REFERENTIAL VOWEL DURATION RATIO AS A FEATURE FOR AUTOMATIC ASSESSMENT OF L2 WORD PROSODY

Tsuneo Kato[†]*, Quy-Thao Truong*[‡]*, Kohei Kitamura*[†] *and Seiichi Yamamoto*[†]

[†] Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan
[‡] Graduate School of Engineering, École Centrale de Nantes, Nantes, France
tsukato@mail.doshisha.ac.jp

ABSTRACT

This paper proposes a referential vowel duration ratio for a pair of vowels in consecutive syllables and a weighted mean of the referential vowel duration ratios on a logarithmic scale as a feature for automatic assessment of second-language (L2) word prosody. In addition to contours of fundamental frequency (F0) and energy i.e. suprasegmental information of speech, segmental duration of syllables or phonemes provides important information for assessing L2 prosody. For L2 learners, the first step of learning prosody is to put accents or stresses on appropriate syllables in words. A syllable with a stress should be produced longer and one without a stress should be produced shorter. To achieve this, we propose taking a duration ratio for every pair of consecutive vowels in reference to duration contrast of the same vowel pair produced by native speakers. Furthermore, we propose a weighted mean of the ratios on a logarithmic scale in consideration of local importance within a word. In evaluation with English word utterances produced by Japanese learners, the introduction of the weighted mean of the ratio significantly improved the correlation coefficient with subjective scores.

Index Terms— duration, prosody assessment, L2 speech

1. INTRODUCTION

Prosody, which refers to aspects of speech related to intonation, stress and rhythm, is an essential part of learning a language. Computer assisted pronunciation training (CAPT) systems have been developed to provide language learners an environment to self-train speech production. Although most of the technology developed for CAPT focused on pronunciation of phoneme segments, attention must be paid to learners' prosody as well.

A basic approach to automatic prosody assessment is comparing learners' read-aloud utterances with natives' reference utterances of the same text. Recently, an automatic assessment method with prominence estimates of syllables obtained by continuous wavelet analysis was proposed [1]. Another study compared tones and break indices (ToBI labels [2]) labeled at both learners' and natives' utterances with mutual information [3]. Many studies have compared contours of F0 and energy [4, 5, 6]. Arias et al. proposed an automatic stress assessment as well as an automatic intonation assessment based on a correlation between a learner's and a reference contours with dynamic time warping (DTW) alignment [5]. They got a high subjective-objective score correlation ($r = 0.88$). Cheng proposed direct comparison of a normalized F0 or energy contour of a learner's utterance with multiple reference contours and evaluated the method with real L2 assessment data collected in a large-scale English read-aloud test [6]. The method achieved a subjective-objective correlation ($r = 0.80$) higher than the correlation between human raters ($r = 0.75$). Motivated by Cheng's study, 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 number of references reflecting the diversity of native utterance contours [7].

However, segmental duration of syllables or vowels has not been exploited as much as F0 and energy contours in these reference-based methods. Although the segmental duration provides a primary cue in various linguistic distinctions in English [8], fewer studies have made use of this. Cheng's method trained a statistical duration model of each phoneme in context with density functions. Such a statistical approach should work well in an environment where sufficient training samples are available. However, that statistical model did not model the contrast of long and short duration of stressed and unstressed syllables explicitly.

In measuring speech rhythm, Grabe et al. proposed Pairwise Variability Index (PVI) based on the duration of consecutive vowels or consonants [9]. PVI has been widely used as a metric for quantifying rhythm in speech, and there have been some studies that applied PVI to assessing L2 speech, such as assessing a proficiency level [10, 11], predicting a level of prosodic control using feature selection and linear regression of a number of prosodic features that included PVI [12] and classifying native and non-native speech with an optimized PVI [13]. However, PVI does not consider correctness of the contrast between long and short syllables. On the other

hand, the first step of learning second language speech is to put stresses on appropriate syllables in words.

Therefore, we propose a vowel duration ratio in reference to the magnitude relation of duration between two vowels in the consecutive syllables produced by native speakers. The referential vowel duration ratio captures how correctly a learner distinguishes stressed and unstressed syllables in vowel duration. Furthermore, we propose a weighted mean of the referential vowel duration ratios on a logarithmic scale considering the local importance within a word.

2. REFERENTIAL VOWEL DURATION RATIO

2.1. Referential vowel duration ratio of vowel pair

Alternation between long and short syllables, corresponding with stressed and unstressed syllables, constitutes the rhythm of stress-timed languages. A vowel duration ratio is calculated on a pair of consecutive syllable nuclei 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 ratio switches according to 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. The referential vowel duration ratio $r(i)$ for a pair of the ith and its following vowels is defined as:

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

where $d_i^{(R)}$ and $d_i^{(L2)}$ denote duration of the ith vowel segment in an utterance of a 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 the short syllables of the pair.

To get the ratios, each native and non-native utterance is forced 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. Note, however, that a vowel in the last syllable of a word is excluded from the computation of the ratio if the vowel is the last phone of the word. The reason is that the last vowel of a word tends to be longer than the others regardless of whether or not it is supposed to be stressed when there is no following sound to close the last vowel. Then, the ratios are first calculated on native utterances to determine which of the pair becomes the numerator and which becomes the denominator.

