributions on these two sets are also reported. On the eval set, double-scoring by two human raters was conducted. A quadratic weighted κ of 0.67 was obtained on the eval set, suggesting that a reasonably well-scored data set is used our experiments.

Set	N	N_{SC1}	N_{SC2}	N_{SC3}	N_{SC4}	N_{SC5}
train	1052	231	198	413	140	70
eval	825	162	135	321	122	85

Table 1: Human score distribution of the data sets used for scoring model training and evaluation

4. Method and Metrics

4.1. ASR-based assessment system

In the previous decade, many systems have been built that use automatic speech recognition (ASR) to score non-native speech [10, 11, 12]. [13] provides a comprehensive review of the related technologies. [14] proposed a two-stage approach to address the issue that true spoken content is unknown for automated speech assessment, in particular for spontaneous spoken responses. A speech recognizer with an acoustic model (AM) optimized for non-native speech is used to get word hypotheses of spoken responses, and then another speech recognizer with an AM reflecting target language' characteristics is used to extract pronunciation features.

In our experiments, we used such a two-stage approach as proposed in [14]. A state-of-the-art Hidden Markov Model (HMM) speech recognition system was used in our experiments. For training the AM used for decoding non-native speech responses, a large-sized speech corpus, including roughly 36 hours of reading speech from our data collection, has been used. For the forced-alignment stage, a native AM was trained from about 100 hours speech data extracted from the Switchboard corpus, covering a range of North America dialects.

4.2. Typical non-rhythm features

Language skill comprises multiple dimensions, including fluency, pronunciation, and prosody. In previous research, most speech features were derived based on ASR results. For example, speaking rate and pause profile have been suggested to be useful features in several studies and even in commercial product implementations [10, 11, 12]. A widely used feature for verifying pronunciation is the recognition likelihoods estimated by the ASR system [15, 16, 14]. For reading assessment, accuracy of reading is an important features. For example, correct words per minute (CWPM) was widely used in literacy studies and has been applied on automated reading assessment [17].

Following the two-stage approach [14], we first recognize the read-aloud responses and then run forced-alignment on the obtained ASR hypotheses. Then, we derive non-rhythm features based on the the recognition and forced alignment results. From a large pool of available features we obtained, relying on Pearson correlation between a feature and human scores, which will be described in Section 5, we decided to use the following features to represent several key aspects of read-aloud speech:

- *Pace-word* is the number of words divided by the duration of the response.
- *Pace-type* is the number of unique words divided by the duration of the response; words with different inflections are counted to be different unique words. This feature also helps to measure non-native speakers' vocabulary size to some extent.
- *Pause* is the number of silences (the shortest silence needs to be at least 0.15 seconds long) divided by the duration of speaking segment, which is the duration of a response excluding silences and disfluencies; to deal with negative impact of very short speaking segments and increase the robustness of this feature, we use $log(1 + segment)$.
- *Accuracy* is the number of correctly recognized words based on reading prompts divided by the duration of the response.

- *Pronunciation* is the summation of alignment likelihoods divided by the number of letters of the recognized words, which is quite similar to the method widely used for pronunciation verification, such as the Goodness of Pronunciation (GOP) [15].
- *ASR confidence* is the mean recognition confidence scores among all recognized words. A recognition confidence score, a value in the range of 0.0 and 1.0, represents the recognizer's confidence on its word hypothesis.

4.3. Rhythm features

Table 2 lists the rhythm features suggested in the previous studies that have been summarized in Section 2. To compute these features automatically, we identified vocalic and consonant intervals (Vs and Cs) from the phoneme stream in the forced alignment results. In addition, we used a syllabification tool from the P2Tk [18] to convert phonemes to syllables (Syls). At last, the features listed in Table 2 were extracted.

feature	computation
$\%V$	proportion of Vs
ΔV	standard deviation of Vs
ΔC	standard deviation of Cs
ΔSyl	standard deviation of Syls
VarcoV	ΔV$\times$100/mean(V)
VarcoC	ΔC$\times$100/mean(C)
nPVI-V	$100 \times \sum_{k=1}^{m-1} \lvert\frac{v_k - v_{k+1}}{(v_k + v_{k+1})/2}\rvert/(m-1)$
rPVI-C	$(\sum_{k=1}^{m-1} \lvert c_k - c_{k+1}\rvert)/(m-1)$

Table 2: A list of rhythm features and their corresponding computation methods

5. Experimental Results

First, we will examine how effectively these proposed features as described in Section 4.2 and Section 4.3 can predict human scores. For this purpose, we calculate each feature's Pearson correlation to human scores. Table 3 lists the correlation coefficients (r) on the train and eval sets. From Table 3, we can find that the typical non-rhythm speech features used for assessing reading speech show high correlations to human scores. For example, the *accuracy* feature achieved 0.45 correlation to human scores on the train set. Therefore, our baseline assessment system using these features are built on effective features.

