*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# A Comparison of English Rhythm Produced by Native American Speakers and Mandarin ESL Primary School Learners

*Hongwei Ding*[1], *Binghuai Lin*[2], *Liyuan Wang*[2], *Hui Wang*[1], *Ruomei Fang*[1]

[1]Speech-Language-Hearing Center, School of Foreign Languages
Shanghai Jiao Tong University, China
[2] Smart Platform Product Department, Tencent Technology Co., Ltd, China
hwding@sjtu.edu.cn, {binghuailin, sumerlywang}@tencent.com

## Abstract

Prosodic speech characteristics are important in the evaluation of both intelligibility and naturalness of oral English proficiency levels for learners of English as a Second Language (ESL). Different stress patterns between English and Mandarin Chinese have been an important research topic for L2 (second language) English speech learning. However, previous studies seldom employed children as ESL learners on this topic. Since more and more children start to learn English in the primary school in China, the current study aims to examine the L2 English rhythm of these child learners. We carefully selected 273 English utterances from a speech database produced by both native speakers and Mandarin child learners, and measured the rhythmic correlates. Results suggested that vowel-related metrics (e.g. *nPVI*) are better indexes for L2 rhythmic evaluation, which is similar for ESL adults; pause-related fluency is another indication for prosodic assessment, especially for child ESL learners. This investigation could shed some light on the rhythmic difficulties for Mandarin ESL child learners and provide some implications for ESL prosody teaching for school children.

**Index Terms**: rhythmic pattern, stress-timed, syllable-timed, L2 speech, Chinese learners

## 1. Introduction

Prior studies have shown that suprasegmental (or prosodic) mistakes can lead to degradation in both intelligibility and naturalness in oral proficiency of L2. It has been shown that the suprasegmental measures collectively account for 50% of the variance in oral proficiency ratings [1]. Furthermore, many investigations have demonstrated that the improvement in L2 speech proficiency is more likely to occur with improvement in prosodic proficiency than with a sole focus on phonemic correction [2, 3]. That means an ESL training course with an aim to improve prosody may be more successful [4]. Thanks to the maturing of speech technology, Computer-Aided Pronunciation Tutoring (CAPT) programs can meet the requirements to help Chinese learners to improve their rhythm of ESL [5]. In order to provide accurate feedback information for CAPT users to facilitate their acquisition of near-native English rhythm, it is important for us to capture the characteristic prosodic mistakes in ESL produced by Chinese children.

The acoustic correlates of prosody mainly include pitch patterns (intonation), stress-related durational patterns (rhythm) as well as intensity patterns [6, 7, 8]. Compared with analysis in pitch and intensity, measurements of duration are more reliable with current speech technology. Thus, L2 speech investigations have been mainly focused on durational patterns, including stress, pausing, and speech rate. Some findings have even claimed and emphasized the importance of prosodic timing patterns for L2 learners [9]. As a preliminary investigation into the school children's L2 English, we also employ the more stable parameter of duration in the current study.

Different languages have different stress patterns, which have been classified as *stress-timed* and *syllable-timed* in rhythm. Since a large number of experiments carried out in the 1970s and 80s failed to provide direct correlates for the isochrony [10], other rhythmic indexes were developed. Compared with syllable-timed languages, stress-timed languages have a higher standard deviation of consonantal intervals ($\Delta C$) and a relatively lower proportion of the vocalic intervals ($\%V$) [10], and stress-timed languages also demonstrate a higher variation in vowel durations (e.g. *nPVI_V*, *VarcoV*) [11, 12]. These metrics have become the most widely used rhythmic indexes in classifying languages of different rhythms in many investigations [13, 14, 15, 16, 17, 18]. It is now generally accepted that languages fall along a continuum where they can be classified as being more or less *stress-timed* or *syllable-timed* in rhythm. Since English and Mandarin Chinese occupy opposite ends of the continuum, acquiring L2 English rhythm by Mandarin ESL learners can be a challenging task.

Apart from rhythmic metrics, speech rate and pause are also durational indexes for L2 speech. It has also been found that L2 learners have a slower speech rate due to less coarticulation in L2 target language [19]. Moreover, pause length and placement are also related to proficiency in L2 oral speech. It has been found that silent pauses in nonnative speech are both longer and more irregular than those in the native speakers, which can affect the overall prosodic structure of the discourse [20].

