9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Compressibility of Segment Duration in English and Chinese

Chengxia Wang[1], *Jinsong Zhang*[2], *Yi Xu*[1]

[1]University College London, UK
[2]Beijing Language and Culture University, China

chengxia.wang.15@ucl.ac.uk, jinsong.zhang@blcu.edu.cn, yi.xu@ucl.ac.uk

Abstract

This study is a reexamination of the rhythm class hypothesis through an investigation of isochrony tendency in English, an alleged stress-timed language, and Chinese, an alleged syllable-timed language. We compared the relationship between segment and syllable duration in a corpus from each language. The results show that the correlation of segment and syllable duration is close to 1 in English but much weaker in Chinese. This indicates that English segments are not compressible for the sake of equal syllable duration, while Chinese does have a weak tendency toward equal syllable duration. Combining evidence from other studies, we interpret the current finding as an indication that there is no tendency toward isochrony of stress intervals in English. In contrast, there is an isochrony tendency at both the syllable level and phrase level in Chinese. Compressibility of segments and syllables could therefore be a useful index of cross-linguistic typology of timing and rhythm.

Index Terms: isochrony, compensation, syllable duration, segment duration, rhythm class

1. Introduction

The dichotomy of stress-timed languages and syllable-timed languages is proposed by Pike [1] and Abercrombie [2], according to which languages of the world are either stress-timed or syllable-timed. In stress-timed languages, *inter-stress intervals* tend to be constant, hence, isochronous, whereas in syllable-timed languages, successive *syllables* tend to be equal in duration. English, Russian, Arabic are typical stress-timed languages, while French, Telugu, Yoruba are typical syllable-timed languages.

Following this dichotomy, in syllable-timed languages inter-stress intervals would tend to be longer in proportion to the number of syllables they contain, whereas such a tendency would be absent (or weaker) in stress-timed languages [3,4,5]. This further implies that in a stress-timed language every inter-stress interval should tend to be of the same duration irrespective of the number of syllables it contains [6]. In order to attain a constant duration of inter-stress intervals, the syllables within the interval should be able to be stretched or squeezed [7]. The shortening of syllable duration as the number of syllables increases is called compensation effect [8]. It is claimed that the compensation effect in English adjusts syllable durations within each stress foot so that the durations of a series of stress feet are more similar than they would be if duration were only dependent on the number of syllables or the segmental make-up of the syllables [7].

A great deal of experimental research has been carried out to find evidence for such isochrony. However, no clear evidence has been found [6,9]. Uldall [10] measured a recording of a passage of written English read by David Abercrombie and found that the measured inter-stress intervals did not show marked regularity. Peter Roach [6] tested 6 languages and his data showed that stress-timed languages exhibit a wide range of percentage deviations in inter-stress intervals. What is more, the duration of interstress intervals in English has been shown to be directly proportional to the number of syllables they contain [11,12,13,14].

In the face of the recurrent failures, some researchers explained the paradox of isochrony by positing a fundamental discrepancy between production and perception, and by attributing all the responsibility to the latter [9,15,16,17,18]. For example, Classe [19] found that the tendency towards isochrony was severely counteracted by the number of unstressed syllables between two stressed ones and the segmental composition of the syllable themselves. He therefore hypothesized that in perception the intervals might look more regular than they are acoustically.

The dichotomy of stress-timed and syllable-timed languages received renewed interest in the 1990s due to the proposal of the rhythm metrics, which use consonantal and vocalic variability to quantify the rhythm class of languages. The notable measurements are %V (the proportion of vocalic intervals in an utterance), ΔV (the standard deviation of vocalic intervals within an utterance), ΔC (the standard deviation of consonantal intervals within an utterance) from Ramus [20], VarcoC (Standard deviation of consonantal intervals divided by mean and multiplies 100), VarcoV (Standard deviation of vocalic intervals divided by mean and multiplies 100) from Dellwo [21,22], and the pairwise variability indices nPVI and rPVI (Pairwise Variability Index in their measurements on successive vocalic and intervocalic intervals) introduced by Grabe and Low [23].