2.2. Weighted mean of logarithmic ratios

An automatic score $S^{(dur)}$ related to vowel duration is computed based on the ratios of all the consecutive vowel pairs so that the score has a high correlation with a subjective score. Since the referential vowel duration ratios are distributed in a log-normal distribution, we take a geometric mean G of all ratios in a word on a logarithmic scale, that is an arithmetic mean of the logarithmic ratios:

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

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 which includes a stressed vowel than that of a pair which does not. Although the effect of the local importance on the subjective score should have non-linearity, we assume a linear weight in reference to the logarithmic vowel duration ratio of a native reference. Let the logarithmic vowel duration ratio of a native reference be a weight, and the equation (2) is extended as:

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

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

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

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 referential vowel duration ratio is able to capture if a stressed vowel is produced longer than an unstressed vowel. However, it is not a simple question if the ratio can evaluate vowel insertion into consonant clusters when a canonical phoneme sequence is given for forced alignment. The referential vowel duration ratio 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 in [7], which will be described in the next section.

3. AUTOMATIC ASSESSMENT OF L2 PROSODY WITH CONTOUR COMPARISON

The F0 and intensity contour comparison and the referential vowel duration ratio are consistent in the sense of "reference-based" methods. Here, the improved contour comparison framework is reviewed, followed by score integration.

3.1. Feature extraction of F0 and intensity

F0 and intensity are measured and N_t equally-spaced points are extracted from the measurement to normalize the duration of utterances for comparing the contours. All features are z-normalized to reduce the natural variations between speakers, then further normalized by a sigmoid function to smooth the outliers around high values. We set N_t at 25 and the prosodic contour is denoted as U, such that $U = (u(1), ..., u(N_t))$ is the concatenation of the N_t normalized prosodic values.

3.2. Weighted distance

The normalized prosodic contour of a non-native utterance is compared to that of a reference by measuring a distance between N_t sampled points. In the distance calculation, a "weighted distance" that puts more weight on the squared error between a reference and a non-native contour around high values of the reference contour is used. The weighted distance D^w between a sigmoid-normalized reference contour U_r and a non-native one U_{L2} for a given word is then defined as:

$$D^w(U_r, U_{L2}) = \sum_{t=1}^{N_t} D^w(u_r(t), u_{L2}(t)) \tag{5}$$

where $D^w(u_r(t), u_{L2}(t))$ denotes the weighted squared error between $u_r(t)$ and $u_{L2}(t)$, the tth feature value of the reference and non-native contour, respectively.

3.3. Variable number of references

The F0 and intensity contours of a word are not unique, but they have a considerable variability among native speakers. The method sets a different number of references for each word depending on the variability in the whole set of native contours. The number of references should be appropriate enough to avoid redundancy between the references. The variability among native utterances is measured as:

$$V^w = \frac{1}{{}_nC_2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D^w(U_r^{(i)}, U_r^{(j)}) \tag{6}$$

where $U_r^{(i)}$ is the ith contour in the total set of n utterances.

Then, the number of clusters k for k-means clustering of the native utterances is linearly set in the range between 1 and half the number of the native utterances.

Table 1. 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

When measuring a weighted distance in contour comparison, each U_{L2} is compared to the centroid $U_r^{(j)} (1 \leq j \leq k)$ of each cluster, and the smallest distance D^w is kept to calculate the automatic score.

3.4. Automatic scores and their integration

An automatic score $S^{(F0)}$ (resp. $S^{(int)}$) ranging from 1 to 5 is given by a linear function of $D^w(U_r, U_{L2})$. The final automatic score S is given by an arithmetic mean of the scores, $S^{(F0)}$, $S^{(int)}$ and $S^{(dur)}$.

4. EXPERIMENTS

4.1. Data

We conducted experiments on 910 utterances of isolated English words produced by Japanese learners of English from English Read by Japanese (ERJ) corpus [14]. This non-native data set consists of 36 words with different numbers of syllables and various stress patterns, The words are listed in Table 1. The 910 utterances were produced by 160 Japanese university students, 79 female and 81 male.

The prosodic quality of the utterances was assessed by two native American English teachers who had to evaluate speakers 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 can be accounted for by the fact that there were only two human raters, the rating was made on the word basis and there was a relatively small number of utterances per word. As a result, the subjective score correlation was degraded by small variations in human ratings. This correlation coefficient of the native raters is nonetheless considered a target value of subjective-objective score correlations.

We built a native data set by recording native utterances of the 36 words from online English dictionaries [15, 16, 17, 18, 19, 20]. From 4 to 19 native utterances with an average

Table 2. Subjective-objective score correlations

Method	F0	Int.	Dur.	Corr.
Baseline #1	•	•		0.265
Baseline #2	•	•		0.304
n-PVI			•	0.005
Baseline #2 + n-PVI	•	•	•	0.303
Arithmetic mean of log ratios			•	0.191
Baseline #2 + arithmetic mean	•	•	•	0.346
Weighted mean of log ratios			•	0.266
Baseline #2 + weighted mean	•	•	•	0.381

of 14 utterances were collected for each word depending on the availability of native utterances online. There were speakers with various English accents such as Australian, Irish, Jamaican, Scottish, UK, UK Received Pronunciation, UK Yorkshire, US and US Southern.