Among the rhythm features, the features derived from vocalic intervals (Vs) show more promising predictive ability than the features derived from consonant intervals (Cs) or syllables (Syls). In particular, $\%V$, ΔV, and VarcoV show a correlation with an absolute value more than 0.20. However, in order to include some properties from consonantal intervals, we will include two features derived from Cs in the scoring model, including ΔC, and VarcoC, although their correlations to human scores are low (about 0.12).

Next, we will examine the scoring models' performance using these features. Based on a variety of features reflecting different aspects of speaking, including fluency, accuracy, pronunciation, as well as rhythm (the focus of this paper), we used a data-driven method to train CART tree models and multiple regression (MR) models, which have been widely used to build speech scoring models, to predict human scores. According to the features used in model building, for each machine learning approach, we built three scoring models:

feature	r_{train}	r_{eval}
pace-word	-0.34	-0.44
pace-type	0.35	0.36
pause	-0.35	-0.39
accuracy	0.45	0.52
pronunciation	0.29	0.41
ASR confidence	0.41	0.50
$\%V$	-0.28	-0.34
ΔV	-0.31	-0.33
ΔC	-0.12	-0.11
ΔSyl	-0.05	-0.08
VarcoV	-0.25	-0.24
VarcoC	-0.12	-0.14
nPVI-V	-0.18	-0.20
rPVI-C	-0.08	-0.03

Table 3: Pearson correlation between the speech features and human scores

- *baseline* is a model using only fluency features (i.e., *pace-word*, *pace-type*, and *pause*).
- *improved* is a model using all features used in the *baseline* model plus the *accuracy*, *pronunciation* and *ASR confidence* features.
- *rhythm-added* is a model using all features used in the *improved* model plus several rhythm features, including $\%V$, ΔV, VarcoV, ΔC, and VarcoC.

Both CART tree and MR models are implemented by using the WEKA machine learning package [19]. We used the default setting provided by the software and only enabled pruning for the CART tree models to further improve the classification accuracy. To measure performance of the scoring models, we used three metrics widely used in evaluating assessment systems, namely the quadratic κ, the exact-agreement, and Pearson correlation (r) between human scores and machine predicted scores. Note that the machine scores computed from the regression models are rounded to five score levels ($SC1$ to $SC5$).

Table 4 reports on the experimental results of the scoring models implemented in CART tree and MR models using different speech features. Among human rater pairs, the Pearson correlation is 0.673 and exact-agreement is 45.21% on the eval set. For the CART tree approach, $CART_{baseline}$ has a agreement of 41.09% but quite low κ and r. After adding three non-rhythm features related to pronunciation and accuracy, although the agreement has not been improved, $CART_{imporved}$ shows substantial increases on κ and r. After adding the rhythm features, the agreement between human scores and machine scores from the $CART_{rhythm-added}$ has been improved to 44.36%. Such agreement increase is statistically significant based on a Wilcoxon sign-test ($p < 0.05$). In addition, by using all features reflecting multiple aspects of speaking skills, the agreement shown on $CART_{rhythm-added}$ (44.36%) becomes very close to human-human agreeement (45.21%).

For the multiple regression (MR) approach, we can find that MR models typically have a lower agreement but higher κ and r measures compared to the corresponding CART tree models. In particular, $MR_{baseline}$ shows a lower agreement (35.39%) but higher κ (0.30) and r (0.409) than the $CART_{baseline}$. When adding those three non-rhythm features, which are typically useful for read-aloud assessment, the $MR_{improved}$ shows a substantial increase to achieve κ (0.41) and r (0.522) for its machine scores. After adding the rhythm

features, $MR_{rhythm-added}$ shows a further increased κ (0.42) and r (0.539) between machine scores and human scores. However, compared to $MR_{improved}$, $MR_{rhythm-added}$ does not show a significantly increased agreement to human scores. In summary, compared to the scoring models using only non-rhythm features, adding rhythm features improves assessment performance measured on κ, agreement and correlation. For the CART tree modeling approach, the agreement increase is statistically significant.

