INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

L2 English Rhythm in Read Speech by Chinese Students

Hongwei Ding[1], *Xinping Xu*[2]

[1]Centre for Cross-Linguistic Processing and Language Cognition
School of Foreign Languages, Shanghai Jiao Tong University, China
[2]Shanghai East High School, China

hwding@sjtu.edu.cn, tuesday608@163.com

Abstract

L2 English speech produced by Mandarin Chinese speakers is usually perceived to be intermediate between stress-timed and syllable-timed in rhythm. However, previous studies seldom employed comparable data of target language, source language and L2 interlanguage in one investigation, which may lead to discrepant results. Thus, in this study we conducted a contrastive investigation of 10 Chinese students and 10 native English speakers. We measured the rhythmic correlates in passage readings of Mandarin and L2 English produced by the native Chinese subjects, and those of English by the native British speakers. Comparison of the widely used rhythmic metrics *%V*, ΔC, ΔV, *nPVI*, *rPVI*, *VarcoV*, and *VarcoC* confirmed that Mandarin Chinese is a highly syllable-timed language. Results suggested that vowel-related metrics were better indexes to classify L2 English rhythm produced by Chinese speakers as being more syllable-timed than stress-timed. Analysis showed that vowel epenthesis, non-reduction of vowels, and no stressed/unstressed contrast could contribute to the auditory impression of syllable-timed rhythm of their L2 English. This investigation could shed some light on the Chinese accent of L2 English and provided support to facilitate the rhythmic acquisition of stress-timed languages for Chinese students.

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

1. Introduction

Rhythmic mistakes may lead to foreign accent of L2 speech, but can be successfully improved through individual training. Regarding the large number of Chinese students, individual training courses are impossible for most of the students in China. Thanks to the maturing of speech technology, Computer-Aided Pronunciation Tutoring (CAPT) programs can meet the requirements to help Chinese students to improve their rhythm of L2 English [1]. 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 rhythmic mistakes in L2 English produced by Chinese speakers.

Traditionally, languages have been classified as *stress-timed* and *syllable-timed* in rhythm. Since a large amount of experiments carried out in the 1970s and 80s to provide direct correlates for the isochrony in languages remained without success [3], other rhythmic indexes were developed. Ramus et al. showed that stress-timed languages have a higher standard deviation of consonantal intervals (ΔC) and relatively lower proportion of the vocalic intervals (*%V*); while syllable-timed languages have a lower ΔC and a higher *%V* [3]. Grabe and Low found that stress-timed languages have a higher variation in vowel durations, whereas syllable-timed languages show a lower variation in vowel length [4]. Barry et al. [5] and Dellwo and Wagner [6] found that ΔC correlates negatively with speech rate in stress-timed languages, and White and Mattys [7] proposed to use a rate-normalized metric (*VarcoV*).

These metrics have become the most widely used rhythmic indexes in classifying languages of different rhythms in many investigations in [8], [9], [10], [11], [12], and [13]. It is 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.

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 ([14], [1], [7], [15], [2], [16], and [17]). However, different findings can be found as to which rhythmic indexes or which combination of indexes can best distinguish different categories of languages or different varieties within a language ([13] and [1]). As Arvaniti [13] pointed out that these rhythmic metrics are largely influenced by inter-speaker variation, elicitation, and syllable structures of the materials, which was also confirmed in our previous investigation in [18]. The results suggest that rhythmic classification on the basis of comparison results across different studies may not be reliable. To minimize the undesirable influences, we conducted an investigation with the same English reading material for both native and L2 speakers, and employed the same Chinese subjects to produce L1 Mandarin and L2 English.

The aim of this paper is not only to identify which rhythmic indexes can best classify the rhythms of L1 Mandarin, L2 Chinese English and native English, but also to find out the possible reasons that have triggered the rhythmic deviation in L2 Chinese English from that of the native English.