Since then, a large number of studies have applied the rhythm metrics to different languages and even varieties of non-native accents [24,25,26,27,28]. On the other hand, criticisms of the rhythm metrics also followed, drawing evidence on the computations, their instability due to speech rate, speaking style, within-speaker variation and measurement uncertainty, and their failure to capture the true nature of speech rhythm [5,27,29,30,31,32]. They point out that although rhythm metrics do separate languages based on syllable-timing versus stress-timing to some extent, they do not demonstrate that the perceived differences are the result of either a rhythmic intent on the part of the speaker or a cyclicity underlying the process of speech production [5].

Among the criticisms of the rhythm metrics, however, one thing is conspicuously missing: Demonstration as to whether there is at least a tendency toward isochrony in the form of durational compensation. Such demonstration is critical because rhythm, after all, is about timing, even if it is

10.21437/SpeechProsody.2018-132

only about perceived timing as has been argued sometimes [27]. From the very beginning, the rhythm dichotomy is about regularity of timing, whether the alleged unit is the syllable or stress interval. In other words, timing is of the essence. Interestingly, a lack of any durational compensation had been shown by a number of studies well before the proposal of the rhythm metrics. Nakatani et al. [33] found a linear relationship between foot size and foot duration in reiterant English. Lea [34] also found a linear relationship between the number of intervening unstressed syllables and the interstress interval for real words in sentence context. These findings show that syllable duration is not compressed as the size of the stress interval increases.

Furthermore, there is evidence that there is no durational compensation of *segments* for the syllable in English. O'Connor [13] finds that syllables in English increase in duration as segments are added. Van Santen and Shih [35] find that syllable duration is highly predictable from segmental duration in English and in most cases the slope of the regression line is close to 1.0. Thus the findings of both [13] and [35] suggest that segments are not compressed for the sake of the syllable in English. Interestingly, however, in the same study Van Santen and Shih [35] find that there seems to be some segmental compensation for the syllable in Mandarin Chinese. But they did not make a strong conclusion about the difference between the two languages in this respect.

These durational studies therefore point out a way of testing the rhythm class hypothesis by directly examining its basic assumptions: how much a segment or a syllable can be compressed in a language as a function of the alleged rhythmic unit, syllable in syllable-timed languages, and stress group in stress-timed languages.

In the present study, we will investigate the temporal compressibility of segments at the syllable level in two languages: English as a stress-timed language and Chinese as a syllable-timed language [23,36,37]. We hypothesize that a) English segments are not compressible, thus replicating the finding of [35], and b) Chinese segments are compressible, thus reaffirming, in clearer terms, data reported in [35].

The current study is only a partial test of the rhythm class hypothesis, however, as a full test will also require testing the duration of syllables as a function of stress-interval size or phrase size. That has been done for English as mentioned earlier [11,12,13,14], and for Chinese [38], but need to be tested further in terms of compressibility in future research.

2. Method

2.1. Corpus

2.1.1. English Corpus

For American English, the Boston University Radio News Corpus was used [39]. It consists of news stories recorded by 7 (3 female and 4 male) FM radio news announcers associated with WBUR, a public radio station, during broadcast, and the same four type-B news stories recorded by 6 of the 7 announcers in a lab. Professional radio announcers tend to be more fluent than non-professional speakers, and they produce fewer disfluencies and fewer prosodic errors [39]. The paragraphs are annotated with orthographic transcriptions, phonetic alignments, part-of-speech tags and prosodic labels. The phonetic alignments are generated automatically using constrained speech recognition as described in [40]. Segmentation times and phone durations are provided in units of 10-msec frames. One problem was that words were divided into syllables based on a dictionary that combined MOBY and SRI dictionaries, which did not consider resyllabification. For example, the dictionary divided the word decade into "d eh+1 k" and "ey d". In spoken English, speakers tend to utter it as "d eh+1" and "k ey d", so that "k" is an onset. This may affect the result, so resyllabification was performed according to the following rule: if a coda is followed by a syllable beginning with a vowel, the coda is treated as an onset. For the current study, we applied this rule only within words. Annotation for the lab news portion of the corpus were hand-corrected, while those for radio news were not. All announcers' data except for M4B's, which lacks annotation, were analyzed. In total, 8,806 tokens of CV syllable, 12,702 tokens of CVC syllable and 2,656 tokens of CVCC syllable were analyzed .