It has been suggested that the rhythm of the target language can be influenced by the learner's native language. Many studies have examined the influence of L1 on L2 in rhythm with some of the above rhythmic indexes [5, 12, 21, 22, 23, 24, 25]. However, different findings have been reported as to which rhythmic indexes or which combination of indexes can best distinguish different varieties of language [5, 18]. Since these rhythmic metrics are largely influenced by inter-speaker variation, elicitation, and syllable structures of the materials [18], the results across different studies may not be comparable.

The aim of the current study is to extract the relevant acoustic measures from an L2 English speech database of Mandarin ESL primary school learners, and to identify which durational index(es) can best distinguish different prosodic proficiency levels and which are significant predictors for the prosodic performance of the Mandarin ESL child learners. Because the results are based on acoustic calculations, possible reasons that have triggered the rhythmic deviation in L2 English from that of the native English will be clarified in the discussion.



http://dx.doi.org/10.21437/Interspeech.2020-2207

# 2. Method

We selected the materials, calculated durational measures, and employed ANOVA and logistic regression models for statistics.

## 2.1. Material

The speech materials were taken from an online L2 English training course for Mandarin primary school children, where the children were asked to follow the native American speakers to produce the same English sentence. About 5,000 sentences produced by Mandarin L2 children and over 1,000 sentences by native American speakers were automatically annotated and manually checked, where the prosodic performances of the L2 sentences were manually scored by the English teaching assistants who had evaluated the Spoken English Test of the national College English Test in China (CET-SET). In order to facilitate the investigation of prosody, the raters were asked to assess the prosodic aspect more than the segmental aspect. Since it is difficult to tease these two aspects apart, we referred to the official criteria of fluency, coherence, and pronunciation used in IELTS speaking band descriptors ("www.ielts.org") and CET-SET, and made our own criteria to meet our special purpose. We employed 1-5 scores with 1 indicating very bad prosody, and 5 indicating near-native prosody. After several rounds of trials, the raters could reach an agreement and achieved a minimal correlation of 0.7 in their assessments.

To minimize the undesirable biases brought by speakers and sentence structures, we managed to maximize the coverage of these factors. We selected 91 different sentences with an average syllable length of 5.8. Each sentence was produced by one native child speaker (*EnNa*: *En*glish *Na*tive) and two L2 child learners with one having a score over 3 (*EnHi*: *En*glish learners of *Hi*gher level) and the other below 3 (*EnLo*: *En*glish learners of *Lo*wer level). Non-integer scores were allowed, especially around the medium score. And it is taken for granted that all the native speakers obtained the highest score of 5. In this way, a total number of 273 sentences were selected. Among which *EnHi* sentences were produced by only five native speakers, while *EnHi* and *EnLo* sentences were produced by different Mandarin ESL primary school learners.

## 2.2. Analysis

In order to ensure comparability, the annotation technique used by Ramus [10] was adopted. Annotation was conducted in the following two steps on Praat [26] (1) *phonetic segmentation* of the sentence into phonemes, and (2) *classification* of separate phonemes into vowel and consonant intervals.

In the first step, following the standard of phonetic criteria [27], the trained annotators corrected the automatic annotations manually as accurately as possible by referring to both visual and audio cues. The changes of spectrogram, waveform and formants served as the visual cues for setting the boundary of segmentation. In the second step, the phonemes were then classified as vowel or consonant intervals with the criteria used by Ramus [10]: checked (free) vowels, free (long) vowels and unstressed schwa were coded as V (vowel); plosives, affricates, fricatives, sonorants (nasals and liquids) were coded as C (consonant); pre- and inter-vocalic glides were treated as consonants; post-vocalic glides were treated as vowels.