2. Method

2.1. Subjects

We recruited 10 native Chinese speakers (5 males and 5 females) who grew up in Shanghai in order to minimize the influence of dialects and learning environments on the investigation. They were undergraduate students at Tongji University in Shanghai with an age range between 19 to 23 (20.3 in average). They began learning English from middle school and had been learning English for about 10 years. Their English teachers had been Chinese native speakers. None of the subjects had been to English-speaking countries before, but all of them had passed College English Test Band 4, which is a prerequisite for a bachelor's degree of non-English majors. Since most of the textbooks they used in learning English at high school were Oxford

http://dx.doi.org/10.21437/Interspeech.2016-427

edition, in which the recordings were in Received Pronunciation (RP), we assumed that British rhythm was the target for them to achieve. All the subjects spoke Mandarin, while their Chinese accents were discernible in their L2 English. These Chinese participants formed a homogeneous group in terms of age, L1 background, proficiency, and so on. They represented the average level of Chinese students in L2 English.

2.1.1. Material

In order to make sure that the materials consist of comparable syllable structures and to facilitate the metrics measurements with fewer pauses and hesitations, read speech was selected instead of spontaneous speech in this study. The reading stimuli or texts were based on MULTEXT [19], which is a most notable corpora designed for prosody research. MULTEXT consists of 40 different passages of five thematically connected sentences. The Chinese version was translated from the English texts and adequately adapted [19]. In recent constructed Open Multilingual Prosody Database (OMProDat) [20], the Chinese version had further been modified to facilitate reading [21]. Ten Chinese speakers were recruited to read both the English and Chinese versions of the passages. The English data were taken from OMProDat, which were produced by native British speakers reading the same English passages [20]. Only one passage in English and its corresponding Chinese version were selected for the metrics comparison in the current investigation. The data consists of 3 sets with 5 sentences from each 10 speakers:

1. *Native English* produced by 10 native British speakers.
2. *L2 Chinese English* produced by 10 Chinese speakers.
3. *Mandarin Chinese* produced by the same 10 Chinese speakers.

We obtained 20 English passages (each with 85 syllables), and 10 Chinese passages (each with 94 syllables). Since every passage consists of 5 sentences, the comparison of the rhythmic metrics were based on 150 sentences in total.

2.1.2. Recording

The speech data collection of the Chinese participants was carried out in the recording room at Tongji University in Shanghai. The speakers were asked to familiarize themselves with a printed copy of the sentences before the recording. During the recording the participants were asked to read aloud at their natural pace, and they were also required to repeat any misread or disfluent passages before moving on, so that unnatural final lengthening seldom occurred. Other hesitation disfluencies such as false starts or pauses, which can be separated from speech were excluded in the measurement. The data were recorded at 44.1kHz sampling rate and 16-bit quantization.

2.1.3. Analysis

In order to ensure comparability, the annotation technique used by Ramus [3] was adopted. After the wave files had been automatically labelled with SPPAS [22], annotation was conducted in the following two steps on Praat [23]: 1) *phonetic segmentation* of the sentence into phonemes, and 2) *classification* of separate phonemes into vowels and consonants.

In the first step, following the standard of phonetic criteria [24], the authors corrected automatic annotation manually as accurate 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, phonemes were then classified as vowels or consonants. In order to ensure comparability, the annotation technique of consonant and vowel intervals used by Ramus [3] was adopted: checked (free) vowels, free (long) vowels and unstressed schwa /@/ were coded as V (vowel); plosives, affricates, fricatives, sonorants (nasal 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 Chinese learners. Short pauses before the burst of stops and nasals were labeled as closure part of the corresponding phoneme. If there were some pauses and hesitations, which could not be identified as part of a sound, these breath parts were then marked as "#". Any two consonantal intervals split by "#" (pauses or hesitations) were combined into the same consonantal interval in calculation by subtracting the duration of pause or hesitation. The same approach was used for vowel intervals as well.