2.1.2. Chinese Corpus

The Chinese data were from Annotated Speech Corpus of Chinese Discourse (ASCCD), which was set up and recorded at Institute of Linguistics, Chinese Academy of Social Sciences. There are 18 discourse structures, each containing 300-500 syllables and several paragraphs. Five male and five female Beijing speakers who speak standard Chinese read the discourses naturally [41]. Four layers, including the syllable tier, initial and final tier, break index tier and stress tier, were labeled. The corpus did not include the closure silence for stops and affricates [7]. Following conventions in speech synthesis, 80 ms were added to stops and affricates as the closure duration. In total, 41,673 CV syllables, 18,486 CVC syllables and 10,647 CGV syllables were analyzed.

2.2. Measurement

We followed the method in [35] to analyze correlation between average duration of segments (interpreted as estimates of intrinsic durations) and syllables.

For syllables of the type CV, DUR(c •) is the mean duration of all CV syllables starting with c, DUR(c|c •) is the duration of c averaged over all vowels, and D_{inherent}(v) is the inherent duration of a vowel. This method also works for vowels.

$$DUR(c\bullet) = \alpha DUR(c \mid c\bullet) + (1/V)\sum_{v=1}^{v=V} D_{inherent}(v) + \beta \tag{1}$$

The equation shows the compensation effect in a syllable, as it measures how much the duration of a consonant or vowel depends on the identities of the remaining segments in the syllable. When α is 1, there is no compensation, and when α is 0, there is complete compensation.

2.3. Analysis and Results

2.3.1. English Results

As shown in Figure 1, in CV syllables, syllable duration is highly related to the intrinsic durations of onset and nucleus, as indicated by the Pearson correlation coefficients of 0.906 (p < 0.001) and 0.954 (p < 0.001). The slopes were 1.009 and 1.13, respectively.

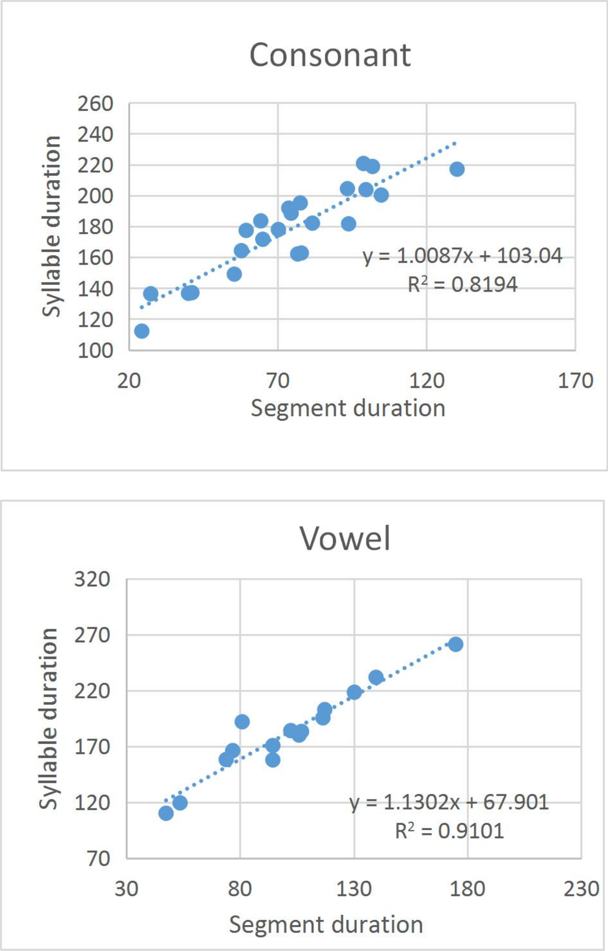

Figure 1: *Effect of consonant and vowel identity on syllable duration in CV syllables, for American English*

In CVC syllables, correlations between syllable duration and segmental durations were 0.915 (p < 0.001) for the onset consonant, 0.843 (p < 0.001) for the vowel, and 0.802 (p < 0.001) for the coda consonant. Slopes were 1, 0.917 and 0.778, respectively.

