Rhythmic Patterns in Native and Nonnative Mandarin Speech

Wentao Gu[1, 2] *and Keikichi Hirose*[2]

[1] Research Center for Language Information Technologies, Nanjing Normal University, China
[2] Graduate School of Information Science and Technology, The University of Tokyo, Japan

{wtgu, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Rhythm plays an important role in the naturalness of speech. This study compared rhythmic patterns of Mandarin speech between native speakers and two groups of L2 speakers whose first languages were Cantonese and English, respectively. The study started from isolated words, but focused on continuous speech, for which eleven durational metrics were used as objective rhythm indicators. The results on continuous speech showed that nonnative Mandarin gave a quite similar rhythmic mode as native one in terms of rate-normalized/independent metrics, but shifted towards the stress-timed class in terms of raw metrics, regardless of the rhythmic class of the L1. This seems to conflict with the L1 transfer effect and the results for isolated words, but it coincides with auditory impression and can be explained by speech rate difference and the lengthening effects associated with the change in prosodic structure.

Index Terms: rhythm, durational metrics, nonnative speech, Mandarin, Cantonese, English

1. Introduction

As an important prosodic property, rhythm plays a key role in the naturalness of speech. There used to be a tradition of classifying spoken languages in the world into three rhythmic classes, i.e., 'stress-timed', 'syllable-timed', and 'mora-timed' [1-3]. This classification was based on the notion of isochrony, which assumes the existence of units of nearly equal duration in speech: syllables for syllable-timed languages such as French and Italian, inter-stress intervals for stress-timed languages such as English and German, and morae for mora-timed languages such as Japanese.

However, experimental studies have failed to find any acoustic evidence to support the existence of isochronous units [4-6]. Thus, the so-called isochrony is only an impressionistic property which correlates with a number of phonological aspects such as syllable structure, vowel reduction, and stress [4]. Instead of searching for isochronous units, many recent studies came to find out acoustic metrics that could correspond roughly to the auditory impression of rhythmic distinctions, by inspecting the durational variability of consonantal and vocalic intervals, such as ΔC (standard deviation of consonantal duration), %V (percentage of vocalic duration) [5], and PVI (Pairwise Variability Index) of vocalic and consonantal durations [6, 7]. Similar metrics were also applied to syllable duration [8, 9]. These studies showed that such durational metrics could categorize spoken languages into different rhythmic classes. Also, they suggested that the difference in rhythm was not categorical; instead, various languages could be on a continuum between extreme rhythmic patterns [4, 6].

It has been widely recognized that English is a typical stress-timed language [2], in which there is perceptually a roughly constant amount of time between successive stressed syllables, and relatively low %V, high ΔC, and high PVIs tend to be measured acoustically [5, 6]. To accommodate the stress-timed rhythm, there is a tendency for unstressed syllables to be shortened. In contrast, Cantonese is deemed a typical syllable-timed language [8], in which successive syllables perceptually have roughly constant duration, and relatively high %V, low ΔC, and low PVIs tend to be measured acoustically. Mandarin (in mainland China) also shows a syllable-timed pattern, at least in read speech, but it is less typical than Cantonese as evidenced by various acoustic measures [8].

Besides rhythmic studies for native speech, there are also a few studies on rhythmic patterns for nonnative speech, e.g., English L2 by Mandarin and Cantonese L1 speakers [9], and the effects of L1 on L2 for Dutch, English, and Spain speakers [10]. This is important as objective rhythmic metrics may be applied in computer-assisted language learning systems.

The findings of these studies, however, are not consistent. In some cases nonnative speech is rhythmically between L1 and L2, coinciding with the general hypothesis of L1 transfer effect, while in other cases nonnative speech is rhythmically almost identical to L1 or L2, or even shows a pattern overshooting L2 – hence opposite to the L1 transfer effect. It is not easy to interpret such inconsistencies, though speech rate variation and selective lengthening have been used for explanation [9, 10]. This may lead us to question the validity of these rhythmic metrics for nonnative speech.