We measured the duration values of V and C:

- *V (vocalic intervals)* : the duration of sequences of consecutive vowels;
- *C (consonantal intervals)*: the duration of sequences of consecutive consonants.

The classification procedure can be observed in Figure 1 with the phoneme annotation in the first tier and the classification of vocalic and consonantal intervals in the second tier for the phrase *give me a firm date* segmented from the speech data.

From the measurements we calculated the following relevant variables of each speaker [7]:

- ΔV: the standard deviation (STDEV) of vocalic interval (VI) duration.
- ΔC: the STDEV of consonantal interval (CI) duration.
- $\%V$: the sum of VI duration divided by the total duration of VIs and CIs and multiplied by 100.
- *rPVI*: the raw Pairwise Variability Index (PVI) for CIs.
- *nPVI*: the normalised PVI for VIs.
- *VarcoV*: the STDEV of VI duration divided by the mean VI duration and multiplied by 100.
- *VarcoC*: the STDEV of CI duration divided by the mean CI duration and multiplied by 100.

The duration values were extracted with praat scripts and the above measurements of metrics were then calculated.

3. Results

3.1. Rhythmic metrics

The overview of the metrics averages are described in Table 1.

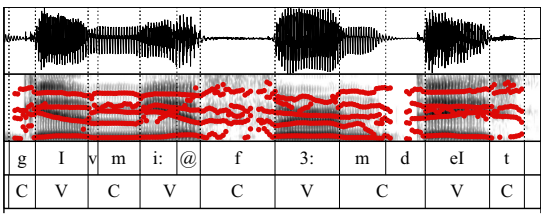


Figure 1: *Segmentation of consonantal and vocalic intervals*

Table 1: Means of rhythmic metrics for English (EE), L2 Chinese English (CE), and native Chinese (CC)

Language	Scores						
	%V*100	ΔC*100	ΔV*100	rPVI	nPVI	VarcoV	VarcoC
EE	38.91	4.86	4.02	55.67	66.74	61.94	50.47
CE	47.13	6.35	5.34	71.62	48.81	47.20	51.48
CC	54.30	3.66	4.77	45.58	48.20	45.56	42.01

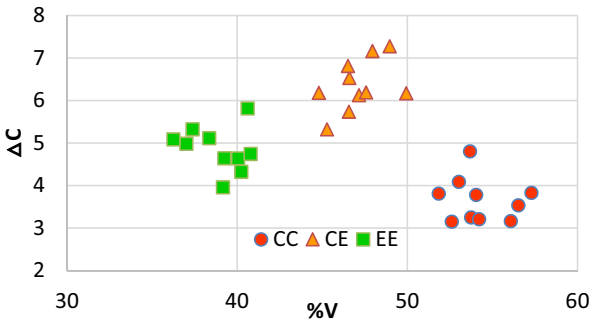


Figure 2: *Values of %V and ΔC of CC, CE, and EE*

The relationship between %V and ΔC, rPVI and nPVI, and VarcoV and %V can be observed in Figure 2, 3, and 4 respectively. All the rhythmic metrics are the averages of 5 sentences of each speaker for native English (EE), L2 Chinese English (CE), and native Chinese (CC).

It can be observed in Figure 2 – 4 that all metrics can distinguish EE and CC, but CE is observed to be intermediate between EE and CC only according to the values of %V, nPVI and VarcoV. %V shows CE lies halfway between EE and CC, while nPVI and VarcoV demonstrate that CE is more like CC than EE.

3.2. Correlation with duration

It is clear that these Chinese learners spoke much slower and made more pauses than the native speakers. The average duration values across the 5 English sentences produced by each speaker of EE and CE are presented in Figure 5, in which speakers 1–10 are British (EE) and speakers 11–20 are Chinese (CE).