The phonetic segmentation was straightforward. The problem of labeling was the pause, especially that of the Chinese learners. Short pauses before the burst of stops and nasals were labeled as closure parts of the corresponding phonemes. If there were some pauses and hesitations, which could not be identified as part of a sound, these breath parts were then marked as *silence*. Any two consonantal intervals not split by *silence* were combined into the same consonantal interval, and the same approach was used for vowel intervals as well. Finally, we measured the relevant durational variables of consecutive vocalic intervals (VI) and those of consecutive consonantal intervals (CI), the description of which are shown in Table 1:

Table 1: *Description of rhythmic and pause variables*

| Variable | Description |
|---|---|
| %V | sum of VI duration divided by total duration of VIs and CIs |
| $\Delta V$ | standard deviation (STDEV) of vocalic interval (VI) duration |
| $\Delta C$ | STDEV of consonantal interval (CI) duration |
| rPVI_C | raw Pairwise Variability Index (PVI) for CIs |
| nPVI_V | normalised PVI for VIs |
| VarcoC | STDEV of CI duration divided by the mean CI duration |
| VarcoV | STDEV of VI duration divided by the mean VI duration |
| rateC | rate of CI (number of CI per second) |
| rateV | rate of VI (number of VI per second) |
| rateCV | rate of C or V intervals (non-pause) |
| nS | number of period of silence within a sentence |
| sumS(s) | sum of all period(s) of silence within a sentence in seconds (s) |

The rhythmic values were extracted with the help of praat plugins (Duration Analyzer) provided by Dellwo ("http://www.pholab.uzh.ch/static/volker/software.html") and other variables were calculated with praat scripts written by the first author. ANOVA and logistic regression analysis were performed in R language [28].

# 3. Results

## 3.1. Means of variables in groups

The relevant durational variables of each sentence were averaged across all sentences and speakers within the same speaker group, which is presented in Table 2:

Table 2: *Means of rhythmic metrics and pause variables*

| **Variable** | **Group** | | |
|---|---|---|---|
| | EnNa | EnHi | EnLo |
| %V | 51.00 | 51.62 | 57.20 |
| $\Delta C$*100 | 7.77 | 8.72 | 8.08 |
| $\Delta V$*100 | 9.66 | 8.43 | 7.48 |
| rPVI_C | 9.19 | 9.50 | 9.29 |
| nPVI_V | 69.48 | 52.73 | 33.38 |
| VarcoC*100 | 48.93 | 55.90 | 54.92 |
| VarcoV*100 | 53.35 | 44.91 | 32.06 |
| rateC | 6.70 | 6.99 | 7.60 |
| rateV | 5.92 | 5.74 | 5.02 |
| rateCV | 6.18 | 6.17 | 5.88 |
| nS | 1.07 | 1.46 | 1.96 |
| sumS (s) | 0.16 | 0.25 | 0.48 |

The twelve variables can be divided into three categories:

- Four (mainly pause-related) increase from *EnNa* over *EnHi* to *EnLo*: %V, rateC, nS, sumS
- Five (mainly VI-related) decrease from *EnNa* over *EnHi* to *EnLo*: $\Delta V$, nPVI_V, VarcoV*100, rateV, rateCV
- Three (mainly CI-related) increase from *EnNa* to *EnHi*, and decrease to *EnLo*: $\Delta C$, rPVI_C, VarcoC*100

We also found epentheses in 38 sentences. Sentences with 2-3 occurrences of epenthesis were all in the lower-level group (*EnLo*). Seven sentences with 1 short epenthesis were in the higher-level group (*EnHi*) but with scores just above 3.

### 3.2. One-way ANOVA across groups

One way ANOVA and post-hoc test (Tukey's HSD tests) were run for each variable separately across three groups, following $p$ values were obtained for each variable and between each two groups, which are shown in Table 3.

Table 3: *Significance levels for rhythmic metrics and pause variables, where [-] = not significant, [*] = $p < 0.05$, [**] = $p < 0.01$, [***] $p < 0.001$.*

| Variable | Significance level across groups | | |
|---|---|---|---|
| | EnNa-EnHi | EnNa-EnLo | EnHi-EnLo |
| %V | - | *** | *** |
| $\Delta C$*100 | - | - | - |
| $\Delta V$*100 | - | *** | - |
| rPVI_C | - | - | - |
| nPVI_V | *** | *** | *** |
| VarcoC*100 | * | * | - |
| VarcoV*100 | *** | *** | *** |
| rateC | - | * | - |
| rateV | - | *** | *** |
| rateCV | - | - | - |
| nS | - | *** | * |
| sumS (s) | - | *** | *** |