We conducted phoneme-level forced alignment of the non-native and native utterances with canonical phoneme sequences in the CMU pronunciation dictionary using Kaldi ASR engine. However, this time we used manually-corrected phoneme segmentation for computing vowel duration ratios to eliminate the degradation by alignment errors.

4.2. Experimental conditions

We compared the proposed methods in a subjective-objective correlation, which was a correlation coefficient between the mean of two subjective scores by human raters and the automatic score. Five types of methods were compared. Baseline #1 was the basic contour comparison of F0 and intensity based on Euclidean distance with a fixed number of references. The number of references was fixed at 4 here. Baseline #2 refers to the improved contour comparison method described in Section 3. As a feature related to vowel duration, three types of scores were evaluated in combination with Baseline #2. The first one was the normalized Pairwise Variability Index (n-PVI) [9] and the second one was the score based on the arithmetic mean of logarithmic vowel duration ratios G. The third one was the score based on the weighted mean of logarithmic vowel duration ratios G^w.

4.3. Experimental results

Table 2 summarizes the subjective-objective correlations. Between the two baseline methods, introduction of the weighted distance and variable number of references improved the correlation coefficient from 0.265 to 0.304 at Baseline #2. There was no correlation between the n-PVI and the subjective score. In contrast, incorporating the arithmetic mean of referential vowel duration ratios to Baseline #2 significantly improved the correlation coefficient from 0.304 to 0.346, although the arithmetic mean itself could not sufficiently predict the subjective score. The weighted mean of the vowel duration ratio in reference to those of native references further improved the subjective-objective correlation to 0.381.

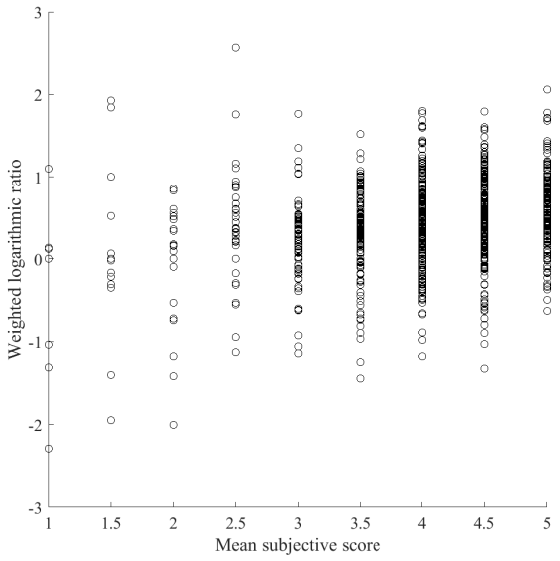

Fig. 1. Scatter graph of 910 isolated English word utterances produced by Japanese learners. The horizontal and vertical axes represent the mean subjective score of two human raters and the weighted mean of referential vowel duration ratio G^w on a logarithmic scale. The correlation coefficient is 0.266.

Fig. 1 shows the scatter graph of all the samples of the mean subjective score of two human raters and the weighted mean of the referential vowel duration ratio G^w. The correlation coefficient was 0.266. As a comparison, the overall correlation of two subjective scores was 0.480.

5. CONCLUSIONS

This paper presented a referential vowel duration ratio and its weighted mean on the logarithmic scale as a feature for automatic assessment of L2 word prosody. Incorporating the weighted mean of the logarithmic ratio into an existing automatic prosody assessment framework based on F0 and intensity contour comparison greatly improved the subjective-objective correlation in an experiment with English word utterances produced by Japanese learners of English.

In future work, we will evaluate the effectiveness of the referential vowel duration ratio with other L2 English corpora spoken by speakers whose mother language is not Japanese, and with words in read-aloud sentence utterances.

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] H. Kallio, A. Suni, P. Virkkunen, and J. Šimko, "Prominence-based evaluation of l2 prosody," in *Proc. Interspeech 2018*. ISCA, 2018, pp. 1838–1842.

[2] 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.

[3] 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.

[4] M. Suzuki, T. Kohno, A. Ito, and S. Makino, "Automatic evaluation system of english prosody based on word importance factor," *J. Systemics, Cybernetics and Informatics*, vol. 6, no. 4, pp. 83–90, 2008.

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] WordReference, "https://www.wordreference.com/," .

[16] Cambridge Dictionary, "https://dictionary.cambridge.org/," .

[17] Collins English Dictionary, "https://www.collinsdictionary.com/," .

[18] Macmillan Dictionary, "https://www.macmillandictionary.com/," .

[19] Oxford Dictionary, "https://en.oxforddictionary.com/," .

[20] Wikitionary: the free dictionary, "https://en.wikitionary.com/," .