Since there have been few rhythmic studies on nonnative Mandarin speech, we attempt to fill the gap. We selected two groups of L2 Mandarin speakers who were native in American English and Hong Kong Cantonese, respectively. The reason for selecting these two groups is that English and Cantonese represent two opposite rhythmic classes as we described above.

2. Speech Data

2.1. Comparison in Phonology

The three languages concerned here, i.e., Standard Mandarin, Hong Kong Cantonese, and American English, contrast sharply in the phonological structure. Mandarin and Cantonese are tone languages of monosyllabic nature in the morpheme sense, while English is a stress language of polysyllabic nature.

Both Mandarin and Cantonese have a very simple syllable structure (C)V(C), and each syllable has a tone. In addition to monosyllabic nature in the morpheme sense, Cantonese has a higher frequency of monosyllabic words than Mandarin, thus further enhancing its monosyllabic nature. Cantonese does not have a contrast in lexical stress while Mandarin does – there is no lexical stress in the phonological sense but there is a neutral tone functioning as an unstressed syllable which usually has a shortened duration.

English, on the contrary, is a polysyllabic language with a more complex syllable structure: with the use of consonant clusters, English syllables can be (C)(C)(C)V(C)(C)(C)(C). Also, English is a stress language, for which the syllables in a polysyllabic word differ in the degree of lexical stress.

10.21437/SpeechProsody.2014-106

2.2. Materials

Read speech was used in the present study for the purpose of a controlled comparison among three groups. Two sets of Mandarin speech materials were designed. Speech Material I consisted of disyllabic and trisyllabic words which were used for analysis of timing in isolated words, while Speech Material II consisted of short stories which were used for investigating rhythmic patterns in continuous Mandarin speech.

Speech Material I included disyllabic and trisyllabic words with neutral tone (T0). The corpus of disyllabic words included four tonal combinations, with four lexical tones in the former syllable and T0 in the latter. With 10 words designed for each disyllabic tonal combination, there were altogether 40 disyllabic words. For trisyllabic words, T0 could occur in the mid and/or the latter syllable, and thus the corpus of trisyllabic words included 16 + 16 + 4 = 36 tonal combinations with T0. With two words designed for each trisyllabic tonal combination, there were altogether 72 trisyllabic words.

Speech Material II included three short stories, each with a length of about 200 syllables. One story was 'North Wind and Sun' as widely used in phonetic study. All three stories were easily understood by L2 Mandarin learners at the intermediate or advanced level, as confirmed by the nonnative subjects.

2.3. Subjects and data collection

Three groups of subjects were recruited: MM, CM, and EM. The MM group, including 3 males and 3 females with an average age of 20, consisted of native speakers of Mandarin, who were undergraduate students majoring in broadcasting and hosting arts, all professional in Mandarin pronunciation. The CM group, including 1 male and 5 females with an average age of 22, consisted of L2 Mandarin learners who were born in Hong Kong and were native in HK Cantonese. The EM group, including 3 males and 3 females with an average age of 22.5, consisted of L2 Mandarin learners who were born in USA and were native in American English. The subjects in both CM and EM groups were L2 Mandarin learners at the intermediate or advanced level. They had studied Mandarin for at least two years, and had already passed Level 5 of HSK, the Chinese proficiency test.

Speech recording was conducted in a sound-proof room after the subjects had got familiar with the reading materials and made sure that they had known the exact pronunciations. The recording was done at each subject's normal speech rate and was monitored by the experimenter. Once there was a mispronunciation or disfluency, the subject would be asked to repeat recording the word or the utterance until success.

2.4. Data analysis

By visual inspection of the waveform and the spectrogram, speech materials were segmented and labelled manually. Each word in Speech Material I was segmented into syllables, while Speech Material II was segmented into consonants, vowels, and pauses. Any pauses in the speech, either silent or not, were excluded from our analysis. On the basis of acoustic instead of phonological criteria, glides were classified as vowels, because the formant transition between a glide and a vowel nucleus in Mandarin is usually smooth, without an abrupt change.