It is clearly shown that the number of variables that can have a significant difference decreases from *EnNa-EnLo* over *EnHi-EnLo* to *EnNa-EnHi*. Finally, only two variables, which are *nPVI_V* and *VarcoV*100*, can distinguish any two groups of speakers at a significant level. A plot with *nPVI_V* against *VarcoV* in Figure 1 can roughly separate these three groups of speakers. There are some overlaps between *EnNa-EnHi* and between *EnHi-EnLo*, but there is almost no overlap between *EnNa-EnLo*. Furthermore, the values for *EnNa* are more homogenous than those of the two learner groups (*EnHi* and *EnLo*), as the diameters of the ellipse of *EnNa* group are shorter.

### 3.3. Binominal logistic regression analysis between groups

To further examine the contribution of each variable to the classification of speaker group, we employed logistic regression

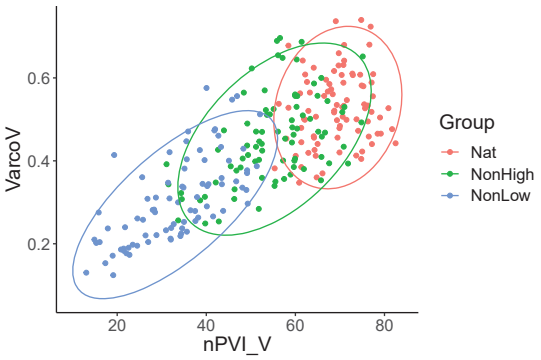



Figure 1: *Scatterplot of nPVI_V against VarcoV of three groups*

analysis. In the regression model, all significant variables listed in Table 3 were entered as independent variables and group as binary dependent variable. Two binominal logistic regressions were performed, one for comparison between *EnHi-EnLo* and another between *EnNa-EnHi*. The number of powerful predictors were further reduced, and only the significant predictors were presented in Table 4 as results of the logistic regressions. It is clear that compared with *EnHi*, *EnLo* is associated with higher *%V* and *sumS(s)*, and lower *nPVI_V*; while compared with *EnNa*, *EnHi* is related to higher *VarcoC* and lower *nPVI_V*.

Table 4: *Results of logistic regression for two comparisons.*

| **EnLo (vs. EnHi)** | | | |
|---|---|---|---|
| Predictors | Coefficient | Z-value | p-value |
| %V | 0.085 | 2.946 | 0.003** |
| nPVI_V | -0.130 | -5.139 | 2.76e-07 *** |
| sumS (s) | 1.944 | 2.223 | 0.026 * |
| **EnHi (vs. EnNa)** | | | |
| Predictors | Coefficient | Z-value | p-value |
| nPVI_V | -0.218 | -6.490 | 8.59e-11 *** |
| VarcoC | 3.359 | 2.403 | 0.016 * |

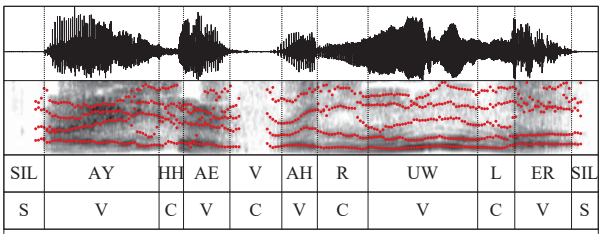
Though *VarcoV* could distinguish different groups when statistics were conducted separately, it did not show a significant difference when combined with other variables. While *%V* and *sumS* could distinguish lower- from higher-level L2 learners, but not higher-level L2 from native speakers because only some lower-level learners used epentheses with increased *%V* and made longer pauses with increased *sumS*. However, it is easier to decrease the occurrences of epenthesis and pauses with appropriate phonetic training [29], it is more difficult to train L2 learners to produce stressed-unstressed contrast [30].

When the higher-level L2 learners approached the native speakers in the rhythmic pattern, they still spoke slower than the native speakers. Therefore, a higher *VarcoC* also distinguished *EnHi* from *EnNa* because *VarcoC* is negatively correlated with speech rate in stress-timed languages [31, 25].