The average duration is divided into two parts: (1) *durations without pauses*, which include vocalic and consonantal intervals, and they were employed for the calculation of *%V* and ΔC; (2) *pauses*, which are the silent or breath periods existing

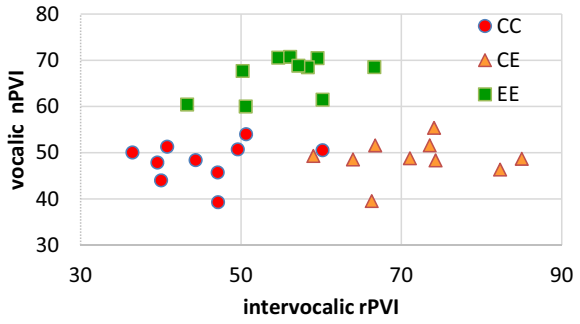


Figure 3: *Values of rPVI and nPVI of CC, CE, and EE*

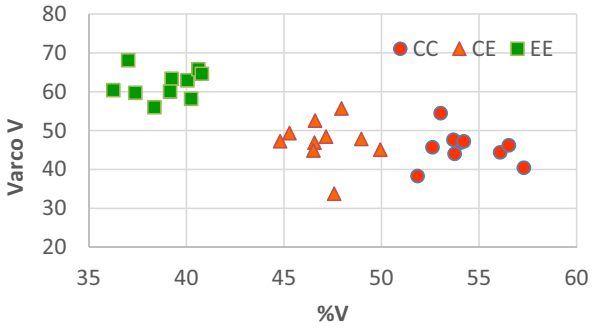


Figure 4: *Values of VarcoV and %V of CC, CE, and EE*

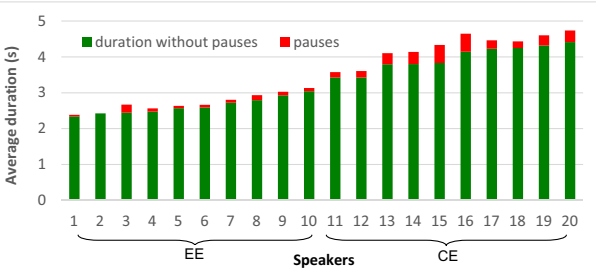


Figure 5: *Average duration values of all speakers*

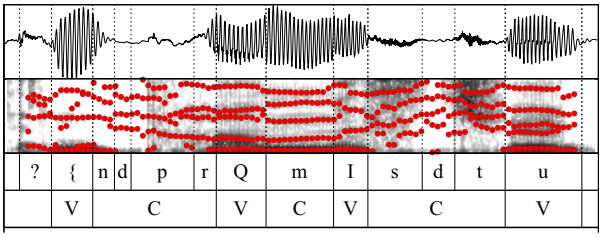
within in the sentences, but were excluded from the calculation. However, these pauses can also reflect the speech rate and the fluency of the learners. One British speaker made no pauses in reading all the sentences and other 9 speakers made one or two pauses after commas. All Chinese learners made more or less pauses: some at appropriate places between prosodic words and some at inappropriate places within the prosodic word.

We further calculated the correlation of the average duration with all the metrics in Table 2.

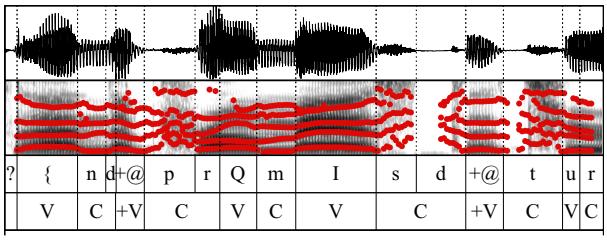
Table 2: Correlation of duration with metrics: [-] = not significant, [*] = $p < 0.05$, [**] = $p < 0.01$, [***] $p < 0.001$. Duration: duration without pause; D+P: duration with pause.