Unlike for Speech Material I, labelling of Speech Material II was conducted not only for segments and tones but also for break indices, following a ToBI-like approach [11]. Three layers of break indices were adopted: prosodic word boundary (B1), minor prosodic phrase boundary (B2), and major prosodic phrase boundary (B3). A prosodic word is a basic unit that is tightly integrated in prosody. B2 differs from B1 in that B2 is accompanied by one of the following perceivable features: a short pause, a pitch resetting, or a final lengthening. B3 tends to be accompanied by a much longer pause than B2.

The labelling of break indices was more subjective than that of segments and tones. To ensure the reliability, we asked four labelers to do the labelling independently after an iterative training and discussion (the cross-labeler consistency reached 96% at the end of this stage). After their labelling, the consistency across the four labelers turned out to be 86%. We then finalized the labelling by double checking those apparently inconsistent labels.

For Speech Material II, the durations of vocalic and consonantal (i.e., intervocalic) intervals were measured. A vocalic interval is the section between the onset and the offset of a series of connected vowels/glides, while a consonantal interval is the section between the onset and the offset of a series of connected consonants. For Speech Material I, we simply measured the durations of individual syllables.

The following seven rhythmic metrics based on vocalic and consonantal interval durations [5, 7, 12] were calculated for each subject, and then were averaged across all subjects.

ΔC: the standard deviation of consonantal durations
ΔV: the standard deviation of vocalic durations
%V: the proportion of vocalic durations in the speech
VarcoC: (ΔC / mean consonantal duration) × 100
VarcoV: (ΔV / mean vocalic duration) × 100

$$rPVI\_C = \left(\sum_{k=1}^{m-1} \left| d_{Con_k} - d_{Con_{k+1}} \right| / (m-1) \right)$$

$$nPVI\_V = 100 \times \left(\sum_{k=1}^{m-1} \left| \left(d_{Vow_k} - d_{Vow_{k+1}} \right) / \left(\left(d_{Vow_k} + d_{Vow_{k+1}} \right) / 2 \right) \right| \right) / (m-1),$$

Here, VarcoC and VarcoV are rate-normalized metrics, which were introduced because ΔC and ΔV were found to be negatively correlated with speech rate [10]. PVI indicates the absolute durational difference between each pair of successive units [6]. Two PVI values, i.e., raw and rate-normalized, were calculated for consonantal and vocalic intervals, respectively, as vocalic duration was more sensitive to speech rate while consonantal duration carried more language variability [7].

Following [8, 9], four syllabic metrics were also calculated on the basis of syllable duration, including ΔS, VarcoS, rPVI_S, and nPVI_S, which were defined in the same way as their counterparts for consonantal and vocalic durations.

3. Results

3.1. Timing in isolated words

We started our study from isolated words in Speech Material I. The average percentages of duration for T0 syllables in a word are shown in Table 1, where X indicates any of the four lexical tones. In the case of trisyllabic words with two T0 syllables, the percentages of all three syllables are presented. In all cases MM gives the minimal durational ratio of T0 (in the case of X+T0+T0 the mid syllable should be the shortest, and hence the ratio of the mid T0 is compared), indicating that nonnative speakers did not shorten T0 adequately and hence decreased the durational variation in a word. In other words, nonnative speakers did not differentiate unstressed syllables from others

appropriately. Thus, it is expected that the rhythmic patterns of their continuous speech might be shifted further towards the syllable-timed extreme, especially for Cantonese L1 speakers.

Table 1. *Percentages of duration of T0 syllable in a word.*

Group	X+T0	X+T0+X	X+X+T0	X+T0+T0		
				X	T0	T0
CM	46.0	26.1	33.6	32.7	33.8	33.5
EM	47.4	25.0	33.3	34.7	30.2	35.1
MM	38.6	23.3	30.0	37.6	28.0	34.4

3.2. Rhythmic measurements for continuous speech

We then investigated rhythmic patterns for continuous speech in Speech Material II, using all eleven metrics described above.

Table 2 gives the statistical results of comparing the means of various rhythmic metrics among three groups. The left shows the *p*-values for one-way ANOVA, while the right shows the results of Bonferroni post-hoc tests only for those metrics showing significant main effects. For ΔC, ΔV, rPVI_C, nPVI_V, and rPVI_S, there are main effects of group, and the differences are significant only between native and nonnative groups – there is no significant difference between CM and EM. Besides, there is a marginally significant effect of group for ΔS, but no significant effects for the other metrics.