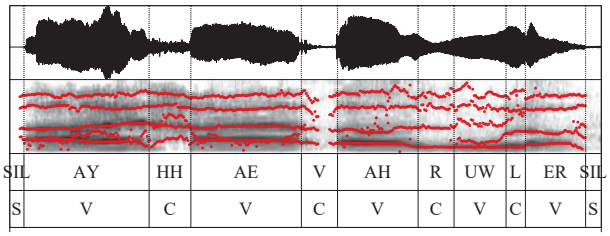
### 3.4. Representing examples for rhythmic metrics

Among these significant predictors, it is easy to understand the disturbance of speech rhythm by long pauses (*sumS*). *nPVI_V* variations can also be observed in the annotation where durational differences of two neighbouring vocalic intervals (VIs) are larger for higher-level learners and shorter for lower-level ones. One phrase "*I have a ruler*" is compared between two learners with score 5 and 2, respectively. In Figure 2 (a), Vs in *I* and r*u*ler, which are represented by *AY* and *UW* are longer than the other three Vs in h*a*ve, *a*, and rul*er*, which are represented by *AE*, *AH* and *ER*; while in Figure 2 (b), all the five Vs are comparably long, with the next one slightly shorter than the previous one, resulting in a smaller *nPVI_V* than that in Figure 2 (a). (Note: the "computer-friendly" ARPABET notation [32] is used here to facilitate corpus processing.)

Epenthesis can also be shown in the annotations, where phrase "*I'd like (some)*" is compared between a native speaker and a lower-level learner. The native speaker demonstrated a longer voiced /d/(D) to /l/(L), which could not be perceived as an audible epenthesis in Figure 3 (a); while in Figure 3 (b) the lower-level learner inserted a long epenthesis (annotated as *+AH*) after both syllable-final stops /d/ and /k/, which highly increased the percentage of vocalic durations (%V).

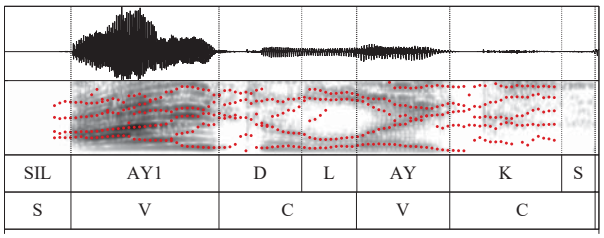



(a) One higher-level learner of larger *nPVI_V* value

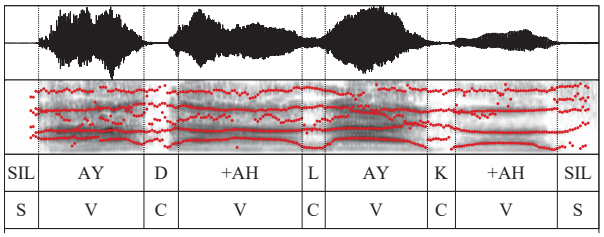



(b) One lower-level learner of smaller *nPVI_V* value

Figure 2: *Waveform and annotation of phrase "I have a ruler"*




(a) One English speaker without epenthesis




(b) One Chinese speaker with epenthesis

Figure 3: *Waveform and annotation of phrase "I'd like (some)"*

## 4. Discussion

Though we employed child speech and different materials, the relevant metrics are comparable to those reported in the previous investigations [5, 10, 12, 21, 22, 23, 25, 31, 33, 34], with slightly higher values of *nPVI_V* and *%V* for our English native speakers. The reason might be that the native speakers exaggerated the stress contrast, and prolonged stressed vowels and made few reductions for the L2 learners to follow, thus increasing pairwise vocalic index and vocalic percentage. Our investigation also confirmed the findings that vowel-related metrics are better indicators than consonant-related metrics for distinguishing stress-timed and syllable-timed languages, which supports the results of some previous research [16, 35].

