Referential Vowel Duration Ratio as a Feature for Automatic Assessment of L2 Word Prosody

Tsuneo Kato†, Quy-Thao Truong‡, Kohei Kitamura† and Seiichi Yamamoto†
†Doshisha University and ‡Ecole Centrale de Nantes

1. Summary

- We propose Referential Vowel Duration Ratio (R-VDR), which explicitly quantifies correctness of English accents and rhythms, for automatic prosody assessment.
- The proposed method was evaluated with 910 utterances of 36 English words from English Read by Japanese (ERJ) corpus.
- The proposed method significantly improved subjective-objective score correlation from 0.30 to 0.38. (cf. Inter-rater correlation coefficient: 0.48)

2. Background and Objective

Background

- Automatic prosody assessment of L2 speech has been based mainly on fundamental frequency (F0) and energy contours.
  - J. P. Arias et al., "Automatic intonation assessment for computer aided language learning", Sp. Com. 2010.
  - J. Cheng, "Automatic assessment of prosody in high-stakes English tests", Interspeech 2011.
  - Q. Truong, T. Kato, S. Yamamoto."Automatic assessment of L2 English word prosody using weighted distences of f0 and intensity contours", Interspeech 2018.

- Long and short syllables constitute the rhythm of English, a stress-timed language.
- Segmental duration of syllables or vowels should provide important information for assessing rhythm of speech.

  - Pairwise Variability Index (PVI) E. Grabe and E. L. Low, 2002.

$$nPVI = \frac{100}{M-1}\sum_{i=1}^{M-1}\frac{|d_i - d_{i+1}|}{(d_i + d_{i+1})/2}$$

(d_i: duration of ith vowel segment, M: # of vowel segments)

✓ PVI does not consider correctness of the contrast between long and short syllables.

Objective

- Develop a metric having high subjective-objective correlation in prosody assessment based on segmental duration.

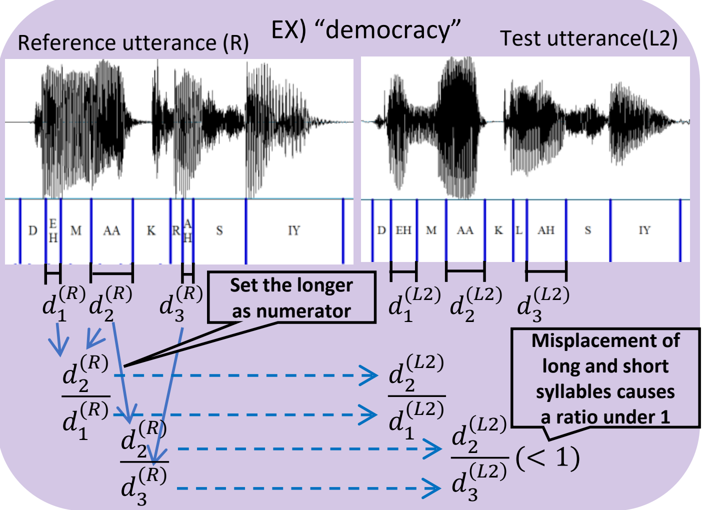
3. Referential Vowel Duration Ratio

Referential Vowel Duration Ratio of Vowel Pair

- To score how correctly a speaker distinguishes stressed and unstressed syllables regardless of a speech rate, a vowel duration ratio is calculated on a pair of consecutive syllable nuclei referring to the pair produced by natives.
- The numerator and denominator switch according to a magnitude relation of durations between the two vowel segments in a native reference utterance.

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

EX) "democracy"



Set the longer as numerator

Misplacement of long and short syllables causes a ratio under 1

Geometric mean of ratios on log scale (arithmetic mean of log ratios)

$$G = \frac{1}{M-1}\sum_{i=1}^{M-1} sgn\left(\ln\frac{d_{i+1}^{(R)}}{d_i^{(R)}}\right)\ln\frac{d_{i+1}^{(L2)}}{d_i^{(L2)}}$$

Weighted mean with the ratio of a native reference utterance

$$G^w = \sum_{i=1}^{M-1}\left(\ln\frac{d_{i+1}^{(R)}}{d_i^{(R)}}\ln\frac{d_{i+1}^{(L2)}}{d_i^{(L2)}}\right)\Bigg/\sum_{i=1}^{M-1}\left|\ln\frac{d_{i+1}^{(R)}}{d_i^{(R)}}\right|$$

4. Experiments

Data

- Test data:
  910 word utterances from English Read by Japanese corpus.
- Subjective assessment scores:
  Rated by two native English teachers. Inter-rater corr.: **0.480**
- Reference data:
  504 English native utterances from online dictionaries.

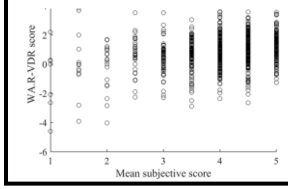
Conditions

- Baseline #1: F0 & intensity contours comparison (Cheng IS2011)
- Baseline #2: Improved contours comparison (Truong IS2018)

Subjective-objective correlations

Method	F0	Int.	Dur.	Corr.
Baseline #1	●	●		0.265
Baseline #2	●	●		0.304
nPVI			●	0.005
Baseline #2 + nPVI	●	●	●	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**

Weighted mean of the ratios



Part of word list for assessment

accessory	dessert
kangaroo	percent
technology	spaghetti
escalator	volunteer

5. Future work

- Evaluating L2 English speech corpora other than ERJ corpus.
- From referring to L1 speech to referring to an accent dictionary.
- From evaluating isolated word utterances to sentences.