10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Tree-based Clustering of Vowel Duration Ratio Toward Dictionary-based Automatic Assessment of Prosody in L2 English Word Utterances

Kohei Kitamura, Tsuneo Kato, Seiichi Yamamoto

Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan
ctwd0126@mail4.doshisha.ac.jp

Abstract

Placing correct accents in producing a word is the first step for second language (L2) learners to acquire the rhythm of a language. To evaluate correctness of the contrast between long and short syllables, we have proposed a referential vowel duration ratio (R-VDR), which takes the ratio of the segmental duration between two vowels in consecutive syllables in reference to the magnitude relation of the duration between the same vowels in the same word uttered by native speakers. The R-VDR significantly improved the correlation between the objective and subjective assessment scores on prosody (subjective-objective score correlation). However, it requires a native speaker's reference utterance of the same word. To migrate from referencing native speakers' utterances to referencing a pronunciation dictionary, we applied tree-based clustering to the weights for computing the objective score. A preliminary experiment showed that rational clusters were formed by the resulting decision tree, although a weighted mean of log VDR with the clustered weights improved the subjective-objective score correlation slightly compared with the arithmetic mean of log VDR.

Index Terms: duration, prosody assessment, L2 speech

1. Introduction

Correct prosody control is an important part of learning a second language because L2 learners' accents, intonation, and rhythm greatly affect the intelligibility of their speech. To learn the prosody control of a language, the first step is learning correct lexical stress. Thus, computer-assisted pronunciation training (CAPT) systems are focusing on automatic assessment of prosody control next to pronunciation of phoneme segments.

A basic approach to automatic prosody assessment is comparing learners' read-aloud utterances with natives' reference utterances of the same text. For example, Silverman et al. compared tones and break indices (ToBI [1]) sequences labeled at L2 learners' and natives' utterances, and showed that mutual information was a good estimator of subjective scores on prosody [2]. Arias et al. proposed an automatic intonation assessment as well as an automatic stress assessment based on a similarity measurement of a F0 trend between learner's and reference utterances with dynamic time warping (DTW) alignment [3]. Cheng proposed directly comparing a normalized F0 or energy contour of a learner's utterance with multiple reference contours [4]. They evaluated the method with real L2 assessment data collected in a large-scale English read-aloud test and achieved a subjective-objective correlation higher than the correlation between human raters. Motivated by Cheng's method, we proposed an improved contour comparison method with a weighted distance that put more weight on a frame-level distance around high values of a reference F0 or intensity contour and with variable numbers of references reflecting the diversity of native reference contours [5].

F0 and energy contours were mainly used in these automatic prosody assessment methods, and the segmental duration of syllables or phonemes has not been used as much. However, the segmental duration provides a primary cue in various linguistic distinctions in English [6], and the correct contrast of long and short syllables is particularly important for novice L2 learners [7, 8].

Segmental duration has been exploited in stress detection studies [9, 10, 11]. Tepperman and Narayanan proposed a two-step primary stress detector on the basis of a Gaussian mixture model (GMM) and maximum posterior probability with the duration of syllable nuclei, F0, and RMS energy as features. Deshmukh and Verma proposed word-independent syllable stress classification by introducing CART-based clustering to both acoustic and duration features [10]. Syllables with acoustically similar nuclei are grouped together and a separate stress classifier is trained for each group. Ferrer et al. presented a GMM-based system for detecting lexical stress in English words spoken by L2 learners with both prosodic and spectral features and showed that the duration feature as well as mel-frequency cepstral coefficients (MFCC) log-posterior probability and energy features were important in an experiment with real English speech data spoken by native Japanese children.

As a measure of rhythm established from combinations of duration and timing of consecutive syllables, the Pairwise Variability Index (PVI) proposed by Grabe et al. [12] has been widely used to quantify the rhythm of languages. Some studies have applied the PVI to assessing L2 speech, such as assessing a proficiency level [13, 14], predicting levels of prosodic control using feature selection and linear regression of a number of prosodic features that included the PVI [15] and classifying native and non-native speech with an optimized PVI [16]. However, the PVI does not consider correctness of the contrast between long and short syllables or rhythm formed by a longer series of syllables. Recently, Kyriakopulos et al. proposed a deep learning approach to automatically assess rhythm in non-native English speech [17]. They explored deep rhythm features with a recurrent neural network with attention mechanisms based on a set of conventional rhythm features in the literature.