As we know that in stress-timed languages, inter-stress intervals are far from equal. Furthermore, there is no fixed stress pattern for a sentence, and we could not know if the stress or pause made at the appropriate place on the basis of acoustic statistics without semantic analysis. However, the fact that stressed and unstressed syllables appear alternatively in stress-timed languages enables *nPVI_V* to be a good predictor for rhythms of stress-timed versus syllable-timed. One prominent feature of Chinese-accented English is that Mandarin ESL learners do not lengthen the stressed vowels and do not reduce the unstressed vowels, which can best be captured by the pairwise variability index for vocalic intervals (*nPVI_V*).

A special difficulty for school-aged children is that they may differ in chunks planning due to immature articulatory timing control [36]. Therefore, unexpected long pauses within short sentences (*sumS*) is also a good predictor for non-fluency for lower-level ESL learners. After they have overcome the difficulties with epenthesis and unnecessary pauses and progressed to a higher level, the rhythmic stress pattern still remains to be an obstacle for them to achieve native-like prosody, which is characterized with a lower *nPVI_V* with a higher *VarcoC*.

There is still much room for future improvements of our current study. Firstly, the speech data were collected online, only rough demographic information of the learners could be obtained. Studies with more control on the ESL speakers could be followed for more accurate findings. Secondly, only pure acoustic analysis was conducted. In the future research, we could combine phonetic-acoustic investigations with linguistic information, which should be able to provide improved predictions. For example, it was found that many speakers with full scores also made pauses, but their pauses were found between phrases, which were regarded as appropriate and did not destroy fluency. Therefore, calculations of pause-related indexes combined with annotations of syllable and phrase boundaries can improve the accuracy of prediction. And the same approach can also be applied to vowel-related variables to evaluate where stress is inappropriate. The contribution of this paper is that we have built a framework for L2 rhythm prediction, where many optimizations can be expected.

We have observed that some L2 children could demonstrate near-native rhythm, and recent findings have also shown that bilingual children are able to correctly assign word stress in both languages [37]. This means that CAPT systems for primary school children with an emphasis on prosody should be very promising, and our current work should also be rewarded.

## 5. Conclusion

In this study, we employed child speech to investigate durational variables of L2 English by Mandarin ESL learners and found that normalised pairwise variability index for vocalic intervals is a robust significant predictor for their stress-timed/syllable-timed rhythm, which can account for their prosodic performances in L2 English. These findings may provide implications for automatic prosodic evaluation and CAPT development. In the future, we will use more speech data to optimize our prediction model and extend our model to incorporate pitch and intensity patterns to represent speech prosody comprehensively.

## 6. Acknowledgements

The work was jointly supported by Shanghai Social Science project (2018BYY003), and the Major Programs of National Social Science Foundation of China (18ZDA293, 15ZDB103 and 13&ZD189).

# 7. References


[1] O. Kang, D. Rubin, and P. Lucy, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *The Modern Language Journal*, vol. 94(4), p. 554–566, 2010.

[2] T. Derwing, M. Munro, and G. Wiebe, "Pronunciation instruction for "fossilized" learners: Can it help?" *Applied Language Learning*, vol. 8, pp. 217–235, 1997.

[3] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL Quarterly*, vol. 39, pp. 399–423, 2005.

[4] Y. Saito and K. Saito, "Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners," *Language Teaching Research*, vol. 21, no. 5, pp. 589–608, 2017.

[5] P. P. Mok and V. Dellwo, "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English," in *Speech prosody 2008*, P. A. Barbosa, S. Madureira, and C. Reis, Eds., 2008, pp. 423–426.

[6] H. Kallio, A. Suni, J. Šimko, and M. Vainio, "Analyzing second language proficiency using wavelet-based prominence estimates," *Journal of Phonetics*, vol. 80, pp. 1–12, 2020.

[7] T. Kato, Q.-T. Truong, K. Kitamura, and S. Yamamoto, "Referential vowel duration ratio as a feature for automatic assessment of L2 word prosody," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6595–6599.

[8] C. Davis and J. Kim, "Characterizing rhythm differences between strong and weak accented L2 speech," in *Interspeech*, Hyderabad, 2018, pp. 2568–2572.

[9] L. Polyanskaya and M. Ordin, "The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language," *Applied Psycholinguistics*, vol. 40, pp. 795–819, 2019.

[10] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, pp. 265–292, 1999.

[11] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton, 2002, pp. 515–546.

[12] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, pp. 501–522, 2007.