Except for nPVI_V, all those metrics that differentiate native and nonnative speech are non-rate-normalized; most of the rate-normalized metrics, however, have not provided any distinction between native and nonnative speech. The average speech rates turned out to be 4.14, 4.16, and 5.05 syllables per second for CM, EM, and MM, respectively. There was a main effect of group ($p < 0.001$). Bonferroni post-hoc tests showed that there was no significant difference between CM and EM; there was, however, a significant difference between MM and CM at $p < 0.01$, as well as between MM and EM at $p < 0.001$. This not only verifies that native speech is generally faster than nonnative speech, but also shows that the two nonnative groups have almost equal language proficiency in Mandarin.

It has been shown that ΔC, ΔV, and rPVI_C are negatively correlated with speech rate [10]. Therefore, the observed differences in ΔC, ΔV, and rPVI_C between native and non-native speakers might be the results of speech rate differences. While using VarcoV instead of ΔV helps capture rhythmic patterns better, rate-normalization of consonantal metrics such as ΔC and rPVI_C removes phonotactic differences between languages and hence cannot discriminate rhythmic classes, as shown in [6, 10]. Hence, we adopt raw metrics for consonants and rate-normalized/independent metrics for vocalic intervals.

Table 2. *Significance levels for the differences in rhythmic scores between the three groups.*

Metric	ANOVA	Bonferroni post-hoc test		
		CM–EM	CM–MM	EM–MM
ΔC	0.010**	0.614	0.013*	0.005**
ΔV	0.041*	0.536	0.017*	0.057†
%V	0.475	–	–	–
VarcoC	0.489	–	–	–
VarcoV	0.564	–	–	–
rPVI_C	0.013*	1.000	0.062†	0.016*
nPVI_V	0.024*	0.847	0.208	0.023*
ΔS	0.058†	–	–	–
VarcoS	0.499	–	–	–
rPVI_S	0.037*	1.000	0.142	0.046*
nPVI_S	0.275	–	–	–

* indicates $0.01 < p < 0.05$, while ** indicates $p < 0.01$.
† indicates marginally significant.

To compare the results with previous findings, in Figure 1 we plot the average values of ΔC and %V for different languages, including Mandarin [13], Cantonese [8], and seven other languages [5]. It can be seen that native Mandarin and Cantonese should both be classified as syllable-timed, though the distance in %V from these two to the other four syllable-timed languages is even farther than between the four syllable-timed languages and the three stress-timed languages. For non-native Mandarin, CM and EM are quite close to each other. They share almost the same %V with MM, but both have a much larger ΔC than MM (even larger than the stress-timed languages), showing a tendency of shifting towards stressed-timed languages in the dimension of ΔC, regardless of the L1. On the whole, all the languages in Fig. 1 can be clustered into four sets in the ΔC vs. %V space; native Mandarin/Cantonese, as well as nonnative Mandarin, looks distinctly deviated from other languages, probably due to their monosyllabic nature.

In Figure 2, we plot the average values of nPVI_V and rPVI_C for different languages, including Cantonese [8] and five other languages [6]. It can be seen that native Mandarin and Cantonese should both be classified as syllable-timed (the position of Spanish in [6] is somewhat suspicious because it differs from the results in [10, 14] substantially). For non-native Mandarin, CM and EM are close to each other. They differ from MM mainly in rPVI_C, showing a tendency of shifting towards stressed-timed languages in the dimension of rPVI_C, regardless of the L1.