We previously proposed the referential vowel duration ratio (R-VDR), which considers correctness of long and short syllables by referencing those of native's speech, and the weighted mean of the ratios as a feature for automatically assessing prosody in L2 word utterances [18]. The R-VDR significantly improved the subjective-objective score correlation, but, it requires native speakers' reference utterances of the same text. To migrate from referencing native reference utterances to referencing a pronunciation dictionary, (i.e., to obtain the weights on the basis of vowel pairs, their stress, and their phoneme context information without a native reference utterance), we introduce tree-based clustering to the weights for computing an objective score in this paper.

10.21437/SpeechProsody.2020-200

2. Clustering of Vowel Duration Ratio

2.1. Referential vowel duration ratio

The R-VDR is calculated on a pair of consecutive syllable nuclei as in equation (1) to score how correctly a speaker distinguishes the stressed and unstressed syllables regardless of the speech rate. The numerator and denominator of the vowel duration switch in accordance with the magnitude relation of durations between the two vowels in a native reference utterance of the same word so that a good contrast of long and short syllables results in a ratio greater than 1.

$$\begin{aligned} r(i) &= \begin{cases} d_{i+1}^{(L2)}/d_i^{(L2)} & \text{if} \quad d_i^{(R)} \leq d_{i+1}^{(R)} \\ d_i^{(L2)}/d_{i+1}^{(L2)} & \text{if} \quad d_i^{(R)} > d_{i+1}^{(R)} \end{cases} \\ &= \left(\frac{d_{i+1}^{(L2)}}{d_i^{(L2)}} \right)^{sgn\left(d_{i+1}^{(R)} - d_i^{(R)}\right)} \qquad (1) \end{aligned}$$

where $d_i^{(R)}$ and $d_i^{(L2)}$ denote duration of the ith vowel segment in an utterance of the same text by a native reference speaker and a non-native speaker to assess, respectively. If the ratio is below 1, the non-native speaker is likely to have misplaced the long and short syllables of the pair.

To obtain the ratios, each native and non-native utterance is forcedly aligned at the phoneme level using an automatic speech recognition (ASR) engine, and durations of phonemes corresponding to vowels are extracted. Each vowel in a word is paired with a vowel in its following syllable with the exception of the last vowel of a word because the last vowel tends to be longer than the others when there is no following sound to close. The ratios are first calculated on native utterances to determine which of the pair becomes the numerator and which becomes the denominator.

Next, a representative value of the word utterance is calculated on the basis of $r(i)$ of all the pairs of consecutive vowels. The first one is a geometric mean G of all ratios in a word on a logarithmic scale, which is an arithmetic mean of the logarithmic ratios.

$$\begin{aligned} G &= \frac{1}{M-1} \sum_{i=1}^{M-1} \ln r(i) \\ &= \frac{1}{M-1} \sum_{i=1}^{M-1} \operatorname{sgn} \left(\ln \frac{d_{i+1}^{(R)}}{d_i^{(R)}} \right) \ln \frac{d_{i+1}^{(L2)}}{d_i^{(L2)}} \qquad (2) \end{aligned}$$

where M denotes the number of vowels in an utterance.

Considering the correlation with a subjective score, it is reasonable to put more weight on the ratio of a pair that includes a stressed vowel than on that of a pair that does not. Let the logarithmic vowel duration ratio of a native reference be a weight and the equation (2) be extended as:

$$\begin{aligned} G^w &= \frac{\sum_{i=1}^{M-1} \left| \ln \frac{d_{i+1}^{(R)}}{d_i^{(R)}} \right| \ln r(i)}{\sum_{i=1}^{M-1} \left| \ln \frac{d_{i+1}^{(R)}}{d_i^{(R)}} \right|} \\ &= \frac{\sum_{i=1}^{M-1} \left(\ln \frac{d_{i+1}^{(R)}}{d_i^{(R)}} \ln \frac{d_{i+1}^{(L2)}}{d_i^{(L2)}} \right)}{\sum_{i=1}^{M-1} \left| \ln \frac{d_{i+1}^{(R)}}{d_i^{(R)}} \right|} \qquad (3) \end{aligned}$$

Table 1: *List of binary questions for tree-based clustering*

Anterior (posterior) vowel has the primary stress.
Anterior (posterior) vowel has a stress
Anterior (posterior) vowel is a diphthong.
Anterior (posterior) vowel is a lax vowel.
Anterior (posterior) vowel is a specific vowel [1]
Anterior (posterior) vowel has /l/ or /r/ to follow.
Anterior (posterior) vowel has a nasal consonant to follow.
Anterior (posterior) vowel has a voiced consonant to follow.