[13] H. Wang, P. Mok, and H. Meng, "Capitalizing on musical rhythm for prosodic training in computer-aided language learning," *Computer Speech and Language*, vol. 37, pp. 67–81, 2016.

[14] T. V. Rathcke and R. H. Smith, "Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies," *The Journal of the Acoustical Society of America*, vol. 137, pp. 2834–2845, 2015.

[15] M. Ordin and L. Polyanskaya, "Perception of speech rhythm in second language: the case of rhythmically similar L1 and L2," *Frontiers in Psychology*, vol. 6, no. 316, 2015.

[16] R. S. K. Tan and E.-L. Low, "Rhythmic patterning in Malaysian and Singapore English," *Language and Speech*, vol. 57, no. 2, pp. 196–214, 2014.

[17] J. Krivokapić, "Rhythm and convergence between speakers of American and Indian English," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013.

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

[19] H. Ding and R. Hoffmann, "A durational study of German speech rhythm by Chinese learners," in *Speech Prosody 2014*, 2014, pp. 295–299.

[20] J. Anderson-Hsieh and H. Venkatagiri, "Syllable duration and pausing in the speech of Chinese ESL speakers," *TESOL Quarterly*, vol. 28, pp. 807–812, 1994.

[21] M. Ordin and L. Polyanskaya, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 533–544, 2015.

[22] P. M. Carter, "Quantifying rhythmic differences between Spanish, English, and Hispanic English," in *Theoretical and experimental approaches to Romance linguistics: Selected papers from the 34th linguistic symposium on romance languages*, R. S. Gess and E. J. Rubin, Eds. Amsterdam, The Netherlands: John Benjamins Publishing Company, 2005, pp. 63–75.

[23] H. Lin and Q. Wang, "Vowel quantity and consonant variance: A comparison between Chinese and English," in *Proceedings of Between stress and tone*, Leiden, 2005.

[24] U. Gut, "Prosody in second language speech production: The role of the native language," *Fremdsprachen Lehren und Lernen*, vol. 32, pp. 133–152, 2003.

[25] N. Whitworth, "Speech rhythm production in three German–English bilingual families," in *Leeds working papers in linguistics and phonetics*, D. Nelson, Ed., 2002, pp. 175–205.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2019, version 6.1, retrieved 31 July, 2019. [Online]. Available: http://www.praat.org

[27] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *Journal of the Acoustical Society of America*, vol. 32, no. 6, pp. 693–703, 1960.

[28] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018, version 3.5.1, retrieved on 23 September, 201. [Online]. Available: http://www.R-project.org

[29] H. Ding and R. Hoffmann, "An investigation of vowel epenthesis in Chinese learners' production of German consonants," in *Proceedings of Interspeech*, 2013, pp. 1007–1011.

[30] H. Ding and X. Xu, "L2 English rhythm in read speech by Chinese students," in *Interspeech 2016*, 2016.

[31] V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2003, pp. 471–474.

[32] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in speech recognition*, W. A. Lea, Ed. Prentice-Hall Englewood Cliffs, NJ, 1980.

[33] U. Gut, *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties*, ser. English Corpus Linguistics. Peter Lang GmbH, 2009, vol. 9.

[34] A. Arvaniti, "Rhythm, timing and the timing of rhythm," *Phonetica*, vol. 66, p. 46–63, 2009.

[35] V. Dellwo, "Rhythm and speech rate: A variation coefficient for deltac," in *Language and language processing*, P. Karnowski and I. Szigeti, Eds. Frankfurt: Peter Lang, 2006, pp. 231–241.

[36] M. A. Redford, "Grammatical word production across metrical contexts in school-aged children's and adults' speech," *Journal of Speech, Language, and Hearing Research*, pp. 1–16, 2018.

[37] D. Zembrzuski, M. Marecka, A. Otwinowska, E. Zajbt, M. Krzemiński, J. Szewczyk, and Z. Wodniecka, "Bilingual children do not transfer stress patterns: Evidence from suprasegmental and segmental analysis of L1 and L2 speech of Polish–English child bilinguals," *International Journal of Bilingualism*, vol. 24, no. 2, pp. 93–114, 2020.