	EE		CE		CC	
	Duration	D + P	Duration	D + P	Duration	D + P
Interval measures						
%V	-	-	-	-	-	-
ΔC	***	***	**	**	**	**
ΔV	-	-	-	-	**	**
VarcoV	-	-	-	-	-	-
VarcoC	***	***	***	***	***	***
Pairwise variability indices						
rPVI	***	***	***	***	-	*
nPVI	-	-	-	-	-	-

It can be obviously observed that *%V*, *VarcoV*, and *nPVI* have no correlation with duration; most values of ΔC, *VarcoC* and *rPVI* have correlation with duration; ΔV have no correlation with EE or CE but some correlation with CC. It seems that vowel-related metrics have no correlation with speech rate.

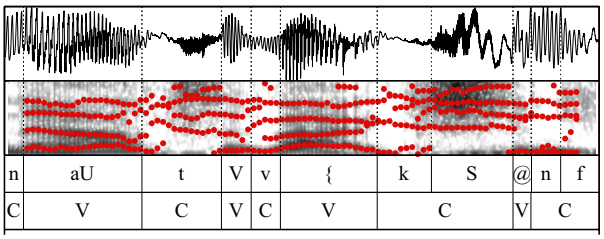


(a) *One English speaker without epenthesis*

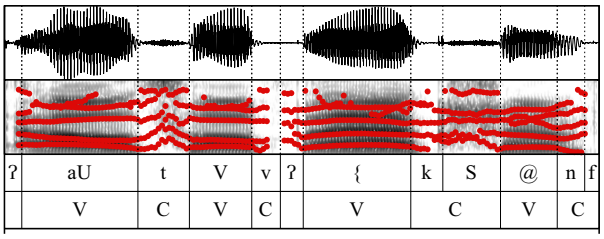


(b) *One Chinese speaker with epenthesis*

Figure 6: Waveform and annotation of phrase '*and promised to*'



(a) *One English speaker with more stressed/unstressed contrast*



(b) *One Chinese speaker with less stressed/unstressed contrast*

Figure 7: Waveform and annotation of phrase '*out of action*'

3.3. Vowel-related phenomena of CE

One prominent feature of CE is the occurrence of epenthesis. It is a common phenomenon that Chinese learners add vowels, especially schwa after consonant finals, which creates additional syllables for them to produce. English speakers do not release a burst of the stop /d/ when it is followed by another stop /p/ or /t/, as shown in Figure 6 (a), and no epenthesis can be identified between these two stops. While most Chinese speakers do explode the first stop /t/, resulting in an epenthesis (@) after *and* and *promised*, as shown in Figure 6 (b). Among the 10 Chinese speakers, 2 produced no epenthesis, the average was 3.2 epentheses per speaker in 85 syllables.

Another feature of CE is that unstressed vowels are seldom shortened or reduced, and stressed vowels are not lengthened. The stressed vowels /aU/ and /{/ are thus much longer than the unstressed vowels /V/ and /@/ produced by an English speaker in Figure 7 (a), while the vowels produced by the Chinese speaker have comparable durations in Figure 7 (b). The differences can thus be manifested in vowel-related indexes.

4. Discussion

As previously mentioned, this study aims to find out which indexes can best identify the intermediate position of Chinese English and which phenomena of L2 lead to that deviation.

- Though we employed a different reading passage other than "The North Wind and the Sun", all our metrics fall in the reasonable areas reported in the previous investigation ([14], [25], [1], [7], [15], [2], [6], [17], and [3]).
- Our investigation shows that almost all the classical rhythmic metrics can distinguish English from Mandarin Chinese clearly, which reinforces the previous findings.
- As all the speakers read the same syllables, the longer the duration, the slower the speech rate. Therefore, a high positive correlation with duration means a high negative correlation with speech rate. Our investigation confirmed the findings that consonant-related metrics are influenced by speech rate. On the other hand, L2 learners demonstrated a lower speech rate in comparison with native speakers [26], so consonant-related indexes cannot be found intermediate between source and target languages. Vowel-related metrics are thus better indicators for rhythmic deviation, which supports the results of some previous researches ([11], [27])
- On the basis of this investigation we further point out that occurrences of epenthesis, non-reduction of unstressed vowels, and no contrast between stressed-unstressed vowels are responsible for the classification of L2 Chinese English as being syllable-timed rather than being stress-timed in rhythm. Though it is easier to decrease the occurrences of epenthesis with appropriate phonetic training [28], it is more difficult to train stressed-unstressed contrast. Even though a lower *%V* may not indicate a stress-timed L2 speech, *VarcoV* and *nPVI* can still capture the missing stress contrast. Therefore, in this investigation *VarcoV* and *nPVI* rather than *%V* demonstrated that L2 English produced by the Chinese subjects is more syllable-timed than stress-timed.