[1] "A specific vowel" stands for one of monophthongs, /aa/, /ae/, /ah/, /ao/, /eh/, /er./, /ih/, /iy/, /uh/, /uw/ or diphthongs, /aw/, /ay/, /ey/. /ow/, /oy/.

Then, the weighted mean G^w is scaled up to a score $S^{(dur)}$ that ranges from 1 to 5 by linear interpolation:

$$S^{(dur)} = \frac{S_{min}^{(dur)}(G_{max}^w - G^w) + S_{max}^{(dur)}(G^w - G_{min}^w)}{G_{max}^w - G_{min}^w} \qquad (4)$$

where G_{max}^w and G_{min}^w denote the maximal and minimal values of the mean. $S_{min}^{(dur)}$ and $S_{max}^{(dur)}$ are 1 and 5, respectively.

The R-VDR is able to capture if a stressed vowel is produced longer than an unstressed vowel. However, it is not a simple question of whether the ratio can evaluate vowel insertion into consonant clusters when a canonical phoneme sequence is given for forced alignment. The R-VDR is considered complementary with the F0 and energy contour comparison. Hence, the weighted mean of the logarithmic ratio is evaluated in combination with the improved contour comparison framework proposed by Truong et al. [5],

The duration score $S^{(dur)}$ on the basis of G^w significantly improved the subjective-objective score correlation [18]. However, this method requires a native reference utterance of the text.

2.2. Clustering natives' vowel duration ratios

If the weight $\ln d_{i+1}^{(R)}/d_i^{(R)}$ in equation (3) is correctly estimated on the basis of a pronunciation dictionary, the automatic evaluation of duration can go without a native reference utterance. To this end, we apply tree-based clustering, which is a popular technique for tying context-dependent phoneme models in hidden Markov model (HMM)-based speech recognition or speech synthesis systems, to the weights for various pairs of vowels in consecutive syllables.

Let $w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))$ denote a weight for a class of the combination of the ith vowel and its following vowel, where $\boldsymbol{v}_i$ represents phonemic, contextual and prosodic information of the ith vowel. The weight $w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))$ corresponds to $\ln d_{i+1}^{(R)}/d_i^{(R)}$ in equation (3). The method initially collects all the samples of $\ln d_{i+1}^{(R)}/d_i^{(R)}$ into a single root node, then splits the node by applying binary questions on the basis of phonemic, contextual, and prosodic knowledge in a stepwise manner, and finally lets the weights in every leaf node share a value. More details are described below.

The method hypothesizes a normal distribution of the weights for every node in the tree. Initially, all the weights of pairs of vowels are put into a root node. A list of binary questions on two vowels in consecutive syllables is prepared as shown in Table 1. A binary question classifies every pair of vowels into either the 'yes' or 'no' class. The method tests all the questions to the root node estimating the increase in the sum of log likelihoods by the split and applies the one producing the greatest increase. After the first split, the method greedily finds the best combination of a node to split and a binary question that produces the greatest increase in the sum of log likelihoods. The method repeats the binary split until the greatest increase falls below a predefined threshold.

The sum of log likelihoods of node S_k is approximated as:

$$\begin{aligned} L(S_k) &= \sum_{m=1}^{M_k} \log N(w_m; \mu_k, \sigma_k) \\ &\approx -\frac{M_k}{2}(\log 2\pi + 2\log\sigma_k + 1) \end{aligned} \tag{5}$$

where w_m, $N(w_m; \mu, \sigma)$ and M_k denote the mth weight in the node S_k, a probability that a normal distribution with a mean μ and a standard deviation σ produces a weight w_m, and the number of weights in the node, respectively. The increase in the sum of log likelihoods is calculated as follows:

$$\Delta L(S_k q) = L(S_{k,y(q)}) + L(S_{k,n(q)}) - L(S_k) \tag{6}$$

where $L(S_{k,y(q)})$ and $L(S_{k,n(q)})$ denote the sums of log likelihoods for the 'yes' and 'no' nodes split by a question q.

Finally, all the weights in a leaf node share their mean value. We denote the weight of a leaf node as $w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))$ hereafter. The advantage of tree-based clustering is that the decision tree assigns an appropriate class (leaf node) of a weight to every pair of vowels by following a series of binary questions even if a word contains a pair of vowels unseen in the training data. This method provides necessary weights to score a learner's utterance for every word in a dictionary.