Among four syllabic metrics, Mok [8] found that rPVI_S gave the best separation between stress-timed and syllable-timed languages, ΔS gave the second best separation, while nPVI_S and VarcoS were not good separators. As shown in

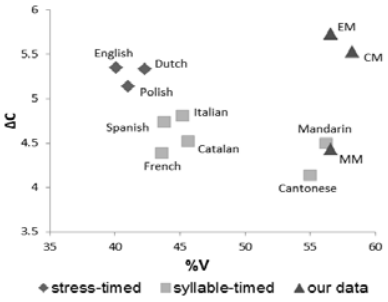

Figure 1: *ΔC and %V for different languages.*

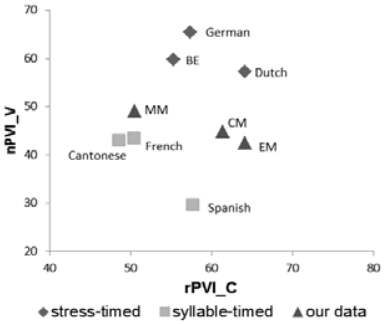

Figure 2: *nPVI_V and rPVI_C for different languages.*

Table 2, the syllabic metric that gives the best distinction among three groups is rPVI_S, followed by ΔS. This is in line with the result for separating stress-timed and syllable-timed languages in [8]. To compare with previous findings, Table 3 lists the average values of rPVI_S and ΔS, for three groups of subjects here and for six languages examined in [8] – English and German are stress-timed while others are syllable-timed.

We have noticed that MM in the present study and Mandarin in [8] give substantially different values. This is quite possibly due to speech rate difference between the materials used in the two studies – hence caution should be made in comparing absolute rhythmic scores between studies, as suggested in [10]. In spite of this, CM and EM showed a tendency of shifting from MM towards the stressed-timed class in terms of both rPVI_S and ΔS. This also coincides with the results for ΔC and ΔV (the latter is not illustrated here). It should be noted that rPVI_S, ΔS, ΔC, and ΔV showing the same tendency are all non-rate-normalized metrics.

Table 3. *Syllabic metrics for different languages.*

Language	rPVI_S	ΔS
English	115.50	88.74
German	99.62	80.78
Mandarin	86.08	75.80
Italian	82.68	67.61
French	75.89	55.30
Cantonese	63.62	57.48
EM	75.70	63.41
CM	72.44	60.32
MM	60.20	51.91

4. Discussion

Comparison of rhythmic characteristics for CM, EM, and MM on the basis of durational metrics has shown that nonnative Mandarin gives almost the same rhythmic mode (i.e., syllable-timed) as native Mandarin in terms of rate-independent (%V) or rate-normalized metrics, but it is shifted towards the stress-timed class in terms of raw durational metrics, regardless of the rhythmic class of the L1. Surprisingly, this does not accord with our expectation from the observation on isolated words, and also conflicts with the general hypothesis (i.e., L1 transfer effect) that the rhythmic pattern of nonnative speech should be intermediate between L1 and L2. However, this does coincide with our auditory impression of nonnative Mandarin speech – basically syllable-timed but meanwhile accompanied by mistaken stress assignment and relatively frequent breaks.

The first conflict can be explained by the fact that stress in continuous speech is not a simple copy of lexical stress pattern but is affected by many other factors. This has been confirmed by our auditory impression that some nonnative subjects assigned stress mistakenly in their utterances, enlarging the contrast of stressed vs. unstressed syllables. It is more difficult to interpret the second conflict, for which the differences in speech rate (viz., nonnative speech is generally slower due to the lack of fluency) can be the best explanation, especially for the larger raw metrics for nonnative speech. In addition, a complex and selective lengthening effect may also account for the observed results, as already mentioned in [9, 10].

This can be further analyzed from the view of prosodic structure. Table 4 shows the numbers of prosodic boundaries at different layers for all subjects. While there is little difference between CM and EM, it is obvious that nonnative speakers produced more prosodic boundaries than natives, especially for B1. The total numbers of higher-layer prosodic boundaries (B2 and B3) are similar, but nonnative subjects have fewer B2 and more B3 than natives.

The results in Table 4 are in line with our auditory impression that nonnative subjects were less fluent in speech and tended to divide an utterance into more chunks, sometimes with longer pauses in between, and even produced some words syllable by syllable. Table 5 lists the distribution of prosodic words with different numbers of syllables for all subjects, showing that CM and EM produced monosyllabic and disyllabic prosodic words more frequently than MM. Besides the lack of fluency, the stronger monosyllabic nature of Cantonese (than Mandarin) may also contribute to the large number of monosyllabic prosodic words for CM.