5. Conclusion

In this investigation we employed the widely used rhythmic metrics to investigate English, L2 Chinese English, and Mandarin Chinese with comparable reading materials. It was thus possible for us to identify better indexes for classifying the rhythmic pattern of L2 Chinese English and to find the possible reasons for the syllable-timed patterns. These findings can be employed for teaching of acquisition of near-native English rhythm for teachers and CAPT developers. In the future, we can recruit Chinese learners of different levels of rhythm to test whether these distinguishing metrics are proficiency dependent.

6. Acknowledgements

The work is sponsored jointly by the National Social Science Foundation of China (13BYY009, 10CYY009, and 13&ZD189) and the Interdisciplinary Program of Shanghai Jiao Tong University (14JCZ03). We thank Daniel Hirst for his praat scripts to calculate the values of all the rhythmic metrics.

7. References

[1] 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.

[2] 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.

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

[4] 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.

[5] W. J. Barry, B. Andreeva, M. Russo, S. Dimitrova, and T. Kostadinova, "Do rhythm measures tell us anything about language type?" in *Proceedings of the 15th International Congress of Phonetics Sciences*, Barcelona, 2003, pp. 2693–2696.

[6] 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.

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

[8] 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.

[9] T. V. Rathcke and R. H. Smith, "Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies," *J. Acoust. Soc. Am.*, vol. 137, pp. 2834–2845, 2015.

[10] 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.

[11] 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.

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

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

[14] M. Ordin and L. Polyanskaya, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 533–544, 2015.

[15] P. M. Carter, "Quantifying rhythmic differences between Spanish, English, & Hispanic English," in *Theoretical and experimental approaches to romance linguistics: Selected papers from the 34th linguistic symposium on romance languages (Current issues in linguistic theory 272)*, S. R. Gess and E. J. Rubin, Eds. Amsterdam, The Netherlands, Philadelphia, Pennsylvania.: John Benjamins, 2005, pp. 63–75.

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

[17] 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.

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

[19] M. Komatsu, "Chinese MULTEXT: Recordings for a prosodic corpus," *Sophia Linguistica*, vol. 57, 2010.

[20] D. Hirst, B. Bigi, H. Cho, H. Ding, S. Herment, and T. Wang, "Building OMProDat: an open multilingual prosodic database," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 2013, pp. 11–14.

[21] H. Ding and D. Hirst, "A preliminary investigation of the third tone sandhi in standard Chinese with a prosodic corpus," in *Proc. ISCSLP*, 2012, pp. 436–439.

[22] B. Bigi, "SPPAS - multi-lingual approaches to the automatic annotation of speech," *The Phonetician: a publication of ISPhS, International Society of Phonetic Sciences*, vol. 111-112, pp. 55–69, 2015.

[23] P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program]. Version 6.0.05, retrieved 01 January 2016. [Online]. Available: http://www.praat.org

[24] 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.

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

[26] H. Ding, O. Jokisch, and R. Hoffmann, "A phonetic investigation of intonational foreign accent in Mandarin Chinese learners of German," in *Proc. 6th International Conference on Speech Prosody*, Q. Ma, H. Ding, and D. Hirst, Eds., vol. 1, Shanghai, China, 2012, pp. 374–377.

[27] 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.

[28] 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.