2.3. Weighted mean of learner's vowel duration ratios with clustered weights

A representative value of a learner's word utterance is calculated by taking the weighted mean of the logarithmic ratios with the clustered weights $w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))$. The proposed weighted mean replaces $\ln d_{i+1}^{(R)}/d_i^{(R)}$ in equation (3) with the clustered weight $w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))$. The new weighted mean is:

$$G^c = \frac{\displaystyle\sum_{i=1}^{M-1} \left\{ w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1})) \ln \frac{d_{i+1}^{(L2)}}{d_i^{(L2)}} \right\}}{\displaystyle\sum_{i=1}^{M-1} |w(cl(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}))|} \tag{7}$$

The duration score $S^{(dur)}$ is obtained by scaling up the weighted mean G^c in the same way as equation (4).

3. Experiments

3.1. Data

We conducted experiments on 910 utterances of isolated English words produced by Japanese learners of English from the English Read by Japanese (ERJ) corpus [19]. This non-native data set consists of 36 words with different numbers of syllables and various stress patterns, The words are listed in Table 2. The 910 utterances were produced by 160 Japanese university students: 79 female and 81 male.

ERJ corpus includes subjective scores of the prosodic quality of each utterance rated by two native English teachers from America. They rated each utterance on a scale of 1 ("very poor") to 5 ("excellent"). Overall, the subjective score correlation (the correlation between the native raters) equaled 0.480. The low correlation was mainly due to different criteria for each scale between the raters. One rater seldom rated 1, whereas the other rated from 1 to 5 dispersedly. Furthermore, 18% of the utterances were rated differently with a score gap equal to or greater than 2 between the raters. This correlation coefficient of the native raters is nonetheless considered a target value of subjective-objective score correlations.

Table 2: *List of English words for assessment*

accessory	electric	academician
kangaroo	electronic	epistemology
technology	desert	differentiate
escalator	pattern	intercommunicate
dessert	control	totalitarian
percent	economic	inferiority
spaghetti	gorilla	theatricality
volunteer	orchestra	instrumental
penalty	cigarette	geology
influenza	millionaire	geological
delicate	dialect	computer
democracy	innovation	computation

We used the same native speech data set of 504 utterances of the 36 words used for computing R-VDR [18] to obtain clustered weights through tree-based clustering. To compare the performance with the original R-VDR, the evaluation is text closed.

3.2. Experimental procedure

First, the duration of vowel segments is measured on both the learners' and natives' utterances on the basis of manually-corrected phoneme segmentation obtained by forced alignment with the Kaldi ASR engine. The forced alignment is based on canonical phoneme sequences in the CMU pronunciation dictionary.

Then, tree-based clustering is conducted on a set of all the data of logarithmic vowel duration ratio $\ln d_{i+1}^{(R)}/d_i^{(R)}$ measured on the natives' utterances. The list of binary questions is summarized in Table 1. We use the tree-based clustering function of the Hidden Markov Toolkit (HTK).

Finally, an automatic assessment score is computed on the basis of the clustered weights with and without averaging with the F0 and intensity scores proposed by Truong et al [5]. The performance is compared with the original R-VDR and a weighted mean with manually-designed weight clusters in terms of subjective-objective score correlation. The subjective score is the mean of the two raters' scores. Two baselines of the original R-VDR are scores on the basis of the arithmetic mean G and the weighted mean G^w of logarithmic ratios. The weighted mean with the manually-designed weight clusters is calculated with 9 classes of weights on the basis of 3x3 combinations of the primary stress, the secondary stress and unstressed for the anterior and posterior vowels, respectively.

3.3. Decision tree

Figure 1 shows the resulting decision tree produced by tree-based clustering. A binary question is labeled at the root node and every intermediate node. The left and right branches stretched from the node correspond to 'no' and 'yes' of the bi-