model	κ	agreement(%)	r
$CART_{baseline}$	0.21	41.09	0.278
$CART_{improved}$	0.39	41.09	0.421
$CART_{rhythm-added}$	0.40	44.36	0.442
$MR_{baseline}$	0.30	35.39	0.409
$MR_{improved}$	0.41	39.03	0.522
$MR_{rhythm-added}$	0.42	39.15	0.539
human-human	0.67	45.21	0.673

Table 4: A comparison of scoring models based on CART tree and multiple regression (MR) approaches, including the *baseline* model using only fluency features, the *improved* model using all non-rhythm features, and the *rhythm-added* model using rhythm features on the basis of the *improved* model.

6. Discussion

In this paper, we reported our experiments on a large-sized reading speech corpus collected from non-native English speakers using Mandrain as their own L1. We used ASR technology to extract the speech features reflecting multiple aspects of speaking (e.g., fluency, accuracy, and pronunciation). In addition, rhythm features described in past literature [1, 2, 3, 4], especially from applications on speech assessment [7, 9, 8], were extracted. First, on a large-sized non-native speech corpus from Mandarin language background, we found that some rhythm features show promising correlations to human scores. Secondly, we compared automated speech assessment models on two conditions: using rhythm features vs. excluding rhythm features. For the CART tree modeling approach, the model using rhythm features showed a statistically significant increase of agreement between machine-predicted scores and human scores, compared to the model without using rhythm features.

Compared to previous studies, there are several contributions from this paper. First, on a much larger speech corpus than the one used in [9], by directly comparing assessment results when using vs. excluding rhythm features, our experiments showed that rhythm features play an important role for achieving higher assessment performances. Secondly, in our experiments, we used the speech recognition hypotheses rather than manual transcriptions as used in [9]. Our finding that rhythm features derived from this setting still help improving speech assessment can provide a fully automatic way to utilize rhythm features in speech assessment. In addition, the usefulness of rhythm features found in our experiments could provide a quick and robust way compared to the methods relying on finding stressed syllables (e.g., [6]).

In future work, we will expand the non-native speakers to a wider native language spectrum. It will be interesting to see whether rhythm features also are useful on non-native speakers from stress-timed language. Secondly, although most of the applications of rhythm features focused on reading, we will investigate the application on spontaneous speech, as well.

7. References

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

[2] L. E. Ling, E. Grabe, and F. Nolan, "Quantitative characterizations of speech rhythm: Syllable-Timing in singapore english," *Language and Speech*, vol. 43, no. 4, p. 377, 2000.

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

[4] V. Dellwo, "Rhythm and speech rate: A variation coefficient for deltac," *Language and language-processing*, p. 231241, 2006.

[5] M. O'Dell, M. Lennes, S. Werner, and T. Nieminen, "Looking for rhythms in conversational speech," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 1201–1204.

[6] J. J. Liscombe, "Prosody and speaker state: paralinguistics, pragmatics, and proficiency," Ph.D. dissertation, Columbia University, 2007.

[7] T. Y. Jang, "Speech rhythm metrics for automatic scoring of english speech by korean efl learners," *Malsori (Speech Sounds) The Korean Society of Phonetic Sciences and Speech Technology*, vol. 66, pp. 41–59, 2008.

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

[9] T. Y. Jang, "Automatic assessment of non-native prosody using rhythm metrics: Focusing on korean speakers' english pronunciation," in *Proc. of the 2nd International Conference on East Asian Linguistics*, 2009.

[10] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, p. 401, 2010.

[11] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, p. 355, 2010.

[12] K. Zechner, D. Higgins, and X. Xi, "SpeechRater: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech," in *Proc. SLaTE*, 2007.

[13] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[14] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[15] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[16] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," in *InSTiLL (Intelligent Speech Technology in Language Learning)*, Dundee, Stotland, 2000.

[17] K. Zechner, J. Sabatini, and L. Chen, "Automatic scoring of children's read-aloud text passages and word lists," in *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 2009, pp. 10–18.

[18] (2011) Penn phonetics toolkit. [Online]. Available: http://www.ling.upenn.edu/phonetics/p2tk/

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 1018, 2009.