In CVCC syllables, correlations between syllable duration and segmental durations were 0.628 (p < 0.001) for the onset, 0.815 (p < 0.001) for the nucleus, 0.082 for the first coda and 0.258 for the second coda. Slopes were 0.95, 0.87, 0.27, 0.54, respectively.

2.3.2. Chinese Results

As shown in Figure 2, in CV syllables, syllable duration was closely related to the intrinsic durations of the onset and the nucleus, as measured by Pearson correlation coefficients of 0.893 (p<0.001) and 0.853 (p<0.001). The slopes were 0.725 and 0.756, respectively.

In CVC syllables, the coda is not analyzed, as the codas in Chinese share similar duration [35]. Correlations between syllable duration and segmental durations were 0.797 (p < 0.001) for the onset, 0.440 for the vowel. Slopes were 0.656 and 0.512, respectively.

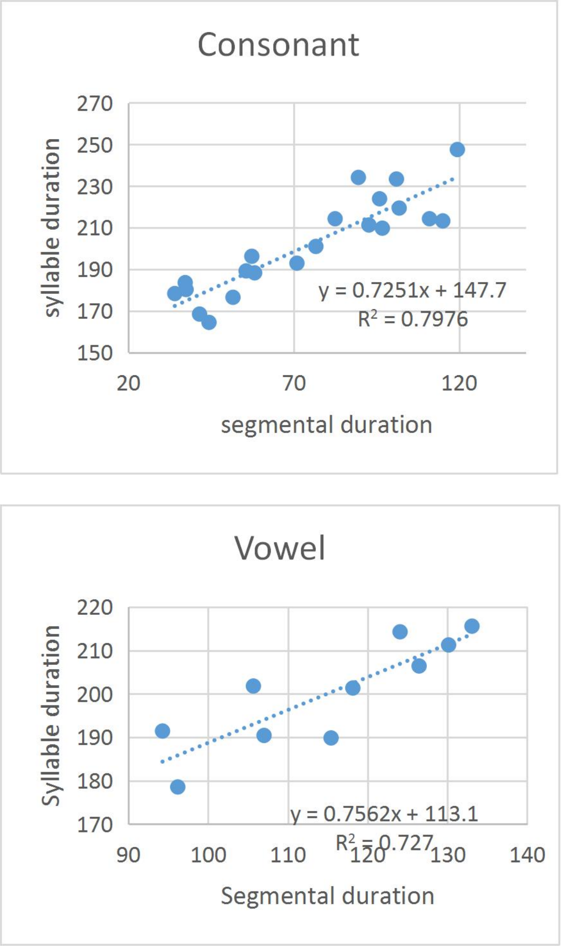

Figure 2: *Effect of consonant and vowel identity on syllable duration in CV syllables, for Chinese*

In CGV syllables, the glides were not analyzed because of their limited number. Correlations between syllable duration and segmental durations were 0.673 (p < 0.001) for the onset, -.041 for the nucleus. Slopes were 0.673 and -0.0255, respectively.

3. Discussion and conclusion

The English results show that α is close to 1 except for codas in CVCC. This is because 1) some codas only have few combinations, so syllable duration is strongly affected by the duration of onset or nucleus, 2) some unreleased stops did affect the syllable duration. If codas in CVCC are not considered, there is no duration compensation in English segments. This means that English segments are not compressed or stretched to make syllables equal in duration. Note that, this cannot be taken as evidence that English is not a syllable-timed language and so should be a stress-timed language. This is because, to realize stress-timing, syllables need to be flexible in duration so as to compensate for the changes in the number of syllables in the stress interval. In other words, compressibility is needed for both syllable timing and stress timing.