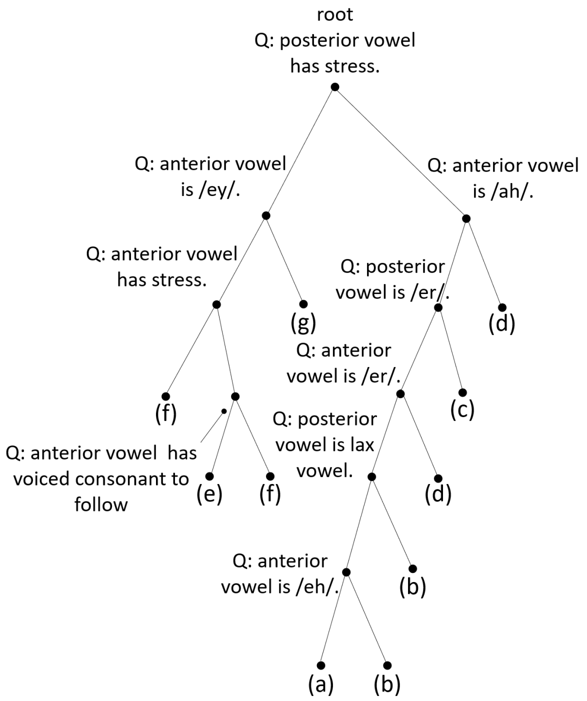
root
Q: posterior vowel has stress.
Q: anterior vowel is /ey/.
Q: anterior vowel is /ah/.
Q: anterior vowel has stress.
(g)
Q: posterior vowel is /er/.
(d)
Q: anterior vowel is /er/.
(f)
(c)
Q: anterior vowel has voiced consonant to follow
(e)
(f)
Q: posterior vowel is lax vowel.
(d)
Q: anterior vowel is /eh/.
(b)
(a)
(b)

Figure 1: *Resulting decision tree. Left and right branches stretched from a node correspond to 'no' and 'yes' of the question, respectively. Symbol at each leaf node represents a cluster of weight. Common symbol at separate nodes stand for a merged class.*

nary question, respectively. A symbol at each leaf node represents a class of weight.

The root node at the top was split by the question of whether the posterior vowel has a stress or not. The 'no' node on the left side was then split by the question of whether the anterior vowel is /ey/ or not. The 'yes' node on the right side was split by the question of whether the anterior vowel is /ah/ or not. Finally, seven classes (i.e. leaf nodes) were formed by the clustering. The normal distributions with a mean and a standard deviation of the classes are shown in Figure 2. The normal distributions were arranged in the $\ln d_{i+1}^{(R)}/d_i^{(R)}$ axis in a rational way, although the standard deviations were not very small.

3.4. Results of subjective-objective score correlation

Table 3 shows the correlation coefficients of the objective scores with the mean subjective scores of the two raters. The objective scores from 1) to 4) were $S^{(dur)}$ without averaging with F0 and intensity scores $S^{(F0)}$ and $S^{(int)}$. Those from 5) to 8) were the average scores of $S^{(dur)}$, $S^{(F0)}$ and $S^{(int)}$.

Without averaging with $S^{(F0)}$ and $S^{(int)}$, $S^{(dur)}$ on the basis of the weighted mean G^c showed a correlation coefficient (0.219) higher than those on the basis of G (0.191) and the weighted mean with the manually-designed weight clusters (0.132) but lower than that on the basis of the original weighted mean G^w (0.266). With averaging, the score on the basis of the weighted mean G^c showed a correlation coefficient (0.329) higher than that on the basis of the weighted mean of the manually-designed weight clusters (0.309) but lower than those on the basis of G (0.346) and G^w (0.381).

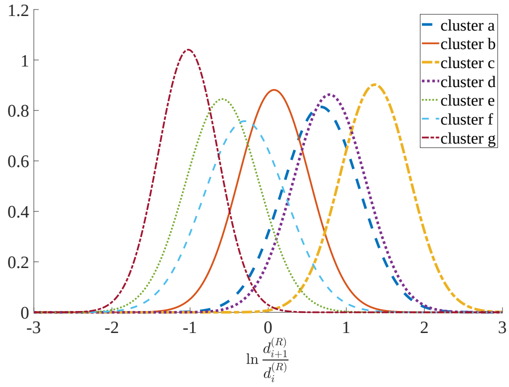

Figure 2: *Normal distributions of clustered logarithmic vowel duration ratios.*

Table 3: *Subjective-objective score correlations*

	method $S^{(dur)}$ is based on	averaging with $S^{(F0)}$ & $S^{(int)}$	subj.-obj. score corr.
1)	Arithmetic mean G	-	0.191
2)	Weighted mean G^w	-	0.266
3)	Manually-designed	-	0.132
4)	Weighted mean G^c	-	0.219
5)	Arithmetic mean G	●	0.346
6)	Weighted mean G^w	●	0.381
7)	Manually-designed	●	0.309
8)	Weighted mean G^c	●	0.329