Because prosodic boundaries are usually accompanied by segmental lengthening such as word-initial and phrase-final lengthening, insertion of more prosodic boundaries leads to more lengthening, which may contribute to a higher variation of duration, causing the larger raw durational metrics.

Table 4. *Numbers of prosodic boundaries.*

Prosodic boundary	CM	EM	MM
B1	1047	1001	801
B2	144	165	247
B3	271	295	196
Total	1462	1461	1244

Table 5. *Numbers of prosodic words in various lengths.*

Group	1-syl	2-syl	3-syl	4-syl	5-syl	≥ 6-syl
CM	226	820	310	84	13	9
EM	184	849	338	74	12	4
MM	106	585	316	166	44	27

5. Conclusions

We have compared the rhythmic patterns of Mandarin speech between native speakers and two groups of L2 speakers who were native in Cantonese and English, respectively. Study on isolated words showed that nonnative speakers did not reduce unstressed syllables adequately. For continuous speech, eleven durational metrics were used to analyze rhythmic patterns. It was shown that nonnative Mandarin gave a quite similar rhythmic mode as native Mandarin in terms of rate-normalized or rate-independent metrics, but shifted towards stress-timed languages in terms of raw metrics, regardless of the rhythmic class of the L1. The result can be explained by speech rate difference and lengthening effects associated with the change in prosodic structure. It coincides with our auditory impression of L2 speech, for which the perceived rhythm may not be classified in a traditional binary/ternary way, but is usually a mixture. In this case, we may need to find better metrics, and the relationship between rhythmic metrics and other measures related to fluency and naturalness needs to be further studied.

6. Acknowledgements

This work is supported jointly by the National Social Science Fund of China (10CYY009 and 13BYY009), the Major Programs for the National Social Science Fund of China (13&ZD189), and the key project funded by the Jiangsu Higher Institutions' Key Research Base for Philosophy and Social Sciences (2010JDXM024).

7. References

[1] Pike, K.L., The Intonation of American English, Ann Arbor: University of Michigan Press, 1945.

[2] Abercrombie, D., Elements of General Phonetics, Edinburgh: Edinburgh University Press, 1967.

[3] Ladefoged, P., A Course in Phonetics, New York: Harcourt Brace Javanovich, 1975.

[4] Dauer, R.M., "Stress-timing and syllable-timing reanalyzed", Journal of Phonetics, 11: 51-62, 1983.

[5] Ramus, F., Nespor, M., and Mehler, J., "Correlates of linguistic rhythm", Cognition, 73: 265-292. 1999.

[6] Grabe, E. and Low, E.L., "Durational variability in speech and the rhythm class hypothesis", in N. Warner and C. Gussenhoven [Eds], Papers in Laboratory Phonology, 7: 515-546, Berlin: Mouton de Gruyter, 2002.

[7] Low, E.L., Grabe, E., and Nolan, F., "Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English", Language and Speech, 43: 377-401, 2000.

[8] Mok, P., "On the syllable-timing of Cantonese and Beijing Mandarin", Chinese Journal of Phonetics, 2: 148-154, 2009.

[9] Mok, P. and Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", Proc. Speech Prosody, pp. 423-426, Campinas, Brazil, 2008.

[10] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics, 35: 501-522, 2007.

[11] Li, A., "Chinese prosody and prosodic labeling of spontaneous speech", Proc. Speech Prosody, 39-46, Aix-en-Provence, France, 2002.

[12] Dellwo, V., "Rhythm and speech rate: A variation coefficient for ΔC", in P. Karnowski, & I. Szigeti [Eds], Language and language Processing, 231-241, Frankfurt: Peter Lang, 2006.

[13] Lin, H. and Wang, Q., "Mandarin rhythm: An acoustic study", Journal of Chinese Language and Computing, 17(3): 127-140, 2007.

[14] Ramus, F., "Acoustic correlates of linguistic rhythm: Perspectives", Proc. Speech Prosody, 115-120, Aix-en-Provence, France, 2002.