The Chinese results show compensation of segmental duration for the sake of syllable. Compared with the slops of regression lines in English, slopes in Chinese are shallower, which means that there is compensation of consonants and vowels in syllables and that Chinese has a tendency to compress segments to shorten an otherwise longer syllable. Although this might be seen as a sign of syllable timing, there is already evidence that syllable duration itself is compressible as a function of phrase size in Chinese [38]. That is, as the number of syllables in a phrase increases the duration of the phrase does not increase in proportion. Together with the present finding, it seems that unit duration in Chinese is compressible at both the segmental and syllabic levels.

The classic discussion of rhythm dichotomy is about whether syllables or interstress intervals are isochronous. To achieve isochrony, elements in syllables or interstress intervals need to be squeezed or stretched to make compensation possible. However, the present results demonstrate that English, a typical stress-timed language, does not compress segments to equalize syllable duration. This may further explain why no evidence of isochrony in English production has been found. Together with the findings mentioned earlier that there is a linear relationship between foot size and foot duration in English [32, 33], English does not even seem to have a tendency toward stress-timing. In contrast, given the compressibility of segments for syllable found in the present study and the compressibility of syllables for phrase [38], Chinese shows a tendency toward both syllable timing and phrase timing.

The finding of this paper is still preliminary, but it raises serious questions about the rhythm class hypothesis, either in the classic form or as measured by the rhythm metrics. At the same time, both the finding of the present study and those of many others may suggest that compressibility of segments and syllables could be a more useful index of cross-linguistic typology of timing and rhythm.

4. Acknowledgement

We would like to thank Elizabeth Eden and Hao Liu for their help on Python scripts.

5. References

[1] K. L. Pike, The intonation of American English, 1945.

[2] D. Abercrombie, Elements of General Phonetics, l967.

[3] D. Abercrombie, “Syllable quantity and enclitics in English,” na, 1964.

[4] D. Abercrombie, “A phonetician’s view of verse structure,” *Linguistics*, vol. 2, no. 6, pp. 5-13, 1964.

[5] F. Nolan and H. S. Jeon, “Speech rhythm: a metaphor?” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* vol. 369, no. 1658, pp. 20130396, 2014

[6] P. Roach, “On the distinction between ‘stress-timed’and ‘syllable-timed’languages,” *Linguistic controversies*, pp.73-79, 1982.

[7] C. Hoequist Jr, “Durational correlates of linguistic rhythm categories,” *Phonetica*, vol. 40, no. 1, pp. 19-31, 1983.

[8] C. A. Fowler, “Timing control in speech production,” *Indiana University Linguistics Club*, 1977.

[9] I. Lehiste, “Isochrony reconsidered,” *Journal of phonetics*, vol. 5, no. 3, pp. 253-263, 1977.

[10] E. T. Uldall, “Isochronous stresses in RP,” *Form and substance*, pp. 205-201, 1971.

[11] D. L. M. Bolinger, “Forms of English: Accent, morpheme, order,” *Harvard University Press*, 1965.

[12] W. A. Lea, "Prosodic aids to speech recognition: IV," *A general strategy for prosodically-guided speech understanding. Univac Report PX10791, St. Paul, Minnesota: Sperry Univac*, 1974.

[13] J. D. O’Connor, “The duration of the foot in relation to the number of component sound-segments,” *Progress Report*, vol. 3, pp. 1-6, 1968.

[14] Y. Shen and G. G. Peterson, “Isochronism in English,” Department of Anthropology and Linguistics, *University of Buffalo*, 1962.

[15] P. M. Bertinetto, “Reflections on the dichotomy ‘stress’ vs.‘syllable-timing’,” *Revue de phonétique appliquée*, vol. 91, no. 93, pp. 99-130, 1989.

[16] G. D. Allen, “Speech rhythms: Its relation to performance universals and articulatory timing,” *Journal of phonetics*, vol. 3, no. 2, pp. 75-86, 1975.

[17] G. D. Allen, “Formal and statistical models of speech timing: past, present and future” *Proc. IX Int. Cong. Phon. Sci*, 1979.