4. Conclusions and Future Work

To migrate the automatic prosody assessment with the vowel duration ratios from referencing native utterances to referencing a pronunciation dictionary, we applied tree-based clustering to a class of vowel duration ratios collected from native reference utterances. The resulting decision tree with the normal distributions was formed in a rational way. The objective score based on the clustered weights showed a subjective-objective score correlation higher than those on the basis of the arithmetic mean and the weighted mean with the manually-designed weight clusters in an assessment condition without averaging with F0 and intensity scores, although it did not exceed the score based on the arithmetic mean in the assessment condition with averaging.

In future work, we will reestimate the weights by maximizing the objective function of the subjective-objective score correlation because the logarithmic vowel duration ratios of native reference utterances are not considered the optimal weights. Furthermore, we will collect more polysyllabic English words with native reference utterances to obtain a more robust decision tree through greater-scale tree-based clustering. The subjective-objective score correlation is to be evaluated in a text-open manner.

5. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 17K02954.

6. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrhumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proc. ICSLP 1992*. ISCA, 1992, pp. 867–870.

[2] D. Escudero, C. González, L. Aguilar, and E. Estebas, "Automatic assessment of non-native prosody by measuring distances on prosodic label sequences," in *Proc. Interspeech 2017*. ISCA, 2017, pp. 1442–1446.

[3] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, pp. 254–267, 2010.

[4] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Proc. Interspeech 2011*. ISCA, 2011, pp. 1589–1592.

[5] Q. Truong, T. Kato, and S. Yamamoto, "Automatic assessment of L2 English word prosody using weighted distantces of f0 and intensity contours," in *Proc. Interspeech 2018*. ISCA, 2018, pp. 2186–2190.

[6] D. Klatt, "Linguistic uses of segmental duration in English: acoustic and perceptual evidence," *J. Acoust. Soc. Amer.*, vol. 59, pp. 1208–1221, 1998.

[7] D. Kim, M. Clayards, and H. Goad, "A longitudinal study of individual differences in the acquisition of new vowel contrasts," *Journal of Phonetics*, vol. 67, pp. 1–20, 2018.

[8] K. Yazawa, J. Whang, M. Kondo, and P. Escudero, "Language-dependent cue weighting: An investigation of perception modes in L2 learning," *Second Language Research*, pp. 1–25, 2019.

[9] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. ICASSP 2005*, vol. 1. IEEE, 2005, pp. 937–940.

[10] O. D. Deshmukh and A. Verma, "Nucleus-level clustering for word-independent syllable stress classification," *Speech Communication*, vol. 51, pp. 1224–1233, 2019.

[11] L. Ferrer, H. Bratt, C. Richey, H. Franco, and V. Abrash, "Classification of lexical stress using spectral and prosodic featuresfor computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[12] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Laboratory Phonology*, vol. 7, pp. 515–546, 2002.

[13] L. Chen and K. Zechner, "Applying rhythm features to automatically assess non-native speech," in *Proc. Interspeech 2011*. ISCA, 2011, pp. 1861–1864.

[14] C. Lai, K. Evanini, and K. Zechner, "Applying rhythm metrics to non-native spontaneous speech," in *Proc. Speech and Language Technology in Education 2013*. ISCA, 2013, pp. 159–163.

[15] F. Honig, A. Batliner, K. Weilhammer, and E. Noth, "Automatic assessment of non-native prosody for English as L2," in *Proc. Speech Prosody 2010*. ISCA, 2010.

[16] S. Gharsellaoui, S. A. Selouani, W. Cichocki, Y. Alotaibi, and A. O. Dahmane, "Application of the pairwise variability index of speech rhythm with particle swarm optimization to the classification of native and non-native accents," *Computer Speech and Language*, vol. 48, pp. 67–79, 2018.

[17] K. Kyriakopoulos, K. Knill, and M. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native english speech," in *Proc. Interspeech 2019*. ISCA, 2019, pp. 1836–1840.

[18] T. Kato, Q. Truong, K. Kitamura, and S. Yamamoto, "Referential vowel duration ratio as a feature for automatic assessment of L2 word prosody," in *Proc. ICASSP 2019*. IEEE, 2019, pp. 1836–1840.

[19] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "English speech database read by Japanese learners for CALL system development," in *Proc. LREC 2002*. ELRA, 2002, pp. 896–903.