[18] I. Lehiste, "Temporal relations within speech units," *Proceedings of the ninth international congress of phonetic sciences*, vol. 3, pp. 247-254, 1979.

[19] A. Classe, “The rhythm of English prose,” *B. Blackwell*, 1939.

[20] F. Ramus, M. Nespor and J. Mehler, “Correlates of linguistic rhythm in the speech signal,” *Cognition*, vol. 73, no. 3, pp. 265-292, 1999.

[21] V. Dellwo and P. Wagner, “Relationships between rhythm and speech rate,” *15th International Congress of the Phonetic Sciences*, *Barc*elona, pp. 471-474, 2003.

[22] V. Dellwo, “Rhythm and speech rate: A variation coefficient for ΔC,” *Language and language-processing*, pp. 231-241, 2006.

[23] E. Grabe and E. L. Low, “Durational variability in speech and the rhythm class hypothesis,” *Papers in laboratory phonology 7,* pp. 515-546, 2002.

[24] J. Dankovičová and V. Dellwo, “Czech speech rhythm and the rhythm class hypothesis,” I*nternational Conference of Phonetic Sciences*, pp. 1241-1244, 2007.

[25] P. Mok, “On the syllable-timing of Cantonese and Beijing Mandarin,” *Chinese Journal of Phonetics*, vol. 2, pp. 148-154, 2009.

[26] F. Nolan and E. L. Asu, “The pairwise variability index and coexisting rhythms in language,” *Phonetica*, vol. 66, no. (1-2), pp. 64-77, 2009.

[27] A. Arvaniti, “The usefulness of metrics in the quantification of speech rhythm,” *Journal of Phonetics*, vol. 40, no. 3, pp. 351-373, 2012.

[28] E. O’Rourke, “Speech rhythm variation in dialects of Spanish: applying the pairwise variability index and variation coefficients to Peruvian Spanish,” *In Proc. Fourth Conf. on Speech Prosody*, 2008.

[29] D. Deterding, “The measurement of rhythm: A comparison of Singapore and British English,” *Journal of Phonetics*, vol. 29, no. 2, pp. 217-230, 2001.

[30] P. M. Bertinetto and C. Bertini, “On modeling the rhythm of natural languages,” *In Proceedings of the Fourth International Conference on Speech Prosody*, 2008.

[31] D. Gibbon, “Computational modelling of rhythm as alternation, iteration and hierarchy,” *In Proceedings of ICPhS,* vol. 15, 2003.

[32] R. A. Knight, R. A. “Assessing the temporal reliability of rhythm metrics,” *Journal of the International Phonetic Association*, vol. 41, no. 3, pp. 271-281. 2011.

[33] L. H. Nakatani, K. D. O’Connor and C. H. Aston, “Prosodic aspects of American English speech rhythm,” *Phonetica*, vol. 38, no. (1-3), pp. 84-105, 1981.

[34] W. A. Lea, “Isochrony and disjuncture as aids to syntactic and phonological analysis,” *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S33-S33, 1975.

[35] J. P. Van Santen, and C. Shih, “Suprasegmental and segmental timing models in Mandarin Chinese and American English,” *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1012-1026, 2000.

[36] H. Lin, and Q. Wang, “ Mandarin rhythm: An acoustic study,” *Journal of Chinese Language and Computing*, vol. 17, no. 3, pp. 127–140, 2007.

[37] P. Mok, and V. Dellwo, “Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese,

Beijing Mandarin and English," *In Proceedings of Speech Prosody,* vol. 4, pp. 423-426, 2008.

[38] Y. Xu and M. Wang, "Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin," *Journal of Phonetics,* vol. 37, pp. 502-520, 2009.

[39] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1-19, 1995.

[40] O. Kimball, M. Ostendorf, and I. Bechwati, "Context modeling with the stochastic segment model," *IEEE Transactions on signal processing*, vol. 40, no.6, pp. 1584-1587, 1992.

[41] A. Li, X. Chen, G. Sun, W. Hua, Z. Yin, Y. Zu, F. Zheng, and Z. Song, "The phonetic labeling on read and spontaneous discourse corpora," *In proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2000.